Finding Conflicting Statements in the Biomedical Literature

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2011

Farzaneh Sarafraz

School of Computer Science

Table of contents

Number of words: 85646

List of figures

List of tables

Abstract

The main archive of life sciences literature currently contains more than 18,000,000 references, and it is virtually impossible for any human to stay up-to-date with this large number of papers, even in a specific sub-domain.

Not every fact that is reported in the literature is novel and distinct. Scientists report repeat experiments, or refer to previous findings. Given the large number of publications, it is not surprising that information on certain topics is repeated over a number of publications. From consensus to contradiction, there are all shades of agreement between the claimed facts in the literature, and considering the volume of the corpus, conflicting findings are not unlikely. Finding such claims is particularly interesting for scientists, as they can present opportunities for knowledge consolidation and future investigations.

In this thesis we present a method to extract and contextualise statements about molecular events as expressed in the biomedical literature, and to find those that potentially conflict each other. The approach uses a system that detects event negations and speculation, and combines those with contextual features (e.g. type of event, species, and anatomical location) to build a representational model for establishing relations between different biological events, including relations concerning conflicts. In the detection of negations and speculations, rich lexical, syntactic, and semantic features have been exploited, including the syntactic *command* relation.

Different parts of the proposed method have been evaluated in a context of the BioNLP 09 challenge. The average F-measures for event negation and speculation detection were 63% (with precision of 88%) and 48% (with precision of 64%) respectively. An analysis of a set of 50 extracted event pairs identified as potentially conflicting revealed that 32 of them showed some degree of conflict (64%); 10 event pairs (20%) needed a more complex

biological interpretation to decide whether there was a conflict.

We also provide an open source integrated text mining framework for extracting events and their context on a large-scale basis using a pipeline of tools that are available or have been developed as part of this research, along with 72,314 potentially conflicting molecular event pairs that have been generated by mining the entire body of accessible biomedical literature.

We conclude that, whilst automated conflict mining would need more comprehensive context extraction, it is feasible to provide a support environment for biologists to browse potential conflicting statements and facilitate data and knowledge consolidation.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

It is neither entirely inappropriate for a thesis on negations and contrasts to start with a negated statement, nor, perhaps, for it to end with speculation. And, between the beginning and the end, throughout the highs and lows, achievements and uncertainties, false positives and lucky finds, I never lacked the support and inspiration to keep me stumbling along this obscure path that is research.

I am indebted to my supervisor, Dr Goran Nenadic, for his endless patience, encouragement, and inspiration. Goran was a fantastic supervisor: he understood my strengths and weaknesses from early on, and the implausible faith he showed in me is what has kept me motivated throughout these years. He helped enormously in the planning of this research, and did not let a single page pass by without clarifying and improving—and sometimes unceremoniously dismembering and redistributing—its contents with his infamously brutal red pen treatment. Four years after I first met Goran, the rate at which I learn from him has not decreased even slightly.

It was an honour to have Dr Ian Pratt-Hartman as my advisor. He made sure that I was on track all the way through my research.

Martin Gerner was a great colleague and a fabulous friend. He contributed enormously to parts of Chapter 5. Each of my ideas were bounced off him before reaching anyone else during the final year of this research, and he was the default victim of my occasional bouts of frustration.

I must also acknowledge Reza Mohammadi's critical contributions to parts of Chapter 3. Evemole is the product of us hacking together on a single Vim screen from thousands of miles apart.

This research would not have been possible without the voluntary contributions of the open source community. The entirety of this research, including the thesis itself, has been produced using the powerful tools and systems that

they have created. I also wish to thank the authors and publishers who provide free knowledge for all by granting open access to their full-text articles.

It must be hard to be the sysadmin in a computer science department and still preserve one's sanity, but Tony Curran was unfailingly prompt, patient, and approachable in his delivery of top notch technical support.

My colleagues at the GN TEAM provided a fun and vibrant environment for work punctuated by healthy doses of music, science fiction, films, and games. I am especially grateful to Mark, Hammad, George, Geraint, Azad, and Mona, as well as Shaobo. My fellow School of Computer Science Research Mentors—Grace, Jasmin, Kawther, Tianyi, Alex, and Salil—gave the place a homely feel.

Kaave Lajevardi initially persuaded me to do this PhD and, with characteristic equanimity, provided sound advice and support at every step of the way.

I am grateful to Fran Boon whose loving support started even before I moved to Manchester, and to his family who made me feel at home in the UK from day one.

Jonathan Caruana led me by the hand through the process of thesis writing, providing invaluable writing-up survival tips, insightful comments on the presentation of results and discussion, and a keen eye for aesthetic detail. His attentive care saved me from the ravages of self doubt and starvation.

I would like to thank all those whose friendship kept me sane and happy throughout my years in Manchester: Chris and Judith Lukey, George Stoppani, Francho Melendez, Lorelei Loveridge, Tom Sharp, Arash Eshghi and Raha Farazmand, Rosie Evans, Laura Marino, Philip Rafael, everyone in the Manchester Beethoven Orchestra, and especially Soroush Rafiee Rad.

Finally, I am forever indebted to my mother, who instilled in me a love of science, and to my father, who taught me to think critically. Throughout my life and in particular my time as a research student, they have been a constant source of emotional, moral, and of course financial support. To them I dedicate this thesis.

To my parents.

Chapter 1
Introduction

Text is the most common form in which human knowledge is stored. It is the primary means of communication among scientists, where knowledge is mainly communicated via research papers published in scientific journals and is widely available electronically.

Text is unstructured data. It relies on the readers' prior knowledge of the language and the specific subject to convey information by means of natural language. Text mining methods are designed in order to extract concise and structured information from natural language documents. Some text mining systems also aim to infer information that is not explicitly stated in the text.

Biomedical scientists use a particularly large and growing body of textual knowledge (Hunter et al. 2006). The main archive of life sciences literature currently contains more than 18,000,000 references and approximately 2,000 are added to this archive every day.[1] It is virtually impossible for any human to stay up-to-date with this large number of papers, even in a specific sub-domain.

With such a large and growing body of literature, and with the advances of technologies to store and process this data, life scientists are increasingly using automated technologies to access related work in their discipline. In addition to advanced search engines to search for and retrieve relevant documents, scientists have started to rely on text mining tools and methods to extract information from this pool of textual data.

The task of extracting information from text is done both manually and automatically, with various speeds and accuracies. Professional curators annotate biomedical papers and commit the reported facts into knowledge repositories. But with the vast amount of biomedical research recorded in

1 MEDLINE Fact Sheet, retrieved 30 September 2011 http://www.nlm.nih.gov/pubs/factsheets/medline.html

textual form, and with its rate of increase, automatic text mining tools and methods have become increasingly interesting to researchers.

The goal of text mining is to retrieve and extract information from text, and present it in a more concise and structured way to the user. Its domain stretches from lexical and syntactic analysis (parsing, part-of-speech tagging, named entity recognition, etc.) to semantic analysis (extracting roles and relations). The extracted information is typically inserted into databases (e.g. the STRING database (Szklarczyk et al. 2011)), or used as an input to other tools, or as support for manual curation.

Besides the enormous volume of the literature, the challenges of text mining particular to the biomedical domain include the language used by the scientists. Biomedicine is a dynamic area of science, and the language used in biomedical discourse evolves along with the development in methods and changes in experiments. Qualitative and quantitative descriptions, observations, and measurements are not always accurate in biomedical experiments, and accordingly, appropriate language is developed to reflect this characteristic. Claims are highly context-dependent, and therefore are described in long and speculative sentences. Other issues involve the variation in the terminology amongst individuals and across research groups, and the ambiguity of the language used by them (Ananiadou et al. 2005).

It is a well-known fact that there is a bias in the research that is shared with the scientific community through publication (Easterbrook et al. 1991); (Butler 2009). There is a tendency on the side of the researchers, editors, and pharmaceutical companies to handle the reporting of experimental results that are positive (i.e. showing a significant finding) differently from results that are negative (i.e. supporting the null hyphothesis) or inconclusive, leading to bias in the overall published literature. It has been found that statistically significant results are three times more likely to be published than papers affirming a null result (Dickersin et al. 1987).

This effect, referred to as “publication bias”, subsequently leads to

different linguistic styles to be used to report positive and negative results. It is expected for negated information to predominantly be reported in comparison or in contrast with similar affirmative information. In other words, when a negated statement is reported, it is likely that its significance is in comparison with other conflicting claims, or otherwise similar but slightly different positive claims.

The sentence in Example 1.1 is a clear example of several affirmative and negative reports of the production of three genes/proteins in different populations. It also demonstrates the information-richness and the complicated structure of some of these sentences, and the complexity of reasoning required to infer all the meaning expressed in them.

> **Example 1.1.** *"Although 21 out of 503 (4%) CD4+ T cell clones produced IL 4, but not IFN-gamma or IL 2, and 208 (41%) produced IL 2 and/or IFN-gamma, but not IL 4, a total number of 185 (37%) CD4+ clones showed the ability to produce IL 4 plus IL 2 and/or IFN-gamma."*
> (From PMID 2969818)

Of course, not every fact that is reported in the literature is novel and distinct. Scientists report repeat experiments, or refer to previous findings. Given the large number of publications, it is not surprising that information on certain topics is repeated over a number of publications. However, not all the mentions of a topic agree on every contextual detail.

From consensus to contradiction, there are all shades of agreement between the claimed facts in the literature, and considering the volume of the corpus, contrasting findings are highly expected to appear. Finding conflicting claims is particularly interesting for scientists, as they can present opportunities for future investigations and consolidation of knowledge. A conflict can be due to different experimental conditions, may suggest a potential contradiction, or

may indicate erroneous results. In any case, these are potential sources of hypotheses and further findings or inconsistencies in the entire body of biological knowledge.

To demonstrate how a person searches for and interprets relevant information, consider this example: a scientists, interested in the interactions between the HIV and host proteins, starts by using PubMed search engine's web interface to search for all the MEDLINE documents that have all of the terms *HIV-1*, *human*, *protein*, and *interactions*. At the time of writing this document, PubMed comes up with 3,049 articles after performing a document retrieval task. If she further wants to know what exact proteins of the HIV-1 virus interact with what proteins of the host and what the types of those interactions are, she would need to perform an information extraction task to extract the desired information. For instance, one of the documents retrieved by the above search is the document with the PubMed ID (PMID) 11336643. In the abstract of this paper she reads:

> *"a disulphide-bridged peptide mimicking the clade B HIV-1 gp120 consensus V3 domain (V3Cs) binds specifically to CCR5 (the major co-receptor of R5 HIV strains) on these cells."*

From the above sentence, she can infer the fact that the HIV protein *gp120* binds with the human protein *CCR5*. Furthermore, she can also conclude that the specific receptor in action is *receptor 5*, with *R5* mentioned in the brackets and also as a part of the protein name.

She then finds alternative (and preferably commonly accepted) names for the two proteins from one of the available databases, such as UniProt. The standard name for the *HIV-1 protein gp120* mentioned in the abstract found in the UniProt database is "*Envelope surface glycoprotein gp120*". Similarly, the name for the human protein would be "*chemokine (C-C motif) receptor 5*". After extracting this information, she then can summarise and represent this

fact as Table 1.1.

Interaction type	**Protein 1 (HIV protein)**	**Protein 2 (human protein)**
Binding	Envelope surface glycoprotein gp120 (UniProt ID: P03375)	Chemtokine (C-C motif) receptor 5 (UniProt ID: P51681)

Table 1.1: The representation of an event.

Suppose now she wants to find whether any other publications also support this interaction. However in the abstract of the article with PMID 22024519, she reads:

> *"N7K significantly increases the distance between V3 position 7 and sulphotyrosine at CCR5 position 14 (crucial for binding to gp120; from 4.22 Å to 8.30 Å), thus abrogating the interaction between these two important residues."*

So, there are cases reported in which this known interaction does not happen, perhaps after treatment or exposure to certain biological processes. This could be a starting point for our biologist to look into this interaction in more detail. Any systematic way of helping her would facilitate knowledge acquisition and consolidation as well as hypothesis generation.

This scenario simplifies how a biologist would analyse the literature and interpret the statements to understand their meanings. A number of activities are assumed when a human performs natural language perception. The purpose of information extraction is to "break" the task down into algorithmic steps so that it can be done automatically.

In this research, we are interested in finding conflicts, contrasts and potential contradictions in biomedical statements presented in literature. We use chemical interactions between certain types of organic molecules as a basic

unit of such biomedical facts. We refer to these interactions as “events”.

As intermediate steps in finding potential conflicts, we need to initially extract these units of information from text. We also need to extract information about whether these facts have been reported affirmatively or negatively, and whether they have been reported speculatively or with certainty.

We apply these methods on large-scale biomedical literature and explore how to extract contrasts and potential contradictions from such data.

1.1 Hypothesis and research question

We hypothesise that automatic extraction of contextualised molecular event information from textual data using state-of-the-art methods can be used to identify conflicting statements. In particular, we hypothesise that the addition of negation and speculation context to extracted event information on a large scale can find conflicting statements including contrasts and potential contradictions in textual research reports. This will be the main research question which this thesis aims to address.

1.2 Aim and objectives

The aim of this thesis is to investigate the way text mining can extract non-trivial and useful information from the biomedical literature by focusing on detecting the conflicting statements and facts. These phenomena are investigated at the event level.

Two event statements are contrasting when they state opposite but not necessarily inconsistent claims. They are contradictory when they also state inconsistent claims. We hypothesise that in at least one of the two contrasting or contradictory statements there appears a form of negation. Therefore, as intermediate steps, we aim to detect negations and speculations.

More specifically, the objectives of this research are the following.

1. Effectively identify biological events and relations among entities with

their context;

2. Design and implement a system that will be able to automatically recognise negated and speculated statements in text, specifically in the domain of molecular interactions;
3. Develop a representation model for establishing relations between different biological events, including relations concerning conflicts. This involves semantically representing a biological event.
4. Design and implement a system that will detect conflicting statements from a database of extracted claims;
5. Evaluate the proposed methodology through a case study on biomedical events;
6. Apply the method on the entire publicly available biomedical literature;
7. Provide the tools and data to the biomedical and text mining research communities, including the contextualised events and the conflicts between them.

The main focus of this research will be on standardised molecular and chemical events involving genes and proteins as examples of biological events. However, the proposed methodology aims to be generic and applicable to any biological fact.

1.3 Contributions

In this thesis, we designed and evaluated rule-based and machine learning techniques to extract events and their context from the literature using publicly available annotated data. We structure this extracted data using a semantic representation form for the event and its context which is an extension of a commonly used representational model. Finally, we propose techniques to find conflicting and contrasting facts in the data extracted from a large scale corpus of publicly available biomedical knowledge.

The research presented in this thesis has made the following contributions.

- A representational model for bio-molecular events and their context, appropriate for the detection of conflicting facts.
- A hybrid machine learning and rule-based method for molecular event extraction using dependency parse trees.
- A novel method to detect negations and speculations, using machine learning and computationally calculated *X*-command features along with other lexical, semantic, and syntactic features.
- A method to identify conflicting statements on molecular events from literature.
- An open source integrated text mining framework for large-scale identification of conflicting biomedical information.
- The large-scale data resulting from this analysis freely available for further biological explorations.

Intermediate results from this research have been presented and published in the following conferences and journals.

- Farzaneh Sarafraz, James Eales, Reza Mohammadi, Jonathan Dickerson, David Robertson and Goran Nenadic. "*Biomedical Event Detection using Rules, Conditional Random Fields and Parse Tree Distances*". Paper presented at the Proceedings of the BioNLP 2009 Workshop Companion Volume for the Shared Task in Event Extraction, 2009.
- Farzaneh Sarafraz and Goran Nenadic. "*Using SVMs with the Command Relation Features to Identify Negated Events in Biomedical Literature*". The Workshop on Negation and Speculation in Natural Language Processing, 2010.
- Farzaneh Sarafraz and Goran Nenadić. "*Identification of Negated Regulation Events in the Literature: Exploring the Feature Space*". Fourth International Symposium on Semantic Mining in Biomedicine

(SMBM), 2010.

- Farzaneh Sarafraz, Martin Gerner, Casey Bergman, Goran Nenadic. *"BioContext: integrated text mining for large-scale information extraction in biology"* (submitted.)
- Daniel Jamieson, Martin Gerner, Farzaneh Sarafraz, Goran Nenadic, David Robertson. *"Towards semi-automated curation: using text mining to recreate the HIV-1-human protein interaction database"* (accepted.)

BioContext was a joint project with Martin Gerner (Faculty of Life Sciences, University of Manchester).

All tools and references are available at http://gnode1.mib.man.ac.uk/.

1.4 Thesis structure

The rest of this thesis is organised in six chapters.

Chapter 2 presents the background and previous research on the topics related to our research. It introduces definitions of the concepts explored in this thesis. It critically evaluates tools, methodologies, and resources that were available at the time of this research.

Chapter 3 describes the research method used for the extraction and contextualisation of molecular events. It starts by the definitions of concepts that are used in the research. Section 3.3 describes the method developed for the automatic extraction of biomedical events from the literature. Sections 3.4 and 3.5 describe the methods developed to extract information about negation and speculation of these events.

Chapter 4 starts with the introduction of the evaluation approach and presents the results from molecular event extraction and contextualisation described in Chapter 3, along with evaluation and discussion.

Chapter 5 is mainly concerned with methods and the framework developed for aggregate analysis of contextualised biomedical events on a

large corpus. Section 5.1 describes the technical details of the text mining framework and the event extraction pipeline. Section 5.4 introduces a method for mining conflicting statements from the aggregate data.

Chapter 6 presents the results and the data of the large-scale text mining and aggregate analysis presented in Chapter 5 . It also evaluates the results and discusses the achievements and limitations of the research, exploring ways in which it can be improved and expanded in future.

Chapter 7 is the summary and conclusion of the thesis.

Chapter 2
Background

The aims and objectives of this research, introduced in Section 1.2, suggest that a wider background needs to be introduced and explored in order to put these objectives into context. In this chapter we introduce the context in which the objectives of this thesis are to be addressed.

Challenges that are of particular relevance to this research will be introduced in Sections 2.4 and 2.5, namely the recognition and extraction of negations, contradictions, and contrasts in general, and in biomedical text mining in particular.

Before that, we shall provide a brief summary of the main challenges in the field of biomedical text mining, and evaluate some of the existing approaches. In Section 2.1 we introduce information extraction as a general problem, with an emphasis on relation detection. Section 2.2 presents an overview of the biomedical literature, the domain which is used as a case study for finding conflicting statements. Section 2.3 explores the challenges in biomedical text mining that are considered to be prerequisites for mining conflicting statements in the biomedical literature. In this section we introduce pre-processing steps such as tokenisation and parsing, and critically discuss previous approaches to the problems of named entity recognition and relation extraction in the biomedical literature.

In recent years increasingly more gold standard corpora have become available to researchers. A selection of these resources are introduced in Section 2.6. They have been used in previous approaches, and will be used in this thesis as well.

Finally, in Section 2.7, we define a number of common evaluation measures and methods that are used in biomedical text mining.

2.1 Information extraction

Natural language, including written text, is unstructured. Although generating and understanding it is intuitive for humans, it is a complicated and non-trivial task to perform computationally. It contains an immense amount of ambiguity ranging from word sense ambiguity where a single word can have several unrelated meanings, to phrase structure and grammatical ambiguity where a word or phrase can have different grammatical roles or sometimes the whole sentence can have different syntactic parses, resulting in the sentence conveying two or more different meanings. On the other hand, a single concept can be expressed with different synonymous words or expressions, or using different grammatical structures. This is the opposite of ambiguity, and is referred to as variability.

Information extraction (IE) refers to the task of extracting facts from text written in a natural language about one or more predefined fact types, and representing those facts in a predefined form (Ananiadou et al. 2005). This "predefined form" is usually a template which is to be filled in with data extracted from text. These templates have the benefit of being more structured, and despite losing some of the context and thoroughness of the knowledge represented in unstructured text, can be used for aggregate processing once in a database. The results of IE are usually stored in a database for subsequent data mining, integrated into knowledge bases for reasoning, or presented to users.

For example, a template for weather reports can have slots for weather temperature, humidity, wind direction and speed, pressure, and weather felt temperature. Similarly, a template for interacting proteins could have the participating molecules, their roles, the type of the interaction, the anatomical location, and other properties of interest.

The manual information extraction task demonstrated in Table 1.1 is an example of how such a template is filled with data extracted from text to form the representation of a fact.

In the following sections we discuss how different parts of this task can

be done computationally.

2.2 Biomedical literature

The United States National Library of Medicine[2] (NLM) maintains a database of biomedical and life sciences scientific literature. The database is known as Medical Literature Analysis and Retrieval System Online (MEDLINE) and currently provides more than 18 million references from more than 5,500 journals in medicine, nursing, pharmacy, dentistry, veterinary medicine, health care and other areas of life sciences and biomedicine. The articles are indexed with NLM's controlled vocabulary, the Medical Subject Headings (MeSH)[3] which contains terms for a wide range of biomedical concepts, from molecular biology to organisms, health care, technologies, people, and more.

The MEDLINE archives go back to the 1940s and cover more journals every year. Although it is not the only archive of life sciences literature (see below) it is considered to be the main one and 2,000 new titles are added to it every day.

The amount of biomedical research information stored in MEDLINE is astonishing compared to most other areas of human knowledge. Figure 2.1 shows the number of new articles added to the database in each year since 1965. More than 600,000 new articles were added to the MEDLINE database in the year 2009 and more are added every year. Figure 2.2 shows the growth in the total number of archived abstracts in MEDLINE since 1980 until May 2010.

2 http://www.nlm.nih.gov/

3 http://www.nlm.nih.gov/mesh/

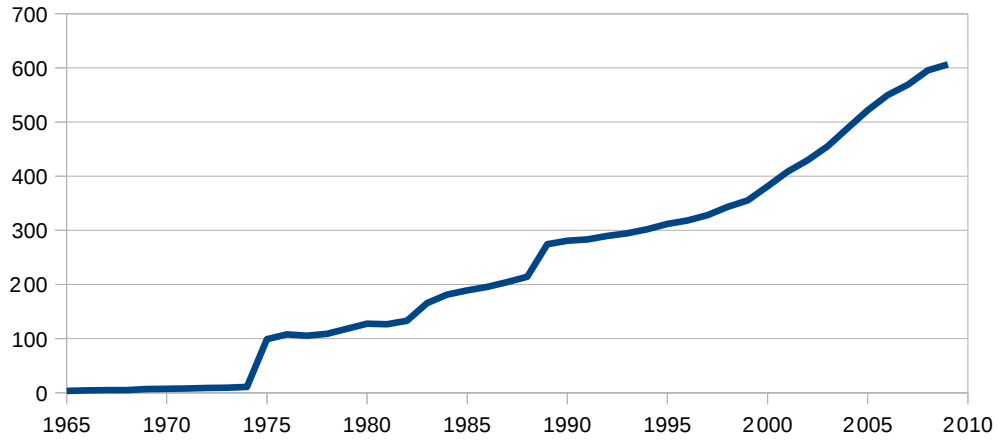


Figure 2.1: Number of additions to MEDLINE

This figure shows the number of additions to MEDLINE since 1980 (in thousands). The slight decrease in the rate of increase at the end of the graphs is due to the release dates being May every year and therefore only containing a subset of the final year's publications.

Note that these figures show the number of references that are in English and contain a title and an abstract. If one includes articles in other languages as well as those that are only referenced without an abstract and sometimes even a title, we will have even a larger corpus.

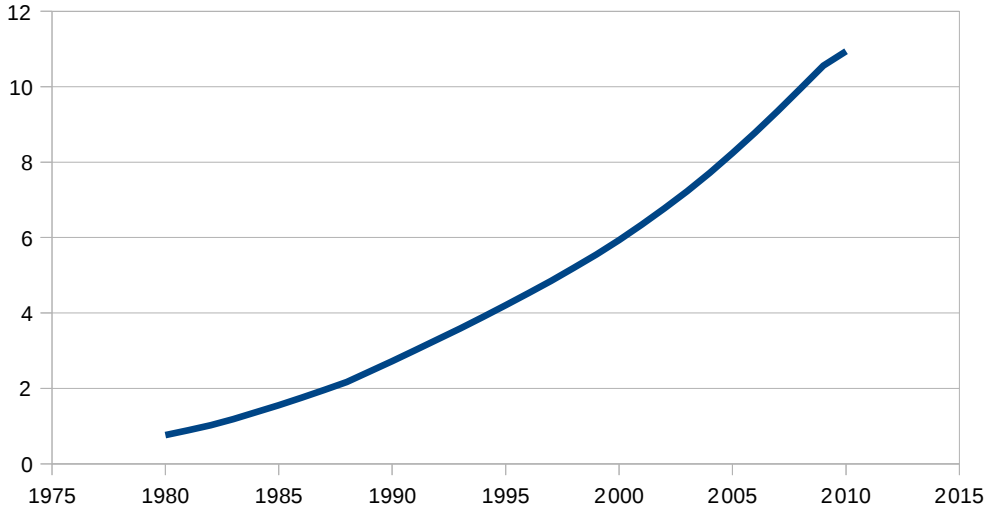


Figure 2.2: Cumulative number of abstracts in MEDLINE

Numbers are in millions.

There are many tools and services running on the MEDLINE database to provide easier and more efficient ways to access a database of this size. PubMed is the search engine to access the database, and is a part of the Entrez information retrieval system; both are provided by the NLM at the National Institutes of Health (NIH)[4]. Entrez Programming Utilities provide programmatic access to the data outside the web query interface of PubMed.

PubMed also provides various tools and services from term counters and entity mappers to alternative formats such as XML. It also makes scripting and pipelining platforms available for further development. MEDLINE is open access and freely available to everyone.

A number of other literature repositories are maintained that provide different ways and levels of access to the literature. PubMed Central (PMC)[5] is another biomedical and life sciences literature repository developed and maintained by the US National Center for Biotechnology Information (NCBI)[6]

4 http://www.nih.gov/

5 http://www.ncbi.nlm.nih.gov/pmc/

6 http://www.ncbi.nlm.nih.gov/

in the National Library of Medicine. PubMed Central provides free and open access to full text articles as opposed to MEDLINE that only provides access to abstracts and references. However, the number of articles that are provided through PubMed Central is more limited compared to that of MEDLINE.

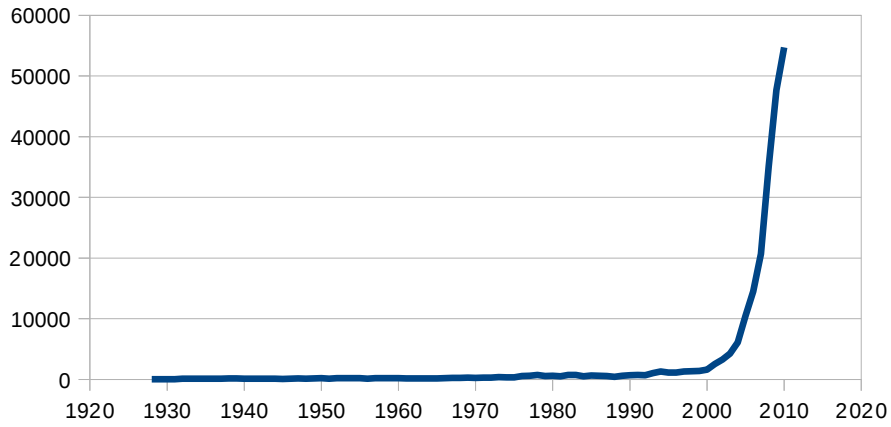


Figure 2.3: Full-text articles in open access PMC

The number of full-text articles in the open access part of the PMC by publication year.

PubMed Central currently hosts 2.2 million full text articles. Most (but not all) of these articles have their abstracts provided by MEDLINE. Although these articles are free to access and read by the researchers, they are not open for automated text mining, data mining or aggregate analysis. Only about 10% of the PMC documents (234,000 articles as in May 2011) are fully available and accessible for text mining research under a creative commons or similar license. Figure 2.3 shows the number of articles in the open access part of PMC, based on the publication year.

2.3 Biomedical text mining

As opposed to data mining which extracts patterns in large structured databases, text mining looks to extract new information and patterns from the data presented as texts written in a natural language.

Among the definitions proposed for the term Text Mining, the one by Marti Hearst (Hearst 2003) is commonly cited (Zweigenbaum et al. 2007) as a strict and conservative definition:

> "Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information [. . .] to form new facts or new hypotheses to be explored further by more conventional means of experimentation."

Although broadly used, this definition requires text mining systems to return knowledge that is not stated (or at least not explicitly stated) in text.

This excludes some valuable efforts such as information extraction or abbreviation handling from the domain of text mining. There are later definitions proposed that allow a broader interpretation of text mining than that of Hearst, allowing the systems to merely extract and link information from the text, or perform functions that are contributory to extracting information from the text. It is becoming increasingly common to use text mining as a facilitating tool to aid manual curation and increase its speed and accuracy (e.g. (Penagos et al. 2007) and (Jaeger et al. 2008)).

Text mining has a huge overlap with the more general domain of natural language processing (NLP), and is closely related to tasks like information retrieval and information extraction.

The one goal of text mining in biology that we discuss in this thesis is to extract facts from text. There are other activities in biological text mining that are not directly relevant to our subject of discussion here, such as text summarizing, question answering, etc.

Information extraction methods are initially aiming to extract explicitly stated facts from the text, and as they get more sophisticated, they are able to assist in what is known as literature-based discovery, as literature can be a

potential source of new hypotheses.

In this section, we introduce the key problems in this area, and discuss the previous efforts and achievements.

2.3.1 General overview of text mining work-flow

Most information extraction systems roughly follow the general work-flow depicted in Figure 2.4 wholly or partially. The relevant documents are selected from a large pool of documents in an initial document retrieval stage. Subsequently, pre-processing is performed on text, which can include anything from extracting the raw text from other formats like PDF to sentence splitting and tokenisation. Depending on the application, further processing is performed, potentially using a combination of tools and resources, to extract the required information in a structured way and store them in databases, provide them to the users, or feed them as the input of other systems.

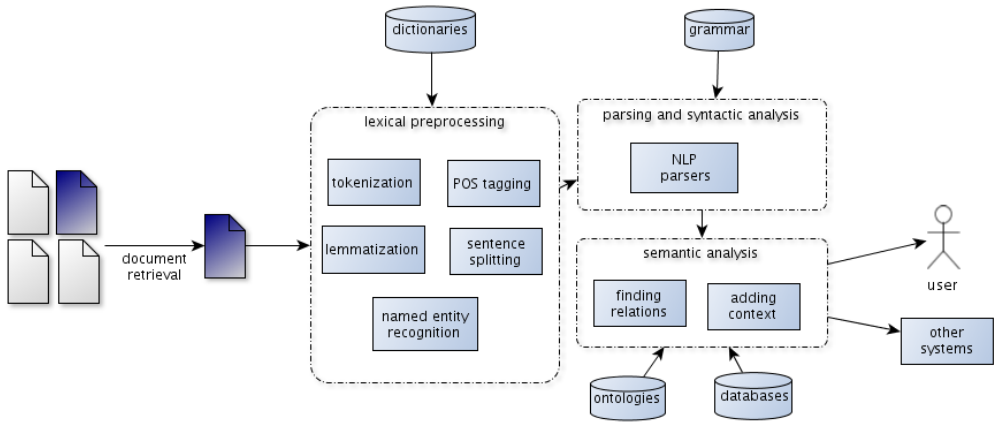


Figure 2.4: General TM work-flow

A schematic view of the general text mining work-flow.

In the next sections we will introduce different stages of this work-flow, and discuss the ones that are most closely related to this research in more depth.

Information retrieval

Information retrieval (IR) is the task of retrieving the documents that satisfy certain criteria from a big pool of documents. Search engines are examples of IR tools, and it is difficult to imagine research without the use of search engines. Besides general-purpose search engines such as Google and Yahoo!, there are specific search engines designed to perform information retrieval on the biomedical data.

One example of such search engines is PubMed which primarily accesses the MEDLINE database of citations and abstracts of biomedical research articles. PubMed is an example of a freely available information retrieval tool, specifically designed to retrieve biomedical documents from a large database. It provides features for specialised queries using MeSH Terms or publication type and year amongst others.

Another information retrieval engine is Entrez, which provides a search interface to many databases and resources including MEDLINE, PMC, and biological databases containing information about genes, proteins, pathways and interacting molecules.

Information retrieval systems play a key role in the text mining architecture. A text mining task typically starts with retrieving documents which are of interest to the task and then applying other processes such as classification and information extraction. It is a very vibrant area of research and specialised search engines are becoming more powerful and intelligent every day. However, although PubMed and other information retrieval systems are useful for retrieving documents of interest and narrowing the search, they do not at this point provide services for identifying and analysing relationships among biological entities.

Despite the recent advances in biomedical information retrieval, it is not yet considered a “complete” task as more development is still being done in this area. In 2010, one of the tasks in the BioCreative III challenge was to retrieve documents ranked in order of relevance to the query of a given gene.

The best performing team achieved the F-score of 61.42% in the ranking task ((Krallinger et al. 2010) (for a detailed definition of F-score and other evaluation measures see section 2.7.1.)

Sentence splitting

One of the first steps before analysing text is to identify the units of analysis, also known as *segmentation*. These units of analysis or segments can be sentences, phrases, words, etc. It is common in information extraction tasks to treat sentences as units of analysis, as they are the smallest syntactically and semantically self-contained unit of language.

Splitting the text into sentences, however, can introduce challenges. Rule-based methods that split the text based on more sophisticated versions of rules such as “period, followed by space, followed by capital letter” are widely used, but there are always exceptional cases for which such rules are not inclusive or exclusive enough.

Tokenisation

Tokenisation is the process of breaking text into linguistic or semantic units (called tokens) that constitute a useful piece of data for processing (Manning et al. 2008). The tokens can be words, symbols, or collocations. Tokenisation is a usual preprocessing step in many Natural Language Processing tasks.

Tokenisation is a computationally non-trivial task. Breaking the string on spaces does not always result in the desired output, as many semantic entities such as “New York” contain a space. Symbols can play several different roles in the English text, and can cause extra complications. An expression as simple as “aren’t” can be tokenised in a number of different ways, and it is not clear which one is the desired one (Manning et al. 2008). Hyphens are used with or without white spaces on either side and can indicate orthographical variation (e.g. “co-operation”), have a grammatical function (“security-checked baggage”), or many other functions.

In the biological language these ambiguities and variabilities become

more pronounced. Many common biological entity names are composed of several words separated by a combination of white spaces, hyphens, and other symbols. Examples include “NF-kappa B” and “TCR-alpha/beta”. Slash is sometimes used to indicate “or”. It can be used to indicate a chemical bond between two entities as in “TCR/CD3 ligation”. It can also be a part of an entity name as is the case in “ERK1/2”.

To address some of these complexities, tokenisation based on the semantic entities—rather than simply splitting on white spaces—have been considered by researchers for some applications. For example, Rinaldi et al. (2002) perform term extraction before tokenisation, and consider each term as a single token, regardless of the number of words contained in the term. Each such “semantic token” is then assigned the syntactic properties of the head of the term. Subsequently, the sentences are automatically parsed, processing multi-word terms as individual tokens. The authors show that this tokenisation improves the parsing process by 50% by removing the ambiguities and the complexities caused by the production of numerous possible parses.

Lemmatisation

Lemmatisation is the process of mapping different inflectional forms to their common base form. For example, *expression, express, expresses, and expressed* could all be mapped to the same base form *express.* The word that is being lemmatised may not have any morphological similarity with its lemmatised form, for example *be* is regarded as the lemma of *am, is*, and *are*. **Stemming** is an approximate computational method to achieve the same goal as lemmatisation, by truncating the end of the word using some rules. Examples of such algorithms are (Lovins 1968) and (Porter 1980).

Part-of-speech tagging

Part-of-speech (POS) tagging refers to the process of marking tokens in text with their lexical categories. The main lexical categories or “parts of speech” are shown in the following list, but most tasks require more refined categories

to also be tagged.

- Noun (N): any abstract or concrete entity
- Pronoun (P): any substitute for a noun or noun phrase
- Adjective (J): any qualifier of a noun
- Verb (V): any action or state of being
- Adverb (RB): any qualifier of an adjective, verb, or other adverb
- Preposition (IN): any establisher of relation and syntactic context
- Conjunction (C): any syntactic connector
- Interjection (UH): any emotional greeting (or "exclamation")

The Penn Treebank Project[7] uses a list of 36 categories (including the above) to mark up the sentences.

In the English language it is very common for words to have more than one possible lexical category. A word like “fiction” can only be a noun, but “secret” could be an adjective, a noun, or a verb, depending on the context. There are also ambiguous sentences which cannot be POS tagged in a unique way, and which can mean different things depending on the POS tagging.

Due to these complexities, automatic POS taggers take into account the dictionary definition of words, as well as the context in which they appear.

2.3.2 Named entity recognition and identification

One of the essential tasks in IE is to recognise the borders of what defines a named entity in text. This is called “Named Entity Recognition” (NER) (Béchet 2011). For example, NER involves recognising the boundaries of the two protein name mentions, *“HIV-1 gp120”* (or *“gp120”*) and *“CCR5”* in the scenario on page 26. As was observed earlier, this is not always a trivial task and needs complicated knowledge of the language as well as the specific domain.

Another more specific and advanced task is “Named Entity

7 http://www.cis.upenn.edu/~treebank/

Identification" (also know as "Normalisation") in which not only the boundaries of named entities are recognised, but the entity is "identified" by being mapped into a unique entry in a database of biological entities. This has immediate practical benefits and has received much attention lately (Morgan et al. 2007); (Hakenberg et al. 2008); (Huang et al. 2011).

In other words, the aim of NER is to identify the boundaries of a sub-string in text and the aim of normalisation is to map the sub-string to a predefined category which in biomedical text mining is usually a biological concept.

NER is a challenging task in general, and biomedical NER is in particular challenging due to the properties of the biomedical literature (Ananiadou et al. 2006). Despite the efforts to gather and maintain the world's scientific knowledge in databases, no complete database is yet available for most types of biological entities. Different research groups have different disciplines in the way they share their findings, funders do not always require the insertion of findings into databases, and institutes value textual publications in peer-reviewed journals more than submission of the data. Numerous initiatives such as NCBI have tried to create comprehensive databases, using centralised or collective efforts, and releasing of data is becoming increasingly important. However, complete databases of biomedical knowledge on which specialists have consensus are not yet available.

Even with the existence of complete databases and dictionaries, a different challenge will be *word sense ambiguity*, i.e. where the same word or phrase refers to different entities. For example, "cat" can be the name of a species of mammals (with the NCBI Taxonomy ID 9685), a human gene (with the NCBI gene ID 847), a protein (with the NCBI protein accession ID NP_001743), and a tomography method (Computerised Axial Tomography). On the other hand, most biological entities have several names among different communities, and even within the same community. The biological entities can have multi-word names, or names containing a combination of upper and lower

case letters and non-alphabetical characters such as numbers, hyphens and brackets. This adds to the complexity of word boundary recognition, overlap of the terms, and disjoint terms where the different parts of a term are separated by another word. For more discussion see (Chen et al. 2005).

Term recognition

Term recognition refers to recognising lexical units from text that correspond to domain concepts (Ananiadou et al. 2006). Single or multiple adjacent words that commonly appear together and convey a certain concept, e.g. “health care”, can be regarded as terms.

Identification of semantic concepts are important in information extraction tasks, as they often have a very specific meaning with colloquial usage. They can be constructed from multiple adjacent words where the meaning of the term is not directly correlated with the meaning of its parts. In some cases, they may not appear in common word lists, and specialised dictionaries need to be used to recognise them.

Automatic Term Recognition (ATR) systems utilise a number of approaches to extract and identify terms. Dictionary-based, rule-based, and machine learning approaches have commonly been used in ATR software. A specific example of term recognition is named entity recognition, discussed in section 2.3.2.

Gene name recognition and normalisation

Many biomedical text mining systems include a module to recognise mentions of biological entities, concepts, and terms in text (Ananiadou et al. 2005). Examples of the categories include genes, gene products, proteins, disease names, drugs, species, and so on. Depending on the particular task, these entities may then be identified by being linked to an ontology or knowledge base. Specifically, due to the varied and complex ways of writing about genes, and with the great number of genes researched and written about, gene name recognition and identification has been of great interest to biomedical text

mining.

Several methods have been developed to tackle the task of NER. Earlier attempts were rule-based, but as more annotated corpora became available various machine learning methods were applied to the task of NER. Lexicon-based approaches have been used for the subtasks where more complete ontologies and terminologies are available. Combinations of the above methods in hybrid systems are also common.

Tool	**Task**	**Availability**		**Performance**
		Binary	**Source (license)**	
ABNER	Gene NER	✓	✓ (CPL)	F = 0.72
BANNER	Gene NER	✓	✓(CPL)	F = 0.85
LingPipe	Gene NER	✓	✓ (own)	F = 0.56
GeniaTagger	Gene NER	✓	✓ (own)	F = 0.73
BioAnnotator	Gene NER	✗	✗	P = 0.94 R = 0.87
Whatizit	General purpose NEI	✗	✗	Depends on the underlying service that is called.
Moara	Gene NEI	✓	✗	F = 0.77 / 0.89 (Normalisation)
GeneTUKit	Gene NEI	✓	✗	TAP-5 = 0.48
GNAT	Gene NEI	✓	✓(BSD)	P = 0.54 R = 0.47 (cross species) P = 0.82 R= 0.82 (known species)
Prominer	Gene NEI	✗	✗	F = 0.80
TaxonGrab	Species NER	✓	✓(BSD)	P = 0.96 R = 0.94
LINNEAUS	Species NEI	✓	✓(BSD)	P = 0.97 R = 0.94

Table 2.1: Existing entity NER tools

Summary of the existing NER tools relevant to this research. The performance numbers are reported on different corpora, using different methods of evaluation.

Gene name recognition, gene normalization, and species name identification were among the most researched tasks in the first, second, and

third BioCreative challenges in 2004, 2006, and 2010 (Cohen et al. 2005); (Morgan et al. 2008); (Lu et al. 2010). The challenges were successful to elicit high performing systems from research groups around the world to the point that the state of the art in the gene name recognition is now determined by the output of some of the participating systems.

Although the applications developed for BioCreative and similar challenges are useful to determine the state of the art, almost all of them were developed for the specific task and not many of them were later available to the public. Table 2.1 summarises the existing NER and NEI systems relevant to this research, mainly gene/protein and species recognisers. Overall, F-score levels in the regions of up to 85% can be expected from most gene NER tools. Expectedly, gene normalisation is a more challenging task. For an evaluation of the existing systems and a summary of achievements see (Lu et al. 2010).

2.3.3 Parsing and syntactic analysis

Parsing is the process of breaking down a sentence into its constructing components (e.g. words, phrases, clauses, etc.) and determining the relations between these components to analyse its grammatical structure (Manning et al. 1999). Parsing is a form of syntactic analysis which helps determine how words or other parts of a sentence (e.g. phrases) relate to each other (Chapman 1988).

Different forms of syntactic representational models have been around for many centuries. Ancient grammarians that are known today include Pāṇini who wrote the formal grammar of Sanskrit around the 4th century BCE, the Greek grammarian Dionysius Thrax (2nd century BCE), and the Latin grammarian Priscian (5th century AD). The first formal theories of Arabic grammar (around the 10th century AD) were based on concepts similar to today's dependency grammar which will be discussed later in this section.

Sentences in natural languages often have syntactic and semantic ambiguities. For example, there are at least two possible ways to interpret the

sentence in Example 2.1.

Example 2.1. *"The chicken is ready to eat."*

It can be very difficult or sometimes impossible to produce a unique and correct parse tree for a given sentence (Aho et al. 1972). However, in order to get closer to understanding natural language sentences, much effort has been done to parse sentences automatically. Several tools have been developed to parse natural language sentences independently or as a part of challenges and shared tasks. We will introduce some of these efforts in this section, and discuss only the tools we have used in the present research.

Shallow parsing

Shallow parsing is perhaps the simplest form of phrase structure analysis. It identifies the boundaries of major syntactic constituents such as noun phrases and verb phrases, but does not specify their internal structure, or the relationships between these phrases in the main sentence.

GENIATagger (Tsuruoka et al. 2005) is a part-of-speech tagger and shallow parser specifically developed for the biomedical domain. The results of testing various trained models show an accuracy in the regions of 90% on the biomedical domain.

Noun Phrase	**Verb Phrase**	**Prepositional Phrase**	**Noun Phrase**	**Prepositional Phrase**	**Noun Phrase**
Several DNA-binding complexes	were detected	on	RAREs	in	undifferentiated cells

Table 2.2: Example of shallow parsing

The shallow parse of the sentence "Several DNA-binding complexes were detected on RAREs in undifferentiated cells" produced by GENIATagger.

Table 2.2 shows an example sentence together with its shallow parse

which is produced by GeniaTagger.

Dependency parsing

Dependency grammars were formally mathematically described by (Duchier 1999). Dependency grammars assume that syntactic structures consist of a *lexicon* and a set of rules called *dependencies* that relate these lexicals (Nivre 2009); (Duchier 2000).

Dependency parsing refers to parsing in the framework of dependency grammars. It determines the grammatical type of the different elements (e.g. words) and the structural relationship between them. Dependency grammar is concerned about how words relate to each other, specifically how pairs of words *depend on* one another. Examples of such relationships include subject, object, compliment, pre-adjunct, and post-adjunct.

For example, in the sentence "*John loves Mary*", "*John*" depends on "*love*" and the type of dependency is SUBJECT. Also, "*Mary*" depends on "*love*" and the type of dependency is OBJECT. This makes "*love*" the head of the sentence, and the root of the dependency parse tree as can be seen in Figure 2.5.

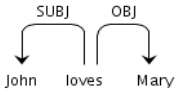


Figure 2.5: Simple example of dependency parsing

The dependency parse tree of the sentence "John loves Mary."

Types of dependency relations that are of interest in dependency parsing include relations such as subject (nominal or clausal subject), object (direct, indirect, or object of preposition), complement, prepositional modifier, noun phrase modifier, punctuation, etc.

The graph representing the dependency parse of a given sentence is in the form of a tree, as the dependency relations do not form a cycle. The

dependency distance between two words (or tokens) is defined as the tree distance between the nodes of the tree.

Figure 2.6 shows the dependency parses of a sentence from the article with PMID 8877104 from the GENIA corpus.

GDep is a dependency parser specifically developed for biomedical text. It combines previously researched probabilistic models with machine learning. It is trained on the GENIA corpus and reports an accuracy of 89% (Sagae et al. 2007b).

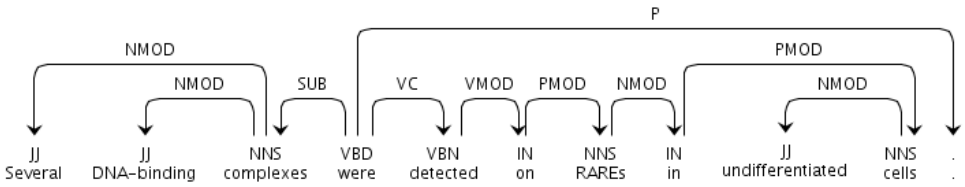


Figure 2.6: Example of a dependency parse tree

The dependency parse of the sentence "Several DNA-binding complexes were detected on RAREs in undifferentiated cells." produced by GDep.

Constituency parsing

Constituency parsing is another form of syntactic analysis of natural language sentences which represents the phrasal structure of the sentence. In a constituency parse tree, like other forms of phrase structure parse trees, only terminal nodes (leaves) are words, and the internal nodes of the parse tree are phrasal nodes. Internal nodes indicate phrases such as verb phrases (VP) and noun phrases (NP).

The constituency relation, like the dependency relation, is not cyclic. The graph denoting the constituency parse of a given sentence is a tree, and **constituency parse tree distance** is defined similarly to the dependency distance.

Constituency trees indicate word order relations along with dominance relations (i.e. which part dominates which), whereas the nodes in a dependency tree can be unordered. Unlike a constituency parse tree, all the nodes in a dependency parse tree are words, some of which are terminal nodes.

Figure 2.7 shows the constituency parse tree of the same sentence as in the previous figure (Figure 2.6).

The types of constituents that are of interest include sentence (S), noun phrase (NP), verb phrase (VP), adjectival phrase (JJ), prepositional phrase (PP), determiner phrase (DT), conjunctive phrase (CONJP), etc.

McClosky-Charniak parser (McClosky et al. 2010) is a statistical parser that recognises the constituency phrase structure of English sentences and performs with an F-score of 67% on the GENIA biomedical corpus.

Bikel (Bikel 2004) is another statistical constituency parser that is based on the Collins' parsing model (Collins 1999) which assigns a probability to each possible parse tree based on some properties of the phrase heads.

Enju is a probabilistic syntactic parser that produces constituency parse trees from English sentences. It has been trained and tested on the GENIA biomedical corpus and reports an accuracy of 87% (Hara et al. 2005).

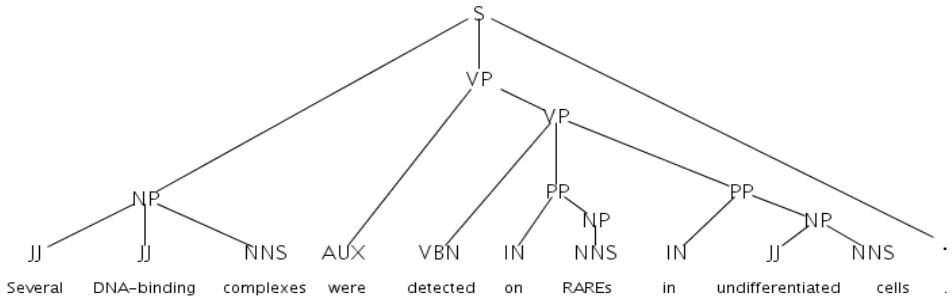


Figure 2.7: Example of a constituency parse tree

The constituency parse tree of the same sentence as in the previous figure (Figure 2.6) produced by McClosky parser: "Several DNA-binding complexes were detected on RAREs in undifferentiated cells."

For a thorough evaluation and comparison of state-of-the-art parsers, see (Miyao et al. 2008).

The command *relation*

Not every phenomenon within a sentence can be reduced to simple dependency or constituency relations, as the former only concerns simple binary relations between two tokens, and the latter discusses the structure of the building blocks of the sentence. Phenomena such as negation and anaphora often affect different and sometimes disjoint parts of the sentence, and run beyond sentence boundaries. Therefore, more in-depth sentence analysis is required in order to understand such phenomena.

The question of which parts of a syntactic structure affect the other parts has been extensively investigated. Langacker introduced the concept of command relation to determine the scope within a sentence affected by an element (Langacker 1969). Langacker originally defined the command relation as follows.

In a tree, and more specifically in the constituency parse tree of a sentence, we say that node *a* 'commands' another node *b* if

1. neither *a* not *b* dominates (i.e. is an ancestor of) the other; and
2. the *S*-node that most immediately dominates *a* also dominates *b*. In other words, the lowest ancestor of *a* with label *S* is also an ancestor of *b*.

Here, *S* refers to the sentence node, and also to any internal node indicating an independent clause. Note that the command relation is not symmetrical. Langacker observed that when *a S*-commands *b*, then a affects the scope containing *b*.

We will refer to this notion of the command relation as "*S*-command', and define a more general "*X*-command" relation similarly for any parse tree tag *X*. For simplicity, we say "command" when we mean *S*-command. Some later uses of the command relation such as (McCawley 1993) have chosen to allow the nodes to dominate each other, and therefore ignore the condition 1 above. According to (Barker et al. 1990), none of these authors gave definite motivation or strong support for this exclusion. Langacker has also observed that, in the case of anaphoric relations, condition 1 automatically holds, and is therefore redundant.

Figure 2.8 shows the command relation in a given parse tree. In this tree, node *a S*-commands node *b*, since the lowest ancestor of *a* with label *S* is also an ancestor of *b*. However, *b* does not *S*-command *a*, as it is placed in a sub-tree with a head labelled *S* which does not contain *a*.

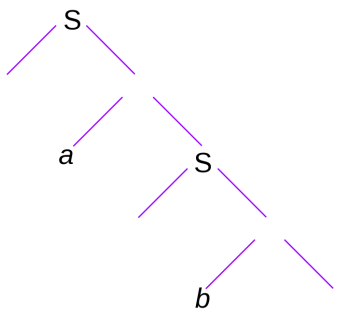


Figure 2.8: The command relation on a sample parse tree.

Node a S-commands node b whereas node b does not S-command node a.

Figure 2.9 shows the (partial) parse tree of Example 2.2.

> **Example 2.2.** *"We now show that a mutant motif that exchanges the terminal 3' C for a G fails to bind the p50 homodimer [...]"*
> (From PMID 9442380)

This sentence contains the word *fails* that indicates the existence of a negation. However, the sentence expresses several concepts, not all of which are affected by the negation cue. The concepts expressed by verbs *show* and *exchanges* are expressed affirmatively, whereas *bind* is negated.

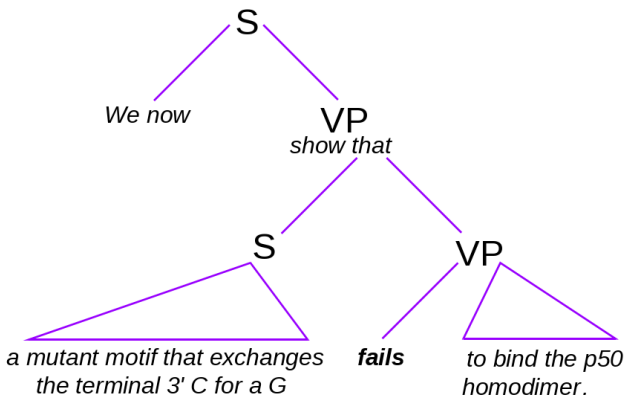


Figure 2.9: The command relation on a sentence.

The schematic parse tree of the example sentence "We now show that a mutant motif that exchanges the terminal 3' C for a G fails to bind the p50 homodimer [...]" The word "fails" VP-commands the interaction trigger "bind" but not the other parts of the sentence.

Figure 2.9 shows that the word *fails* *VP*-commands the sub-tree that contains *bind*, but not the other parts of the sentence.

Variations of the command relation have been proposed to explain and categorise various linguistic phenomena. For example, (Lasnik 1976) explores the connection between the command relation and anaphora by proposing a "kommand" relation which was rephrased by Barker et al. as the intersection of *S*-command and *NP*-command, and suggested that it could be relevant for the description of the constraints on anaphora.

Klima argued that assuming that negation only affects the constituent where the cue appears would not explain the function of negation which is more complex (Klima 1964). He then introduced a relation between two nodes in a constituency parse tree which he refers to as "in construction with", and which others refer to as the command relation. Klima shows that the command relation explains the structure of numerous expressions of negation. Specifically, he speculates that the part of the sentence which is affected by the negation cue is that which is commanded by it.

These definitions and discussions have so far only been proposed theoretically, with no statistical evaluation reported to our knowledge.

2.3.4 Relation Extraction

As introduced in Section 2.1, in information extraction, we are interested in extracting facts from text. These facts are usually relations among entities, and are extracted in the form of templates that need to be filled (as in Table 1.1). The entities are either already known or are recognised in the previous NER stages (see Section 2.3.2). This is an area which has attracted recent research (Cohen et al. 2005).

Many biomedical facts and functions can be formulated as relations between entities. Interactions between proteins (also known as biomedical "events") can be represented as tuples containing the interacting proteins and the interaction type. Medical treatments can be represented as drug-disease pairs, probably with more context added regarding dosage, side effects, duration of treatment, mode of application, etc.

Relations as basic units of scientific facts are widely accepted. Extracting them by means of automatic mining have been increasingly important and several community challenges have been organised to address this problem.

In recent biomedical relation extraction studies, most emphasis has been on the relations between genes and proteins (Cohen et al. 2005). It is believed that detecting common function in a set of genes is useful in identifying functionally interesting ones (Raychaudhuri et al. 2002). Therefore a lot of text mining research has been around grouping genes with similar functions according to the textual clues in the sentences they appear. There has also been research around specific relationships between genes and proteins.

In the following section we discuss one such challenge that is closely related to this research. Subsequently, we study the methodologies used across the literature for the task of relation extraction.

Extraction of molecular events—a community challenge

In 2009, the Genia group[8] together with the BioInfer group[9] and the U-Compare initiative[10] organised a shared task whose main aim was the extraction of bio-events from the literature, focusing particularly on bio-molecular events involving proteins and genes. The BioNLP'09 Shared Task (Kim et al. 2009) was designed to address a semantically rich information extraction problem as a whole, divided into three subtasks.[11] Task 1 required biomedical events and their participants to be detected in text, task 2 involved recognition of location entities and assigning these entities to the events, and task 3 involved further characterising the events as being negated or speculated.

The tasks assumed that named entity recognition was already performed on the text and for the purposes of the challenge, manual gold annotations for gene and gene product entities were provided.

The challenge defined an **event** as a structured collection with the following properties:

1. Every event has a **type** which is the biological type of the process e.g. *regulation* or *gene expression.* A total of nine event types were considered.
2. Every event has a textual **trigger** which is the part of the sentence that indicates the expression of an event.
3. Events have one or more **participants**:

   (a) Every event has at least one **theme**, which is usually the protein

8 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

9 http://mars.cs.utu.fi/BioInfer/

10 http://u-compare.org/

11 More than 40 teams from research groups around the world expressed initial interest in participating in the Challenge. Final submissions were received from 24 teams who completed task 1, and six teams completed each of tasks 2 and 3. The results and methodologies were presented in the BioNLP workshop as part of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009 conference.

entity that is affected by the process.

(b) Some events may have a **cause**, which is usually the protein entity that causes the process.

(c) Themes and/or causes can sometimes be other events, forming **nested events**.

The following nine event types were considered: *gene expression*, *transcription*, *protein catabolism*, *localisation*, *phosphorylation*, *binding*, *regulation*, *positive regulation*, and *negative regulation*. Depending on the event type, the task included the identification of either one (for the first five event types mentioned above) or more (for *binding*) themes. Information requested for regulatory events was more complex: in addition to one theme (an entity or another event), these events could also have a cause (another entity or event.)

Tables 2.3 and 2.4 show two example sentences from the BioNLP corpus. The example in Table 2.3 is a simple event (class I) of type "gene expression" which has one entity theme (IL-2).

Event	**Trigger**	**Type**	**Theme**	**Cause**
Event 1	"induction"	Gene expression	IL-2	-

Table 2.3: Representation of an event from the BioNLP'09 corpus

An example sentence from the BioNLP'09 training data with a gene expression event annotated: "The effect of this synergism was perceptible at the level of induction of the IL-2 gene."

Amongst the four events annotated in the example in Table 2.4, two have participants that are biomedical entities (events 1 and 2) and the other two have participants that are events (events 3 and 4). Note that a sentence can express more than one molecular event, and a string can be the trigger of more than one event. There is no limit on the length of the trigger, and events can span across sentences.

Event	Trigger	Type	Theme	Cause
Event 1	“transcription”	Transcription	FasL	-
Event 2	“Overexpression”	Gene expression	ALG-4	-
Event 3	“Overexpression”	Positive regulation	Event 2	-
Event 4	“induced”	Positive regulation	Event 1	Event 3

Table 2.4: Representation of four event in a sentence from the BioNLP’09 data.

An example sentence from the BioNLP’09 training data with four events annotated, some referencing others:“Overexpression of full-length ALG-4 induced transcription of FasL and, consequently, apoptosis.”

The composition of the data sets is presented in Tables 2.14 and 2.15 in Section 2.6.

We can further categorise different event types into three event classes. Simple or class I events are those that have exactly one theme, and this theme is a named entity (protein). Events of types gene expression, transcription, protein catabolism, localisation, and phosphorylation belong to this class. Class II events are events that have one or more theme. The only type in this class is binding. Finally, class III events are complex events with a theme and an optional cause, which can be an entity or another event. This class includes regulatory events: regulation, positive regulation, and negative regulation.

Studying the BioNLP’09 training data showed that 95% of annotated events are fully contained within one sentence (Björne et al. 2009). Moreover, 92% of the event triggers are a single token, and the other 8% are adjacent tokens (Björne et al. 2009). A token or a group of adjacent tokens in a sentence can act as the trigger for several events, possibly even of different types. The words that act as triggers cannot be recognised by a simple dictionary look-up as there is a high level of word sense ambiguity. The same word or group of words can be the trigger of an event in some cases and not a trigger in others. They can also indicate events of different types across the corpus, so the type

of the event does not directly correlate with the trigger lexicon.

For example (Björne et al. 2009) observed that only 28% of instances of the word *"activates"* in the corpus are triggers for an event, and the instances of the word *"overexpression"* are evenly distributed between gene expression, positive regulation, and no trigger.

BioNLP'09 task 2, which concerns assigning location entities to localisation events, is not directly relevant to the subject of this thesis and will not be discussed here. Task 3 requires further classification of the extracted events in task 1, by determining whether an event is affirmative or negative, and whether it has been stated certainly or speculatively. We will introduce these concepts in more detail in Section 2.4.

To demonstrate the requirements of task 3, consider the sentence from an abstract shown in Example 2.3.

Example 2.3. *"In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain."*

The proteins *TRAF2* and *CD40* are already manually annotated in text with their indices (57, 62) and (88, 92). Task 1 required event annotation, in which the following events will be extracted:

Event	Trigger	Type	Theme(s)	Cause
Event 1	"phosphorylation"	Phosphorylation	TRAF2	-
Event 2	"binding"	Binding	TRAF2 / CD40	-
Event 3	"inhibit"	Negative regulation	Event 2	Event 1

Table 2.5: Events from an example sentence, before negation/speculation

Annotations for the events extracted from the sentence "In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain." Negation and speculation detection task has not yet been performed.

Task 3 required the marking of Event 3 (see Table 2.5) as speculated since it has been expressed as a hypothesis by the authors, and is not a certain

fact. The output of performing this task on the example sentence is shown in Table 2.6.

Event	Trigger	Type	Theme(s)	Cause	Negation	Speculation
Event 1	“phosphorylation”	Phosphorylation	TRAF2	-	0	0
Event 2	“binding”	Binding	TRAF2 / CD40	-	0	0
Event 3	“inhibit”	Negative regulation	Event 2	Event 1	0	1

Table 2.6: Events from an example sentence, after negation/speculation

Annotations for the events extracted from the sentence “In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain.” Negation and speculation detection task has not yet been performed. Event 3 has been annotated as speculative.

Co-occurrence and statistical methods

The simplest way to detect relations between biomedical entities is to collect documents or sentences in which they co-occur. Co-occurrence statistics can provide high recall but typically have poor precision (Kilicoglu et al. 2009), and are now used more as a simple baseline method against which other methods are compared (Cohen et al. 2008).

Statistical methods aim at detecting relations by looking for structures, terms, and patterns that co-occur more frequently in the desired expressions than would be predicted by pure chance. Lindsay et al. describe an example of a predominantly statistical approach in (Lindsay et al. 1999).

Albert et al. focused on a semi-automatic method of retrieving protein-protein interactions (Albert et al. 2003). Their method was to retrieve the co-occurrence of two protein names and one interaction term in one sentence and then manually checking the abstracts containing one such “tri-occurrence”.

Rule-based methods

To increase the precision, several rule-based approaches such as the one described by Yu et al were generated by the domain experts which most commonly use regular expressions (Yu et al. 2002).

Other methods such as that of Friedman et al. rely on the thorough analysis and parsing of the text in order to extract the information from each sentence according to the linguistic semantics of the text (Friedman et al. 2001). These methods generally result in better accuracy especially on smaller corpora, but are costly in terms of the time taken for hand-crafting rules and still can miss out exceptional cases expressed in less common ways.

Spasic et al created a rule-based system which uses morphological, lexical, syntactic, and semantic features to extract information about the medication used by a patient from a medical report (Spasic et al. 2010). The desired information contained name, dosage, route, frequency, duration, and reason of the drug administered. They manually created patterns in which this information appears, and combined them with heuristic context-sensitive rules.

A number of attempts have been made to extract other relations between genes, proteins, and other biological entities using rules. Rinaldi et al. constructed an event extraction system, OntoGene, that uses manually created patterns based on the syntactic parse of sentences (Rinaldi et al. 2006). They initially detect these syntactic patterns. Subsequently, they combine various patterns into a single semantic rule that represents different syntactic phenomena (e.g. passive voice, nominalisation, etc.) Finally they combine these rules with terms and ontologies to extract events.[12]

An example of such a rule is "*A triggers the H of B*" where H is a nominalised verb, such as *activation*, and *A* and *B* are reported as participants. Their systems was evaluated on the GENIA corpus using post hoc validation of the output and reported a precision of around 90% on selected events. Like other post-hoc evaluations, precise recall measures were not reported, but was

12 The system can be accessed at http://www.ontogene.org/

estimated in the range of 38%-50%.

Kilicoglu et al. used a rule-based methodology for the BioNLP'09 Shared Task on event extraction (Kilicoglu et al. 2009). They construct patterns from the known trigger words, and defined a selection threshold to handle ambiguity and term variability.

To associate participants to the triggers to form events, they statistically analysed the dependency paths between the event triggers and their participants. They observed that the distribution of these paths obeyed Zipf's law, with 70% of the paths occurring only once. They constructed a total of 27 hand-crafted rules involving the dependency paths for the most common trigger terms. For example, one such rule was the existence of the direct object (dobj) dependency between verbal event triggers and themes.

Not all the rules involved dependency paths. For instance, NPs with hyphenated adjectival modifiers, such as *"LPS-mediated TF expression"* were reported as a regulatory event with the NP as the cause.

Overall, they achieved P/R/F-score of 61%/33%/43% in the event extraction task.

Machine learning

With the provision of annotated training data, machine learning has become an effective method in all areas of text mining, including biomedical information extraction.

Support Vector Machine is a statistical learning method that has been widely used for relation extraction in biomedical text mining (Burges 1998).

SVMs have been used by many researchers to extract protein interactions. Mitsumori et al. used bag of words features around protein names (Mitsumori et al. 2006). Yakushiji et al. defined patterns on predicate argument structures on the syntactic dependency parse tree of the sentence and used them as SVM features to extract relations between interacting proteins (Yakushiji et al. 2006). Many researchers, including (Sanchez 2007), (Culotta et al. 2004),

(Kilicoglu et al. 2009) and (Swaminathan et al. 2010) have used the properties of the dependency paths between protein names and other indicators of molecular interactions to extract information about them.

Sanchez (2007) trained a maximum entropy model to classify pairs of protein names as to whether they interact or not. They used features including the protein name word forms, the stems of the words between the two protein names and surrounding them in a window of size 5, whether or not the trigger falls between the two proteins, and if so, one or more.

(Culotta et al. 2004) used SVM to detect and classify relations between entities in text. They define a kernel function that returns a similarity score between two trees. They only consider the smallest sub-tree in the dependency tree that includes both of the entities in question, and use their defined kernel to train a classifier to detect relations such as roles, parts, location, etc. from news articles, and show that dependency tree kernel improves the F-score by 20% compared to the usual features such as POS and entity types.

(Kilicoglu et al. 2009) extracted molecular events in the form of the BioNLP'09 Shared Task, using SVMs with features including the vertex walks on the dependency paths between the event trigger and the participants. They included dependency types and word forms, but blinding the trigger term and protein names. They reported P/R/F-score of 33%/52%/41% on the BioNLP'09 test data set.

An important characteristic of the approaches using SVM is their choice of features. Word form, stem, part-of-speech tag, dependency path tags and distances, character and token *n*-grams, token position and length, and membership in lexical and semantic dictionaries are only some of the features commonly used in relation extraction tasks.

Naïve Bayes methods have been used by (Donaldson et al. 2003) for extracting protein-protein interactions. A maximum entropy method was used by (Xiao et al. 2005) as a supervised learning approach to extracting protein-protein interactions.

Conditional Random Fields (Lafferty et al. 2001) are probabilistic models for predicting a collection of class labels (usually a sequence) simultaneously. Unlike Hidden Markov Models (HMM) and Maximum Entropy Markov Models (MEMM), CRFs do not require the instances in a sequence to be independent. HMMs and MEMMs try to assign a label sequence $Y=(y_1, y_2, ...)$ to an observation sequence $X=(x_1, x_2, ...)$ by maximising the conditional probability $P(y_i|x_i)$ as i ranges over the data sequence. These methods require the assumption that the instances in the sequence are independent, and could be observed in any order. This is not usually a correct assumption in the problem of token labelling, as tokens in text are inherently sequential and dependent. CRF addresses this issue by maximising the conditional probability $P(Y|X)$ for the sequence. It does that by modelling state transitions when predicting the sequence of labels as well as the overall probability of states.

CRFs have been shown to be particularly suitable for sequential data such as natural language, since they take into account features and tags of neighbouring tokens when evaluating the probability of a tag for a given token. They have been used in identifying molecular events by (MacKinlay et al. 2009), among other researchers. (Yang et al. 2008) used CRFs to, given a sentence discussing transcription factors (a protein that is involved in the molecular event transcription), identify transcription factors that will affect other proteins. They used features involving the phrase types, a dictionary of known transcription context lexicon, protein and gene names, interaction words, and other biological terms.

The best results in the BioNLP'09 event extraction task were from the University of Turku (Björne et al. 2009) who achieved an overall P/R/F-score of 58.5%/46.7%/51.9%. These results were also the highest for each of the three classes of events, namely simple events (single theme), binding events (multiple themes), and regulation events (recursive, possibly having a cause as well as a theme.) The highest recall/precision/F-score for class I were

64.2/77.4/70.2, for class II they were 40.1/49.8/44.4, and for class III they were 35.6/45.9/40.1.

Towards the conclusion of our research in 2010, the University of Turku researchers made an improved implementation of their system publicly available as an open source code.

The Turku Event Extraction System (TEES) method of detecting triggers is effectively a token labelling problem, similar to named entity recognition. Each token is assigned to one of the nine types, or a negative class for tokens that are not triggers of any event. TEES uses multi-class SVMs to detect triggers. They used token features (e.g. punctuation, capitalisation, stem, character bigrams and trigrams) for tokens in a window of radius 1 and frequency features (e.g. the number of named entities in the sentence and near the token). They also included dependency features including the dependency types and the sequence of dependency types up to a depth of three from the token in question.

Once the triggers are extracted, TEES uses a graph-based method to assign participants to triggers. They use another multi-class SVM to classify any possibly edge between a named entity and a trigger or between two triggers in the case of nested events as either *theme, cause,* or neither. The edges are labelled independently, and then later pruned using rules based on the task constraints.

EventMiner (Miwa et al. 2010) is another system that was developed to extract events in the BioNLP'09 representation. It uses machine learning and dependency tree features to profile events, following a similar work-flow as TEES. EventMiner differentiates between the triggers that affect proteins and those that affect other triggers, and use two separate multi-class SVM classifiers to classify words as triggers. For features, they use lexical features such as capitalisation, numeric characters and punctuations, and character n-grams. They also use dependency features such as n-grams of dependency paths, n-grams of POS and base forms, and lengths of paths.

Similarly to the TEES system, to assign triggers to participants, EventMiner also prunes the trigger-participant edges using two SVM classifiers: one to assign triggers to other triggers in nested events, and another to assign triggers to named entities. In addition to the TEES features, they also use the confidence of participant prediction. In nested events, they also add shortest dependency path features between the participant trigger and the respective entity participants.

They use separate classifiers for each of event classes I, II, and III, participation types (theme and cause), and participant types (protein or event). They report F-scores of 70%/65%/47% for the extraction of events of classes I/II/III using the approximate recursive matching. For more details of these evaluation measures see Section 2.7.1.

As this thesis was being written, The Stanford Natural Language Processing Group released their software that extracts molecular events from biomedical text, redefining the problem to be comparable with the problem of constructing dependency parse trees from the sentence (McClosky et al. 2011). Similar to the previous approaches, they use a multi-class classifier (logistic regression) to detect the event triggers, using features including word form, lemma, membership in a set of known interaction words, surface context of window size 1 on either side, dependency paths down to depth 2, and entity count. Secondly, the participants are being assigned using a method similar to forming the parse tree of a sentence, based on the conversion of the event representation to a tree representation with nodes representing the event trigger and its participants, and with edges labelled with participation type *(theme* or *cause).*

They report P/R/F-score of 59%/49%/53% on the BioNLP'09 development data, and 57%/43%/49% on the test data.

Other approaches

In addition to the above examples that predominantly use one method or

another, many applications use a combination of more than one technique to achieve the best performance.

Chiang and Yu's MeKE system (Chiang et al. 2003) is an ontology-based system that uses semi-automatically constructed patterns to extract the functions of gene products. Their rule-based method is combined with sentence classification (Naïve Bayes) to determine the type of the function.

2.4 Recognition and extraction of negation and speculation

When we extract information from text, an important piece of information is whether the information is expressed in text as negated or affirmative. It is also important whether the information is stated certainly or speculatively. Negation and speculation in information extraction can affect the quality and accuracy of the extracted information, and has been the focus of much research in recent years. For example the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP 2010) brought together researchers working in this field, many of whom with particular interest in biomedical information extraction (Morante et al. 2010b). Many ontologies are currently expanded to include information about negated relations as well as affirmative or 'realistic' relations (e.g. (Ceusters et al. 2007) and (Fleischhacker 2011)).

Negated and speculated statements in text are not trivial to extract and analyse. Negations and speculations are expressed in many forms, including highly complicated and ambiguous forms. Even words like *not* that may seem to always indicate negation can appear in phrases that express no semantic negations. For example, in the phrase "*not only A, but also B*" both concepts A and B are mentioned as present, and therefore no negation or absence is expressed despite the appearance of the word *not* which we will later see that is a strong negation cue (See Figure 3.16).

Multiple negations in a sentence can also introduce more layers of ambiguity. It is not unusual even for humans to have difficulty parsing and

understanding the meaning of a sentence due to the use of negated patterns in it.[13, 14]

2.4.1 Negation and speculation terminology, concepts, and definitions

There have been numerous contemplations on the concepts of negation and speculation. Here we adopt a definition of negation as given by the Cambridge Encyclopedia of Language Sciences: "Negation is a comparison between a 'real' situation lacking some element and an 'imaginal' situation that does not lack it" (Lawler 2010). The imaginal situation is *affirmative* compared with the negative real situation. The element whose polarity differs between the two situations is the negation target.

Negations in natural language can be expressed by the use of negating words such as *no*, *not*, or *never*, or by specific expressions (e.g. *absence of*, *failure*, etc.) The word or phrase that makes the sentence wholly or partially negative is typically referred to as the negation cue and the part of the sentence that is affected by the negation cue and has become negative is the negation scope.

> **Example 2.4.** "*Tandem copies of this 67-bp MnlI-AluI fragment, when fused to the chloramphenicol acetyltransferase gene driven by the conalbumin promoter, stimulated transcription in B cells but* ***not*** *in Jurkat T cells or HeLa cells.*"
>
> (From PMID 1986254 annotated by BioScope corpus annotators.)

In Example 2.4, the word *"not"* indicates a negation, and therefore is the

13 Liberman, M., *Why are negations so easy to fail to miss?*, Language Log, February 24, 2004.

14 Consider, for example, the sentence "There has never been a time when there has been no person in Cornwall without a knowledge of the Cornish language." from Henry Jenner, *Handbook of the Cornish Language* (1904).

negation cue. The part of the sentence that is underlined is the scope of negation.

> **Example 2.5.** *"In vitro translated hGR was capable of selective DNA binding even in the **absence** of glucocorticoid."*
> (From PMID 1944294 annotated by BioScope corpus annotators.)

In Example 2.5, the word *"absence"* is the negation cue, and the human annotators have considered it as a part of the negation scope as well.

Speculative statements, on the other hand, are not necessarily explicitly asserted in the text (Light et al. 2004). They are to some extent true (or false) but there is not a definitive confirmation about their status, which makes them more or less uncertain. Many authors also consider statements that show insufficient knowledge, or express speculative questions or hypotheses as speculation (Medlock et al. 2007); (Szarvas et al. 2008). Like negation, speculation is often indicated by a cue, which could similarly affect all or part of a sentence, i.e. the speculation scope.

> **Example 2.6.** *"This zinc-finger region, which is **thought** to bind DNA in a sequence-specific manner, is similar (greater than 80% on the amino acid level) to two previously described transcription factors pAT 225/EGR1 and pAT 591/EGR2."*
> (From PMID 1946405 annotated by BioScope corpus annotators.)

In Example 2.6, the word *"thought"* is a speculation cue, and the part of the sentence that is underlined is the scope which is affected.

The term *hedging*, also referring to speculative expressions, was originally introduced by (Lakoff 1973).

2.4.2 Tasks and views on negation and hedging

Identification of negations and speculations in the literature has been widely explored by information extraction researchers (Hakenberg et al. 2009); (Morante et al. 2009); (Kilicoglu et al. 2009). We categorise different approaches based on their syntactic and semantic properties. The identification of negations and speculations can include either detecting the cue phrase and its scope or detecting the specific target (i.e. word, phrase, term, concept, or relation) under negation/speculation. Furthermore, some approaches aim at assigning polarity/modality (negation/speculation) at the sentence level. We also differentiate between approaches aiming at detecting affected concepts and those addressing the detection of affected events. A table summarising different tasks and prominent research can be seen on page 82 (Table 2.8).

Sentence polarity detection. Perhaps the simplest approach to negation and speculation detection is to detect the polarity and modality of a whole sentence, based, for example, on whether or not the sentence contains a speculative or negated fragment. Although the results of these sentence-level approaches are valuable in the coarse filtering of the relevant sentences, they seldom provide information on the individual events reported in the sentence, especially if several events and other facts are reported within a single sentence.

Medlock et al. (2007), for example, used a machine learning method to classify a sentence into speculative or non-speculative categories using lexical features automatically extracted from the training data. They applied this method on a set of full text biomedical documents and reported an F-score of 76% (equal precision and recall).

Shatkay et al. (2008) introduced a system performing multi-dimensional classification on a corpus of randomly selected sentences from full text articles, labelling every sentence for negation and speculation as well as three other qualitative contexts (focus, evidence, and trend). The classification was at the sentence-level and achieved an F-score of 71% on detecting speculation and an

F-score of 97% on detection negations. The authors have calculated the F-score values based on multi-class classification, with all correctly predicted affirmative instances (true negative predictions of the negation detection task) also contributing towards the F-score.

The Computational Natural Language Learning (CoNLL) shared task in 2010 (Farkas et al. 2010) involved the recognition of sentence level uncertainty. CoNLL best performing system for sentence classification on biological text (Tang et al. 2010) achieved P/R/F1 of 85%, 88%, 86% using a sequence labelling approach. The best performing system on Wikipedia articles involving uncertainty used a bag-of-word sentence classification and achieved P/R/F1 of 72%, 52%, 60% (Georgescul 2010).

Detecting scopes and targets. A number of approaches have been suggested for the detection of negated targets and scopes ((Chapman et al. 2001b); (Chapman et al. 2001a); (Szarvas et al. 2008); (Ballesteros et al. 2011)). The following manually annotated examples show some examples of what these methods aim to achieve.

> **Example 2.7.** *"Cotransfection studies with this cDNA **indicate that** it **can** repress basal promoter activity."*
> (From PMID 1946405 annotated by BioScope corpus annotators.)

Example 2.7 shows two speculation cues with their scopes. The double under line shows where the two scopes overlap. The sentence of Example 2.8 contains both a negation and a speculation cue. The double under line shows where the scopes of the two cues overlap.

> **Example 2.8.** *"Similarities between the effects of dexamethasone and RU486 **suggest** that the antiglucocorticoid properties of RU486 do **not** occur at the level of specific DNA binding."*

(From PMID 1944294 annotated by BioScope corpus annotators.)

Many of these approaches rely on task-specific, manually constructed rules of various complexities. (e.g. (Chapman et al. 2001a)) to patterns that rely on shallow parsing (e.g. (Leroy et al. 2003)). They differ in the size and composition of the list of negation cues, and in how these lists are utilised. Rule-based methods range from simple co-occurrence based approaches to more complex rules.

The approach which identifies proximate co-occurrences of negation cues and terms in the same sentence, is probably the simplest method for finding negations and provides a useful baseline method for comparison. NegEx (Chapman et al. 2001a), for example, uses two generic regular expressions that are triggered by phrases containing negation cue and target term such as:

<negation cue> * <target term>

<target term> * <negation cue>

where the asterisk (*) represents a string of up to five tokens. Target terms represent domain concepts (e.g. terms from the Unified Medical Language System (UMLS)). Given that NegEx was primarily developed for the clinical domain, the cue set comprises 272 clinically-specific negation cues, including those such as "*denial of*" or "*absence of*". Although simple, the proposed approach showed good results on clinical data (78% sensitivity (recall), 84% precision, and 94% specificity). Tolentino et al. (2006) show that using rules on just a very small set of only five negation cues (*no*, *neither/nor*, *ruled out*, *denies*, *without*) can still be reasonably successful in detecting negations in medical reports (F-score 91%).

Similarly, Negfinder (Mutalik et al. 2001) use hand-crafted rules and a list of 60 negation cues in order to detect negated UMLS terms. Their list of cues includes single-word cues such as *no, without, negative* and phrases such as "*no evidence of*", "*could not be currently identified*". They use simple

conjunctive and disjunctive phrases (e.g. *"and"* and *"or"*) to identify lists of concepts that are negated by a single cue. Therefore, the task is broken down into finding the scope of negation, and determining whether the terms in question are located in that scope. In order to achieve the scope of negation, Mutalik et al. use a method similar to parsing, but without parsing the complete structure of the sentence (Mutalik et al. 2001). They select "negation terminators" from the list of prepositions, conjunctions, personal pronouns, and relative pronouns, based on some rules. Finally, UMLS concepts, negation cues, negation terminators, and sentence terminators are located and negated concepts are identified by determining whether these concepts fall within the scope terminated by a negation terminator.

Negfinder is tested on a corpus of medical narratives (radiology reports) and report specificity/sensitivity of 92%/96%.

In addition to concepts that are explicitly negated by negation phrases, Patrick et al. (Patrick et al. 2007) further consider so-called pre-coordinated negative terms (i.e. concept that semantically indicate a negative situation, e.g. "headache"). These concepts have been collected from the SNOMED CT medical terminology.

Some of the methods rely on shallow parsing (e.g. (Leroy et al. 2003)) or various types of parse trees (e.g. (Sanchez 2007)). For example, (Huang et al. 2007) introduced a negation grammar that used regular expressions and dependency parse trees to identify negation cues and their scope in the sentence. They applied the rules to a set of radiology reports and reported a precision of 99% and a recall of 92%. Techniques developed for speculation identification follow similar approaches as for negation detection (Velldal 2011); (Morante et al. 2010a).

Not many efforts have been reported on using machine learning to detect patterns in sentences that contain negative expressions. Still, Morante et al. (2009),for example, used various classifiers (Memory-based Learners, Support Vector Machines, and Conditional Random Fields) to detect negation cues and

their scope. An extensive list of features included the token's stem and part-of-speech, as well as those of the neighbouring tokens. Separate classifiers were used for detecting negation cues and negation scopes. The method was applied to clinical text, biomedical abstracts, and biomedical papers with F-scores of 80%, 77%, and 68% respectively.

An extended version of this system (Morante et al. 2010a) was applied to the speculation detection task on the cue and scope level which was the second shared task of the Computational Natural Language Learning (CoNLL) in 2010 and achieved P/R/F1 of 60%, 55% and 57.3% on the biological data as the best performing system. The best performing system on the Wikipedia sentences (Tang et al. 2010) achieved 63%, 26%, 36%.

Özgür and Radev used machine learning (SVM) to detect speculation cues, using common features such as stem and part-of-speech tag, and some other features that we will briefly introduce here (Özgür et al. 2009). The authors used certain dependency relations such as clausal complement and auxiliary as binary features. They also used features indicating which position in the article the sentence has appeared in (e.g. title, abstract, etc.). Finally, they used features regarding word co-occurrence and the existence of negation cue in the sentence, as they hypothesised that it can play a role with certain speculation cues.

Agarwal et at. used a biological and medical annotated corpus to train several CRF models to detect negations and their scope in biomedical text (Agarwal et al. 2010). They detect cues and their scopes independently, and replace words with their part-of-speech. Their system, BioNOT, was applied on a large scale corpus of biomedical abstracts and full text articles and was tested on the biological and medical corpus BioScope with F-Score of 92%.

For more discussion on the detection of negation and speculation cues and scopes see (Morante et al. 2011).

Detecting negated and speculated events. While many of the systems

mentioned above focused on identification of negated terms, several approaches have recently been suggested for the extraction of negated events, particularly in the biomedical domain. For example, (Van Landeghem et al. 2008) used a rule-based approach based on token distances in sentence and lexical information in event triggers to detect negated molecular events. Kilicoglu et al. (2009), Hakenberg et al. (2009), and Sanchez (2007) used a number of heuristic rules concerning the type of the negation cue and the type of the dependency relation to detect negated molecular events described in text. For example, a rule can state that if the negation cue is *lack* or *absence*, then the trigger has to be in the prepositional phrase of the cue; on the other hand, if the cue is *unable* or *fail*, then the trigger has to be in the clausal complement of the cue (Kilicoglu et al. 2009). As expected, such approaches typically suffer from lower recall (32%).

(MacKinlay et al. 2009), on the other hand, used ML, assigning a vector of complex deep parse features (including syntactic predicates to capture negation scopes, conjunctions and semantically negated verbs) to every event trigger.

We have estimated that their system achieved an F-score of 36% on the same dataset as used in this paper (Sarafraz et al. 2010).

Task 3 of the BioNLP'09 challenge involved the identification of negations and speculations in biomedical abstracts. The evaluation is done on the performance of the whole pipeline, including event extraction stage and negation/speculation detection stage. The best performing team achieved recall/precision/F-score of 15.0/50.7/23.1 when applied their negation detection system to the automatically extracted events. Unfortunately we do not have access to the performance of the second stage alone, as the performance of the negation and speculation detection stage will inevitably be affected by less-than-perfect performance of the first stage (i.e. event identification). However, by knowing the performance of the whole pipeline and the performance of the event detection (first stage), we can estimate the performance of the

negation/speculation detection (second stage.)

With overall event detection sensitivity of 33% (Kilicoglu et al. 2009) on the test dataset and pipeline recall of 15%, we can estimate that had all events been correctly identified, the recall of their negation detection approach could have been three times higher, and reached 45%. With pipeline precision of around 50%, their projected F-score, again assuming perfect event identification, could have been in the region of 50%.

As part of the effort to add context to extracted events in (Sanchez 2007), negation and speculation information was extracted from sentences containing protein-protein interactions. Sanchez identifies a number of categories of negation and speculation patterns, and constructs heuristic rules mainly based on dependency parse of the sentence to determine whether a given interaction is negated or speculated. Examples of such rules can be seen in Table 2.7. A total of 7 cases for negation and 3 cases for speculation were categorised, and each case was addressed by up to a dozen rules.

Type of negation / speculation pattern	**Rule to detect negation**
Adverbial negation “not”	Trigger is a verb and is connected by a verb chain dependency to auxiliary verb Negation cue and subject depend on auxiliary verb Object depends on the trigger
Inability to interact “cannot”, “unable to”, “inability”	If trigger is postmodifier of “able” And “able” is complement of “to be” in negative form → Subject of “to be” and object of trigger verb are possibly negated
“No” and “lack of”	Trigger is a noun, And “no” is a dependent determiner Or trigger appears in the prepositional complement of “lack of”
Not “have” evidence	Trigger is a verb Trigger is the object of “have” Dependency distance between “not have” and trigger is no more than 4

Table 2.7: Examples of rules used by Sanchez to detect negations and speculations

The first three examples show heuristic rules used by Sanchez to detect negated events and the last example is used to detect speculative events. Similar rules have been derived for other types of negation and speculation patterns including “fail to”, “does not exist”, “no effect on”, “not detect”, etc.

To demonstrate one of the above rules, consider Example 2.9.

Example 2.9. “*The p46 isoform of JNL was not phosphorylated by ORF36.”*

The diagram in Figure 2.10 shows the partial dependency parse of this sentence, and demonstrates how the second rule in Table 2.7 applies, i.e. trigger is a verb and negation cue and subject depend on the auxiliary verb.

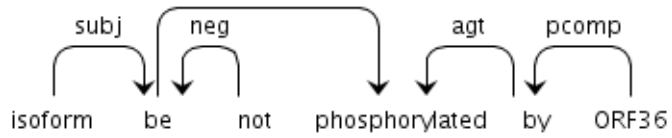


Figure 2.10: Dependency parse satisfying rules for negation

Partial dependency parse of the sentence "The p46 isoform of JNL was not phosphorylated by ORF36." The rule for the negation type with "not" requiring that negation cue and subject depend on auxiliary verb applies.

To evaluate these heuristics, Sanchez et al. selected 185 sentences from Journal of Biological Chemistry articles that contained negation words and protein names. Amongst these sentences, 90 contained one or more negated events and the other 95 sentences contained no negated events. The notion of "event" here is comparable to that of BioNLP'09 corpus, as it refers to molecular events between two proteins. However, the two cannot directly be mapped as the types and structures differ. They report precision/recall/F-score of 89%/67%/76% on data with gold annotated proteins, and 64%/61%/62% after automatically extracting proteins from the text using ABNER-UniProt.

These results cannot directly be compared with other research, as the choice of the evaluation data set makes it rather contrived. On the one hand, by deliberately selecting sentences that contain protein names and negation cues, the task of finding negated events becomes a more difficult task, as there are plenty of false clues in the negative instances. On the other hand, as the sentences have been picked to either do or do not have negated events, the task can be reduced to a sentence classification problem, which as we previously noted, is relatively less complicated. Moreover, there is no discussion on what proportion of the sentences without a negated event do contain an event. This, together with the performance of the automatic event extraction could affect

the results dramatically.

Tables 2.8 and 2.9 summarise the tasks, views, methods and approaches described here, in addition to a few other which we did not cover thoroughly.

This summary shows that the task of detecting negated and speculated events from complex text is considerably more difficult than detecting cue scopes, and therefore more advanced information extraction techniques are required in order to be able to detect negations and speculation more accurately and reliably.

Tool / research	Tasks	View	Approach		Corpus	Performance
			Machine learning	Rule-based		
Medlock et al	Speculation	Sentence polarity	Weakly supervised learning	No	Biomedical	F = 76%
Shatkay et al	Negation, speculation	Sentence polarity	SVM	No	Biomedical	F = 71%
Tang et al	Speculation	Sentence polarity	Yes	No	Biomedical	F = 86%
NegEx	Negation	Scopes and targets	No	Yes	Medical	F = 96%
Negfinder	Negation	Scopes and targets	No	Yes	Medical	F = 96%
Morante et al	Negation, speculation	Scopes and targets	Memory-based Learners, SVM, CRF	No	Biomedical	F = 57%
BioNOT	Negation	Scopes and targets	CRF	No	Medical, biomedical	F = 92%
Kilicoglu et al		Events	No	Yes	Biomedical	F = 43%
Sanchez et al	Negation	Events	No	Yes	Biomedical	F = 77%
MacKinlay et al	Negation, speculation	Events	Yes	No	Biomedical	F = 35% F = 30%
NegHunter (Gindl et al. 2008)	Negation	Targets	No	Yes	Medical (practice guidelines)	P 59% R 67%

Table 2.8: Summary of past efforts on negation and speculation detection

We conclude this section by briefly mentioning an effort in the detection of negations and speculations in a language other than English. Hagege (Hagege 2011) introduced a rule based method for finding negation in French clinical discharge summaries. They used a set of “negative seeds” which are phrases that indicate the absence of something. They focus on the detection of

seven classes of targets whose absence they are interested in, including viral disease, diagnosis, etc. They also use more than 100 nouns and verbs that indicate affirmation or negation, e.g. *existence* or *absence*. They hand craft rules using “negative seeds” and negation indicators to determine whether any of the mentions of the targets are negated.

Method	**Properties**	**Details**
Machine learning	Features	Dictionary
		Orthographical information about the token
		Lemma or stem of the token
		Part-of-speech (POS) tags
		Syntactic chunk information
		Dependency/constituency parsing
		Position in document or sentence
		Handling multiple cues together or independently
	Definition of the ML problem	Sequence labelling (e.g. CRF)
		Bag-of-word feature representation
		Classifying every token in the sentence to detect cue phrase
		Finding scopes: labelling tokens as inside/outside or begin/end
	The ML engine used	SVM
		CRF
		K-nearest neighbours
		Entropy Guided Transformation Learning
		Average perceptron
Rule-based systems	hand-crafted rules, Regular expressions	Surface distances
		POS tags
		Dependencies distances

Table 2.9: Summary of methodologies used in negation and speculation detection

Summary of the methodologies with the most common properties and features.

2.5 Extracting contrasts and contradictions from literature

As a mathematical logical concept, **contradiction** is abstract. Contradiction happens when two logically opposite statements are true simultaneously. In classical logic, the law of non-contradiction (NLC) is number two in Aristotle's three classic laws of thought.

Philosophers and logicians tend to agree that although it is possible to define contradiction as an abstract concept, no real contradiction can exist in the natural world. It is also generally believed that imagining contradiction, perhaps similar to imagining the 4th dimension, is impossibly difficult for the human brain, despite the fact that it is used as a basic tool in mathematics.

Contrast, on the other hand, is an expression used to describe two different concepts. Oxford English Dictionary defines contrast as:

"Comparison of objects of like kind whereby the difference of their qualities or characteristics is strikingly brought out; manifest exhibition of opposing qualities; an instance of this."

In this section we review interpretations of the concepts of contrast and contradiction within the domain of biomedical text mining.

2.5.1 BioContrasts

BioContrasts (Kim et al. 2006), is a database containing pairs of proteins that have appeared in contrasting phrases. It focuses on expressions such as "protein1 but not protein2" that have been extracted from the MEDLINE abstracts and contains about 800,000 non-normalised (~40,000 normalised) contrasting protein pairs. Both of the contrasting proteins appear in the same sentence. The contrast is explicitly expressed using phrases such as *but not* and in a context such as their interaction with a third protein.

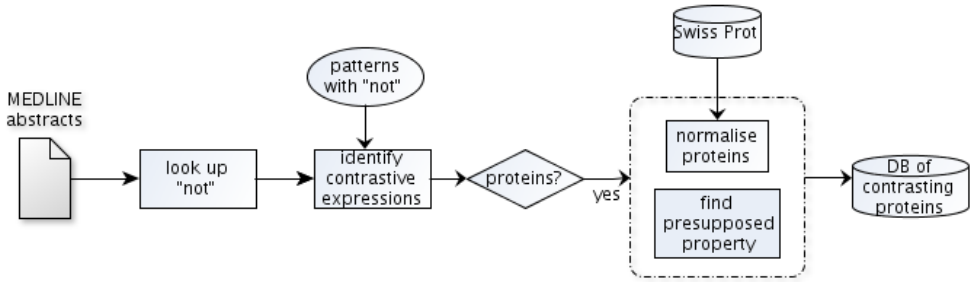


Figure 2.11: Work-fow of the BioContrasts system.

Figure 2.11 shows the work-flow of the BioContrasts information extraction system. It extracts contrasting protein pairs using a rule-based algorithm based on the keyword *not* and the grammatical parse of the sentence. Examples of the rules that are used in the identification of contrasting patterns can be seen in Table 2.10.

The rules that identify contrastive expressions are based on the identification of noun phrases and other grammatical entities such as verbs and prepositions. A POS tagger and a noun phrase identifier have specifically been developed for the task. The rules for contrasting expressions were also manually designed, presumably based on manual analysis of a training corpus of 166 abstracts.

Pattern type	**Grammatical pattern**
A but not B	NP but not NP V NP but not V NP V PREP but not PREP NP
not A but B	not NP but NP not V PREP but V PREP NP not V NP but V NP
A, not B	NP, not NP PREP NP, not PREP NP

Table 2.10: Examples of patterns used by BioContrasts

Example patterns to extract contrasting protein pairs used by BioContrasts

The two proteins reported as contrasting, contrast in a "presupposed property" which is assumed to be the verb of the sentence in question. This identifies the event type in which the two proteins differ, but only if this event is expressed via the verb of the sentence.

For an example that demonstrates this approach, consider the sentence from an abstract shown in Example 2.10.

> **Example 2.10.** *"In contrast, IFN-gamma priming did not affect the expression of p105 transcripts but enhanced the expression of p65 mRNA (2-fold)."*
> (From PMID 8641346)

The sentence is selected at the first stage because it contains the word *not*. After POS tagging and identification of noun phrases, we see that part of it matches the pattern "not V NP but V NP" from Table 2.10:

> *not* + *affect* (V) + *the expression of p105 transcripts* (NP) + *but* + *enhanced* (V) + *the expression of p65 mRNA (2-fold)* (NP)

Once the protein names have been identified, the system compares the verbs in order to determine the similarity between them and therefore determine whether the two phrases are contrastive. The authors use WordNet (Fellbaum 1998) and a hand-crafted list of similar biomedical verbs for this purpose. One of the limitations of this method is that it does not recognise interaction words that are not verbs as presupposed property. For example, the word "*expression*" in a phrase like "the expression of A but not B".

Finally, they insert the two protein names that satisfy the aforementioned criteria in the database. In the case of Example 2.10, *p105* and *p65* will be added as contrasting proteins.

They applied their system to a corpus of 2.5 million MEDLINE abstracts each containing the word "not". They extracted contrasting protein pairs with

normalised protein names using the above method, and selected 100 pairs for post-hoc evaluation. They reported a precision of 97%. As other post-hoc evaluations, the recall value was not reported, but it is expected to be lower than their previous, less strict system with 61% recall (Kim et al. 2006).

Another limitation of this approach is that no context is given beyond the fact that the two proteins are reported to differ in a given interaction. The interaction type is not normalised and no more information about the event itself is reported. In addition, with very strict hand-crafted rules, the recall is expected to be relatively low, as contrasting proteins and interacting proteins in general can appear in many more patterns than investigated in this study.

Finally, the contrast is defined only between the pair of entities. To move towards contrasts between smallest pieces of self-contained information such as events, the methods would need to expand to properties of biomedical events other that participating proteins.

2.5.2 An approach to contradicting events

(Sanchez 2007) has introduced the concept of contradiction in protein-protein interactions and has categorised them into explicit and implicit contradictions. According to her definitions, explicit contradiction refers to the situation when an author reports results and mentions that they contradict or are different from previous findings. An implicit contradiction, on the other hand, are two statements possibly in different documents, one reporting a protein-protein interaction event affirmatively, and the other reporting it either negatively or speculatively.

Sentences that contain both “contradiction” phrases and “finding” phrases are identified and used for training and evaluation purposes. According to their definitions, finding phrases concern the phrases that report some finding in the literature. Contradiction phrases are those that express conflict, presumably with previous views. This is based on the hypothesis that these sentences are associated with explicit contrasts.

Example 2.11. *"An affinity of RRM3 for poly(U) appears to contradict previous reports of poly(A) binding by RRM3."*

In Example 2.11, the word *"contradict"* is a contradiction phrase, and the word *"reports"* is a finding phrase. See an extensive list of these phrases in Table 2.11.

Contradiction words	Finding words
contradict, contradiction, contradictory, conflict, negate, negation, disagree, disagreement, refute, refutation, differ, dissent, discrepancy, inconsistency, inconsistent, contrast, controversy	observation, report, notion, evidence, finding, research, hypothesis, knowledge, interpretation, conclusion, model, data, fact, study, inform, document, work, proposal, result, view, assertion, assay

Table 2.11: List of contradiction and finding phrases used by Sanchez.

Note that in the original thesis, the heading of the left-hand column reads "CONTRAST WORDS", although everywhere else it is referred to as contradiction words or phrases.

To detect **explicit contradictions**, Sanchez looked at the dependency parse of the sentences containing both finding and contradiction phrases, and hypothesised that there may be an explicit contrasting event in the sentence if there is a dependency path between the two phrases that does not pass through the root and is at most of length 3, and a dependency path of maximum length 10 between the contradiction word and the interaction trigger.

The sentence in Example 2.12 from the author will demonstrate this hypothesis.

Example 2.12. *"Although the activation of AMPK by insulin would contradict previous observations (28,29), AMPK activation is known to*

accelerate glucose uptake and utilization in the heart."

A simplified dependency tree for Example 2.12 is shown in Figure 2.12. All the criteria for an explicit contradiction as defined by Sanchez are satisfied:

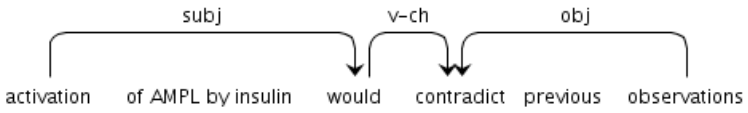


Figure 2.12: Simplified partial example dependency tree

Part of the dependency tree of the example sentence, specifically the part that represents the application of the criteria: "activation of AMPL by insulin would contradict previous observations" In this tree, 'v-ch' represents a verb chain dependency.)

1. There is an event expressed in the sentence, with the trigger "*activation*", possibly involving entities "*AMPK*" and "*insulin*".
2. The dependency path between the finding word "observation" and the contradiction word "*contradict*" is 1 (less than 3).
3. The dependency path between the contradiction word "*contradict*" and the interaction trigger "*activation*" is 2 (less than 10).

Sanchez evaluated her method on 122 sentences containing contradiction phrases derived from 500 Journal of Biological Chemistry articles, manually annotated for explicit contradictions. Amongst these sentences, 61 contained explicit contradictions. Their method of detecting explicit contradiction from sentences containing contradiction phrases achieved recall/precision/F-score of 36%/92%/52%.

Implicit contradictions. To find implicit contradictions, Sanchez introduced a semantic representation of events in order to introduce rules that

make two different events contradictory. The semantic representation of an event involves the following attributes:

> *event type, trigger, participating proteins, polarity, direction, certainty, manner, organism* and *anatomical location*.

Once the events are extracted, attributes of the semantic representation need to be determined. Two of these attributes, namely polarity (negation) and certainty (speculation) have already been discussed in the previous chapters.

To determine the *direction* of an event, Sanchez categorises all possible interaction words (around 50 of them, not including inflected forms) into 13 categories. She then uses a look-up table assigning positive, negative, or neutral directionality to each class. For example, triggers belonging to the *"attach"* class (such as *"bind"* and *"complex"*) have a positive direction, whereas those belonging to the *"inactivate"* class (such as *"block"*, *"down regulate"*, and *"suppress"*) will have a negative direction. Trigger words such as *"translocate"* or *"affect"* have neutral direction.

The *manner* attribute is the adjective or adverb that affects the trigger. In addition to speculation extracted previously, manner words are also used to infer speculation (e.g. *"there is a potential interaction"*).

To demonstrate these attributes see Table 2.12 for an original example and its semantic representation.

Semantic class	inactivate
trigger	inhibit
Protein 1	ATP
Protein 2	AMPK
Auxiliary molecule	-
Polarity	Positive
Direction	Negative
Certainty	-
Manner	Neutral
Organism	-
Location	-

Table 2.12: Semantic representation of an event according to Sanchez's definition

Representation of the event expressed by sentence "ATP inhibition of adenosine monophosphate-activated protein kinase."

They combined the attributes *direction*, *manner*, and *polarity* to assign a single number to every event. The numbers are merely labels assigned to different combinations of these three attributes and do not represent their numerical value. For example, "weak positive direction" is assigned state 2, and "negative polarity and strong or neutral negative direction" is assigned state 9. They introduced around 15 different states to assign to different combinations of attributes of an event.

Subsequently, they manually constructed a decision table with pairs of states that constitute a contradiction and pairs of states that do not. From the table, we learn that there is a contradiction between states 3 and 7, i.e. "weak neutral direction" and "strong neutral direction", but there is no contradiction between states 4 and 7, i.e. "neutral direction" and "strong neutral direction".

Another example of a contradicting pair is states 3 and 11 , i.e. "weak neutral direction" and "negative polarity and weak neutral direction". There are 34 pairs of states that define contradictory or non-contradictory states, and the rest are defined as undecidable.

To demonstrate how this method works, we discuss their original example here:

Example 2.13.

Document A *"Cells treated with hyperosmolar stress, UV-C, IR, or a cell-permeable form of ceramide, C2 ceramide, rapidly down-regulated PI(3)K activity to 10%-30% of the activity found in serum-stimulated control cells"*

Document B *"And fourth, C2-ceramide did not affect the amount of PI 3-kinase activity in anti-IRS-1 precipitates."*

We show the semantic representations of these two events in Table 2.13. As we can see, the states of the two events in the sentences, in Example 2.13 as determined by their direction, manner (manner degree), and polarity, are 9 and 5 respectively, which are "negative polarity and strong (or neutral) negative direction" and "negative polarity and neutral direction". The pair (9,5) do not appear in the look-up table used to define contradiction pairs, meaning that we cannot say anything about whether the two events are contradictory or not.

In none of the publications describing this work, the authors have included any other example that demonstrates how this method detects contradicting or non-contradicting events.

Attribute	Document A	Document B
Event type	Inactivate	Cause
Protein 1	C2-ceramide	C2-ceramide
Protein 2	PI-3K	PI-3K
Trigger	down-regulate	affect
Polarity	positive	negative
Direction	negative	neutral
Manner	rapidly	-
Manner polarity	neutral	neutral
Manner degree	neutral	neutral
state	**9**	**5**

Table 2.13: Semantic representation of two events in the example sentences

The 'states' of the events are determined based on a decision table as well as the values of direction, manner (manner degree), and polarity. Original example and semantic representation by Sanchez.

They applied their method on automatically extracted events as well as gold annotated events. They report recall/precision of 19%/60% on the automatically extracted events and recall/precision of 62%/50% and 80%/53% on gold annotated events against two different gold standard annotations.

They also evaluated their system using inter-annotator agreement measure *kappa*. The agreement between biologists and their system was 0.39 and the agreement between non-biologists and the system was 0.21. See Section 2.7.1 for the definition of this measure.

This was one of the earliest computational approaches to contradictions, and despite the limited availability of annotated data, provided an understanding of the phenomenon. However, there are a number of limitations in this work that we shall point out here.

- No generic definition of contradiction was given. The definition was by instance, and was not exhaustive.
- The decision table to assign a state to an event does not match the list of the state descriptions introduced earlier, and seems limited, ad-hoc, and anecdotal.

- The numerical values assigned to different events have no numerical significance and are only used for coding different combinations of the four attributes contributing to the value. The decision table is incomplete and does not contain every possible combination, including those of some of the examples discussed in the thesis.
- The table does not cover all possible combinations of these attributes. Some of the states are defined in a way that their instances can overlap, for example “neutral direction” (state 4) is a subset of “neutral direction or degree” (state 0).
- No discussion on multi-event sentences are included. It would appear that despite the extensive sub-sentence rules and analysis, the problem is only approached on the sentence level.
- They have evaluated their method on the data specifically selected to pass the bag-of-words test, some of which would have resulted in a false positive in a bag-of-words approach. Although this might result in a tougher evaluation set-up, the evaluation is still not on “natural” data, and is difficult to expand to a given set of documents.
- The method is composed of a number of very specifically tailored rules and extensive look-up tables. Although it is possible to trace every case and how it is classified following the rules, it is not expandable, generic, or applicable to similar problems.
- Despite having access to organism and anatomical location information from the event extraction stage, they have not used these attributes in the characterisation of contradictions.
- They have not differentiated between the contrast expressed about a biological concepts, or one expressed in relation to a finding. For example, the sentence of Example 2.14, taken from the corpus used in the study, contains the former type of contrast.

  Example *2.14.* *“Moreover, in **contrast** with PMA, the effect of thrombin on the tyrosine phosphorylation of SH-PTP1 was hardly*

affected by GF109203X, a specific protein kinase C (PKC) inhibitor. ”

In addition, it is worth noting that protein name normalisation was done manually, and no large-scale application or evaluation of the method was performed beyond the 500 JBC articles.

2.6 Resources

In this section we introduce the resources available for the biomedical text mining particularly related to this research. We introduce the publicly available annotated corpora which is a requirement for many information extraction tasks, and specifically for machine learning approaches.

In recent years, resources that provide manually annotated data are increasing in availability. GENIA corpus (Ohta et al. 2002) was perhaps the first, and the most widely used annotated corpora of biomedical abstracts. It contains 2000 MEDLINE abstracts selected from the search results for the terms *“human”*, *“blood cells”*, and *“transcription factors”*.

The abstracts are annotated for a range of linguistic and biological information. The annotations include POS tags, shallow parses, and co-reference annotations. They also include term annotations for entities from 31 different semantic classes including proteins, DNAs, RNAs, etc. (Kim et al. 2004).

Several types of molecular events are annotated in the GENIA corpus with their types, themes, and causes (Kim et al. 2008). Other annotations include disease-gene association, cellular localisation, and pathways.

The GENIA corpus annotators have assigned “assertion” and “uncertainty” attributes to every event. Assertion indicates whether the event is negated or affirmative and the possible values for this binary attribute are “exist” and “non-exist”. Uncertainty indicates the level of speculation in the reported event. However, unlike assertion, uncertainty is not a binary attribute, with the three possible values of “certain”, “probable”, and “doubtful”.

Challenges such as BioCreative and BioNLP also make valuable high

quality manual annotations publicly available. As a result, research groups publish their results on these data sets and sometimes even release their tools, making it possible to compare approaches and results within the same framework.

A manually annotated corpus was released for the training and testing of the BioNLP'09 Shared Task. The entity annotations were limited to genes and gene products (proteins), manually selected from the following entity tags in the GENIA corpus: *protein molecule, protein complex, DNA domain or region,* and *RNA molecule.* Since these groups contain entities such as protein complexes (e.g. *NF kappa B*), genomic regions that are not genes (e.g. *third intron* or *gene*), or other biological entities, the organisers have removed them in the construction of the BioNLP'09 corpus.

The events were selected from the subset of the GENIA corpus (Ohta et al. 2002) that can be considered as bio-events and involve gene and protein molecules. The following event types from the GENIA corpus were included in the BioNLP'09 corpus: *Positive regulation, Negative regulation, Regulation, Gene expression, Binding, Transcription, Localization, Protein catabolism, Protein amino acid phosphorylation,* and *Protein amino acid dephosphorylation.* The last two event types, *Protein amino acid phosphorylation,* and *Protein amino acid dephosphorylation,* were merged into a single class, *Phosphorylation.*[15] Events whose participants were not genes or proteins were excluded. In addition, a number of new events were also added manually by the curators of the BioNLP'09 corpus. A summary of the number of events in each category can be seen in Table 2.14.

15 For a biological definition of these event types please see Appendix A

Event type	Number of events in training data	Number of events in development data
Gene expression	1738	356
Localization	265	53
Transcription	576	82
Protein catabolism	110	21
Phosphorylation	169	47
Binding	887	249
Regulation	961	173
Positive regulation	2847	618
Negative regulation	1062	196
Total	**8615**	**1795**

Table 2.14: The distribution of the different event types in the BioNLP'09 corpus.

The data came in three data sets. The *training* and *development* data sets were available to the public together with gold annotations for entities and events. The *test* dataset was used to evaluate the challenge participants and was only publicly available with gold annotations for entities. Table 2.15 shows the composition of the first two data sets.

	Training data	Development data
Abstracts	800	150
Sentences	7449	1450
Words	176146	33937
Entities	9300	2080
Events	8597	1809
Negated events	615	107
Speculated events	455	95

Table 2.15: The composition of the events in the BioNLP'09 data.

The number of words, events, and other statistics in the training and development data sets.

In 2011, the BioNLP'11 shared task was organised with tasks and data along the same lines with BioNLP'09, in that it provided a representation of bio-molecular events and called for extracting relations from them. As this thesis was being written, new research groups are releasing tools that address the challenges posed by this commonly available data.

Bio Information Extraction Resource (BioInfer) is another manually annotated corpus available for training and development of biomedical information extraction efforts. The corpus contains 1,100 sentences taken from biomedical abstracts, and are manually annotated for named entities, relationships between the named entities (e.g. equality, membership, anaphora, causal, etc.), and syntactic dependencies (Pyysalo et al. 2007).

A manually annotated corpus that provides annotations for negations and speculations of biomedical and clinical text is BioScope (Szarvas et al. 2008). It contains all the abstracts in the GENIA corpus, five full text articles from FlyBase (Tweedie et al. 2009), and a corpus of radiology reports used in other challenges. The texts were annotated for negation and speculation cues and their linguistic scope, i.e. the part of sentence that is affected by those scopes and has become negated or speculated.

The BioScope corpus has been used in a number of attempts to automatically detect negations and speculations. (Morante et al. 2009) used the BioScope corpus in a scope detector system which uses supervised sequence labelling.

Resource	Type	Size	Annotations
GENIA corpus	Biomedical abstracts	1000 documents	substances and the biological locations involved in reactions of proteins, based on the GENIA ontology
BioNLP'09	Biomedical abstracts	950 documents	Named entities, molecular events, localisation, negation, speculation.
BioInfer	Biomedical abstracts	1100 sentences	Relationships, named entities, syntactic dependencies
BioScope	Medical free texts / biological full papers / biological abstracts	20,000 sentences	negations and speculations and their linguistic scopes
CoNLL'10	Biomedical full-text	15 documents	Speculation cues and their linguistic scopes (to be added to the BioScope corpus)

Table 2.16: Summary of corpora related to this research

The Computational Natural Language Learning shared task in 2010 (CoNLL 2010) focused on the identification of sentences in biological abstracts containing uncertain information (Farkas et al. 2010). It aimed at detecting speculative sentences in two tasks. One task was defined as a binary classification problem on the sentence level, distinguishing *factual* from *uncertain* sentences. A second task was defined as detection of speculation cues and their scope within the sentences. The CoNLL 2010 challenge also used the BioScope corpus as one of the two corpora included in the challenge (the other corpus was about Wikipedia weasels.)

2.7 Evaluation in text mining

2.7.1 Evaluation methods

Many areas of computer science suffer to various degrees from the lack of standard, commonly accepted, thorough and reliable bench marks and test datasets that correspond with the up to date real world problems and IE and IR are no exceptions. However, there have been several "shared assessments" or

"challenges" in biological text mining that have been among the most influential assessments in the field (Cohen et al. 2005) a selection of which were introduced in the previous sections. These are among the leading references to determine the state of the art in various biomedical IE tasks and to provide resources to be used as gold standard data in those areas.

Evaluation of IE and IR systems is important in order to compare the performance of the existent systems, measuring the progress of the field over time, and creating a shared infrastructure to support research. Previous examples have shown that once objective common evaluations become available, there is real eagerness from the research communities to participate in new challenges and improve the solutions to existing problems (Cohen et al. 2005).

In this section after introducing the baseline measure as the minimum requirement for any system to be worthwhile of studying, we describe precision, recall, and F-score measures that are commonly used with or without other specifically developed measures by the above groups. Finally we will briefly introduce other types of measures for IE and IR.

Baseline measure

The baseline measure is the performance of a simple but not necessarily trivial method against which other innovative methods are evaluated. Random assignment is often used as a baseline measure for comparing with more sophisticated information extraction methods.

An information extraction task can be simplified as finding items with certain properties in a pool of items. Assume, for example, that we have a document with a certain number of tokens, some of which are names of species and the rest are other types of words. The task of finding those species names amongst all the tokens is an NER task.

Reporting every token as a positive result, i.e. assigning them to the class of species names will result in 100% recall, as we are obviously finding every

species name correctly. However, the precision would just be equal to the percentage of the species names (positive instances) in the pool.

Categorical assignment of all instances to one class is the simplest baseline measure and is specially useful when the sizes of positive and negative classes are very disproportionate and we are interested in the number of correct classifications in all classes.

An unsophisticated classifier with little discriminative functionality could randomly assign tokens to positive and negative classes. The precision, recall, and F-score can then be computed, given that we have enough information about the composition of the data, specifically, about the percentage of positive and negative instances in the dataset.

In the case when we can safely assume more specific characteristics about the dataset, we can improve the baseline measure to reflect the information we already have about the distribution of the different types of the instances. In the example above, an unsophisticated baseline classification is to randomly label half the tokens as species names. But given the extra information that most tokens in text are articles, adjectives, and other types of linguistic “fillers”, and that terms or named entities are relatively rare, it would be logical to improve the precision of the baseline method to somehow assign a smaller proportion to the category of species names. Any system of species name recognition will therefore have to perform better than this baseline random classifier.

Common evaluation measures

Precision, recall, and F-score evaluate the performance of a system by comparing its output with a gold-standard. Recall measures how much the system has covered the desired output, i.e. how much of the relevant information it has retrieved.

Recall is defined as the number of correct answers given (e.g. relevant documents retrieved by the system, species names correctly recognised)

divided by the total number of correct answers in the dataset or relevant documents in the pool:

$$Recall = \frac{TP}{TP+FN}$$

where TP and FN stand for the number of true positive and false negative answers given by the system respectively. On the other hand, precision is defined as the number of correct answers divided by all the instances retrieved by the system:

$$Precision = \frac{TP}{TP+FP}$$

where FP stands for the number of false positive answers given by the system.

For any system, there is usually a trade-off between the two above measures. The more specific search criteria are and the more narrowly we search for the results to increase the precision, the more likely it is to miss some of the off-centre positive results, and therefore decreasing the recall. To be able to reflect both precision and recall in a single measure, the harmonic mean of the two measures is widely used as an evaluation measure and is referred to as the F-score:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

If, depending on application, we intend to assign β times as much importance to recall as precision,we would use the general formula:

$$F_\beta = (1+\beta^2) \times \frac{Precision \times Recall}{\beta^2 Precition + Recall}$$

Evaluation measures sensitivity and specificity can be used to measure the performance of a binary classification task. Unlike precision/recall where we are only interested in the retrieved information from a pool of data, and only evaluate the quality of the information extracted from that pool, here we are interested in the quality of the discrimination between the two groups

without necessarily preferring one to the other. Sensitivity is defined similarly to recall, whereas specificity is defined as

$$Specificity = \frac{TN}{TN+FP}$$

Specificity measures the proportion of negative instances which are correctly identified.

Cohen's kappa coefficient is a statistical measure of inter-annotator agreement. It tests whether the agreement between several annotators exceeds that expected by chance. Kappa is defined as

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where *Pr(a)* is the relative observed agreement amongst annotators (i.e. the proportion of time that the annotators actually agree) and *Pr(e)* is the hypothetical probability of chance agreement (i.e. the proportion of time they would have agreed if they were guessing based on chance alone.)

2.7.2 Inter-annotator agreement

Inter-annotator agreement studies have shown that it is not uncommon for human annotators to disagree on whether an event is negated or speculative. (Vincze et al. 2011) have shown that after trying to map the negation and speculation annotations between two manually annotated corpora, GENIA Event corpus and BioScope corpus, the agreement rate between is no more than 48%.

The results of the mapping between the two corpora as reported by (Vincze et al. 2011) are shown in Table 2.17.

	Agreement	**BioScope + / GENIA -**	**BioScope - / GENIA +**
Negation	1554	1484	569
Speculation	1295	3761	180

Table 2.17: Inter-annotator agreement between GENIA and BioScope corpora

Inter-annotator agreement after mapping the negation and speculation annotations between GENIA event and BioScope corpora.

Sanchez also compared the annotations by several biologist and non-biologist annotators. They measured the agreement in finding what they define as “implicit contradictions”, introduced in the previous section.

The agreement amongst biologists was quite low (kappa = 0.38), which makes their system conceivably pass as an expert by agreeing with biologists more often than the biologists do with each other (kappa = 0.39).

But the agreement among non-biologists was predictably lower than that of biologists (kappa = 0.22), and the agreement between their system and non-biologists was even lower (kappa = 0.21).

2.8 Conclusion

In this chapter we briefly introduced the general area of text mining in the biological and biomedical domain and discussed some of the main challenges in this area. In particular, since the detection of conflicting statements relies on the recognition of contextualised event information including polarity detection, we critically reviewed the background research in relation extraction, contextualisation of event information including the extraction of negations and speculations, and the previous approaches to detecting various forms of contradictions from the literature.

We showed that the sentence polarity approach to negations and speculations is too rough, as several pieces of information is usually conveyed within a single sentence, and not all are negated and speculated simultaneously.

We also discussed the scope-based approaches to negation and speculation detection and showed that although these approaches have relatively high performance, they do not directly help enrich the extracted information since even when a negation or speculation scope is correctly identified, we still cannot directly infer whether or not an event statement which partly falls within this scope is affected. Consider, for example, phrases in which participants are negated such as *“SLP-76”* in the sentence *“In contrast, Grb2 can be coimmunoprecipitated with Sos1 and Sos2 but not with*

SLP-76."

We noted that the methods discussed in this chapter mainly focus on finding negated triggers in order to detect negated events.

We observed that molecular events between genes and proteins are meaningful pieces of information that are commonly described in text, and they can be expressed in a structured form. However, the performance of the current negation and speculation detection systems on the event level are not nearly as good as other approaches to negation detection (sentence polarity and cue/scope detection).

Although event data as used by event extraction systems contain rich semantic information regarding event types and participant types, this information has not been exploited to improve the results of event negation and speculation detection.

Syntactic properties have always been amongst the important features used in various information extraction tasks. However, although command relations were introduced by linguists decades ago and used commonly in theoretical linguistics, to the best of our knowledge they have not previously been exploited computationally for identification of negation and speculation.

No large-scale analysis of the negation and speculation extraction has been reported, and none of the best performing approaches have made their system publicly available.[16]

All the negation and speculation detection systems we are aware of have reported their performance as part of a text mining pipeline, and therefore an evaluation of the stand-alone system is not always reported.

A number of researchers have explored contradictions and contrasts in the biomedical domain. However, no general and comprehensive method has so far been proposed to detect explicit and implicit conflicting facts from the literature which could potentially lead to knowledge discovery and data

16 Only recently, as this thesis was in its final stages, a few recent attempts were made at large-scale extraction of biomedical events from all the available literature, some of which contain negation and speculation information.

consolidation.

In this thesis we aim to address these issues by taking the following steps:

1. Effectively identify biological events and relations among entities with their context;
2. Design and implement a system that will be able to automatically recognise negated and speculated facts in text, specifically in the sub-corpora of the GENIA event corpus interactions;
3. Develop a representation model for establishing relations between different biological events, including relations concerning conflicts. This involves semantically representing a biological event.
4. Design and implement a system that will detect conflicting facts from a database of extracted facts;
5. Evaluate the proposed methodology through a case study on biomedical events;
6. Apply the method on the entire publicly available biomedical literature;
7. Provide the tools and data to the biomedical and text mining research communities, including the contextualised events and the conflicts between them.

Chapter 3

Molecular event extraction and contextualisation

In this chapter we introduce methodologies to extract and contextualise molecular events from large textual corpora of biomedical literature. The work-flow in Figure 3.1 shows an overview of the tasks required to achieve the aims of this research. The boxes highlighted in darker blue represent tasks for which we use previously existing tools with some modifications, or implement existing methodologies. These tools and methods have been introduced in Chapter 2.

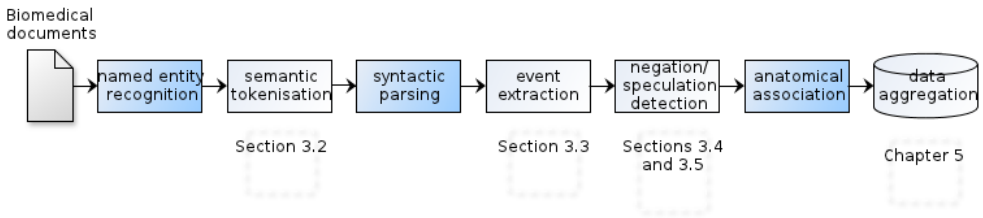


Figure 3.1: An overview of the event extraction pipeline.

The shadowed boxes are tasks for which existing methods have been utilised. The lighter boxes are tasks for which we developed a method which will be described in more detail in the corresponding sections.

The novel tools and methodologies that were created to address some of the challenges of this research will be discussed in this chapter. We start by defining the terms and concepts used in this thesis in section 3.1. In Section 3.2 we introduce semantic tokenisation, a method to reduce parser errors. Section 3.3 describes a hybrid machine learning and rule-based method developed to extract molecular events. In Section 3.4 we describe our methods to detect negated events. Section 3.5 expands the negated detection method to the task of speculation detection of events.

3.1 Definition of terms and concepts

3.1.1 Events and their context

In this thesis we focus on bio-molecular events as described in BioNLP'09. Nine event types are considered, namely, *Gene expression*, *Transcription*, *Localization*, *Protein catabolism*, *Phosphorylation*, *Binding*, *Positive regulation*, *Negative regulation*, and *Regulation*. The first five of these event types have only one participant (theme) and we refer to them as class I events. *Binding* events can have one or more themes (class II events). *Regulation* events have a cause as well as a theme, and can be nested with other events acting as a participant, and are referred to as class II events.

An event is minimally represented with four features: event type, the cause of the event, the target (theme) of the event, and the lexical expression (trigger) that is used to describe the event in text. However, placing the event in a wider context is important from a biological perspective.

An expert biomedical scientist, reading the sentence *"p53 is expressed in lung"* in an article, would understand more from it than it appears to convey. She would use her expert background knowledge of the field as well as other facts stated in other parts of the document and maybe other documents, to put that statement into context and achieve what we think of as an understanding of its meaning.

In an ideal automated information extraction task, we would hope to achieve similar levels of understanding to a human expert by combining vast amounts of background knowledge and mining as rich of a context as possible. This is an ambitious goal that more recent systems such as IBM's Watson (Ferrucci et al. 2010) and various Google services are aiming to achieve.

We distinguish between implicit and explicit statements. We aim to extract only information that is explicitly stated in the natural language text. The Informatics for Integrating Biology and the Bedside (i2b2) challenge in 2008 (Uzuner 2008) divided the information extraction task into two sub-tasks:

extracting information that is explicitly stated in text and that which can be inferred from the context. The distinction has also been recognised by the GENIA corpus, where the annotation guidelines require the annotators to only annotate information that is explicitly stated, and not use their own knowledge to infer from the text information that is only implied (Ohta et al. 2007).

As discussed in more detail in Chapter 2, (Sanchez 2007) also distinguishes between two concepts that she refers to as implicit and explicit contradictions. However, by these she means contradictions that are explicitly talked about in text and those that are found in different documents (see examples on page 92). Therefore, even her implicit contradictions are still what we consider explicit as they do not take into consideration anything that is not explicitly stated in text.

In biomedical text mining and information extraction, it remains an open problem to extract all of what the authors intend to communicate from the text itself. This is because these texts are typically very rich in information, often with many shades and nuances in meaning. Experiments vary in a great deal of detail; species, anatomical locations, experimental conditions such as temperature, and duration of experiments are only a few examples. Findings are reported with various levels of certainty, and are discussed thoroughly in comparison with the previous findings. It is very common that the meaning of whole sentences and paragraphs depend on the not-so-immediate context.

We primarily look at statements at the sentence level. Apart from a few exceptions in the anatomical NER module that will be discussed later, all the information extraction tasks are performed on the sentence level. This means that information stated elsewhere in the text is not taken into consideration when extracting facts and relations from a given sentence, and only what is explicitly stated in an isolated sentence will be considered.

According to this convention, if a sentence states that "*p53 is expressed in lung*" but it is evident from the rest of the document that it is in vivo lung of newborn rats subject to a certain medical procedure, we are in a situation where

we need to define precisely what we aim to extract. We do not aim to extract the intra-sentence context, namely “in vivo”, “new born”, “rat”, and the procedure. We limit our goal to extracting only the sentence-level information and therefore the only context we are concerned with is “lung”.

3.1.2 Event negations and speculations

Negated events. A negated event refers to an event that is reported in text as not happening. We treat negations as an attribute of the event.

According to this definition, several events can be expressed within a sentence, and not all are necessarily negated (or affirmative) simultaneously. The scopes of negation and speculation cues may vary or overlap. Moreover, the parts of the sentence stating a single fact are not always connected and independent. Therefore, sentence-based or scope-based definitions introduced in Section 2.4.2 are not suitable for understanding which facts are actually reported negatively or speculatively. This will also allow us to treat events as abstract concepts whose expression in text is not necessarily with a self-contained phrase or sub-string.

> **Example 3.1.** *“However, while HUVECs contained endothelial NOS protein, no inducible NOS was detected in either tolerant or nontolerant cells.”*
> (PMID 9915779, annotated by the BioNLP’09 corpus curators.)

In Example 3.1, the authors write that *“no inducible NOS was detected”*, effectively describing the lack of NOS expression in certain cells. In the same sentence, the word “contained” is referring to the expression of NOS in a different location. Therefore, the same sentence describes two NOS expression events, one affirmative (i.e. expression of NOS in HUVEC) and the other negated (induction of NOS in tolerant or nontolerant cells).

Speculated events. Speculations (hedges) are defined similarly as extra context on molecular events described in text. Sometimes the authors do not express with absolute certainty whether an event has happened or not. Rather, they speculate the existence of the event. This speculation may be in the form of expressing a hypothesis that they are later testing, or merely lack of enough evidence. Hedges classify events into "speculated" (or "un-asserted") and "asserted" categories (see Section 2.4), depending on how they have been described by the authors.

Example 3.2 shows a sentence which speculates the positive regulation (trigger: *"participate"*) of the regulation (trigger: *"transcriptional regulation"*) of *"IL1-beta"*.

> **Example 3.2.** *"These observations suggest that a so-far-unrecognized SP-1 site in the human IL-1beta promoter may participate in the transcriptional regulation of this gene in keratinocytes."*
> (PMID 8977297, annotated by the BioNLP'09 corpus curators.)

We also consider sentences like Example 3.3 as speculation.

> **Example 3.3.** *"Tumor necrosis factor alpha (TNF-alpha) mRNA production was analyzed by polymerase chain reaction amplification in monocytic U937 cells and in a chronically HIV infected U937 cell line (U9-IIIB)."*
> (PMID 2204723, annotated by the BioNLP'09 corpus curators.)

In Example 3.3, the authors talk about the "*production*" of "*Tumor necrosis factor alpha (TNF-alpha)*". However, they are not asserting whether the production did or did not happen. Rather, it was "*analyzed*", which suggests we can not infer the outcome of this analysis only from this sentence.

In Example 3.4, the authors are declaring that a certain molecular

processes *"have not been studied"*, and therefore, although the interaction is described in detail, neither its existence nor its absence can be inferred.

> **Example 3.4.** *"However, monocyte interactions with activated endothelium in shear flow following gene transfer of the NF-kappaB inhibitor IkappaB-alpha have not been studied."*
> (PMID 10339475, annotated by the BioNLP'09 corpus curators.)

3.1.3 Event representation

In this research we extend the template-based approach shared by BioNLP'09 and BioNLP'11 to defining biomedical molecular events. This model formally represents a molecular event with its attributes as a set of key-value pairs. There are restrictions on the valid values for each key. For example, participants of a relation can only be entities or other events. The restrictions on the values of some keys may vary depending on the values of other keys. The representational model is described in detail in Section 5.2.2.

We introduce two levels of event representation. On the semantic level, we identify every event that is biologically **distinct** by the following features:

1. its molecular type (any of g*ene expression*, *transcription*, *localization*, *protein catabolism*, *phosphorylation*, *binding*, *regulation*, *positive regulation*, or *negative regulation*);
2. the unique (database) identifiers of its cause and theme;
3. the unique identifier of the anatomical entity associated with the event (including the species that anatomical entity belongs to);
4. its polarity (negated or affirmative);
5. its certainty (speculated or asserted).

The representational model is depicted in Figure 3.2.

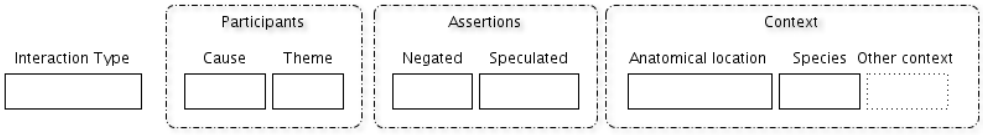


Figure 3.2: The event representational model.

Not every event has all the above properties. For example, only regulatory events could have a cause, and not all entities can be normalised by linking to database entities. In such cases, we use the best approximation possible. We treat the 'cause' field optional in the unique identification of an event. We would consider the word form of the named entities (gene, protein, or anatomical entity mentions) if they cannot be normalised into their database entries.

For example, in Example 3.5, the activation (i.e. positive regulation) of *StatG* by *G-CSF* in *myeloid cells* have been described.

> **Example 3.5.** *"We previously demonstrated that G-CSF activated a distinct Stat3-like protein in immature and mature normal myeloid cells, StatG."*
> (From PMID 9823774 annotated by BioNLP'09 corpus annotators)

We further represent this interaction as in Figure 3.3.

Interaction Type	Cause	Theme	Anatomical location	Negated	Speculated
Positive regulation	G-CSF	StatG	myeloid cells	no	no

Figure 3.3: The example of semantic representation of an event.

In this case, neither of the two protein names, *G-CSF* and *StatG*, can be

normalised to a database identifier and therefore are represented with their textual mentions.

An event can be triggered by a number of different triggers. For example, a positive regulation event can be triggered by *"activate"*, *"effect"*, *"regulate"*, *"depend"*, *"involve"*, *"role"*, or many other words. Despite this lexical variation, we expect two positive regulation event that describe the same mothelecular process, using different terms to be treated as equal. Similarly, named entities are referred to by various lexical terms. We would like this lexical variation not to affect the presentation of an event. In other words, a "regulation" event by any other name would refer to the same underlying molecular process.

In the case that any of the other features are not applicable (e.g. cause for any event type other than regulation) or do not exist (e.g. not mentioned in text), we assign a null value to that field. If an entity exists, but cannot be normalised to a database identifier, the exact string of the term is recorded instead.

The feature "participant" encodes the the gene/protein as well as species (or organism) in the cases that the participant is a gene or protein and it is normalised. For example, the term "p53" appearing in an article that discusses human patients, will be resolved into a unique identifier referring to human p53 (NCBI Entrez Gene ID 7157), which will be different from that of, say, mouse p53 (NCBI Entrez Gene ID 22059). For this reason, wherever we mention a normalised gene or protein, the species will be implicitly included. However, in this research we do not include species as a separate column in the event representation at this stage, as the gene or protein is not always normalised.

This representation only selects a subset of the contextual attributes that could be associated with an event. It can be expanded to include a wide range of contexts related to the reported molecular event and the experimental settings, from population (including species) to features like *in vivo/in vitro*, and the like. On the other hand, the minimum set of features that gives an

identity to an event consists of the event type (feature 1) and the theme (included in feature 2). These features are not optional and must exist in every event. “Event type” can be anything from an event ontology entry to a predefined set of types. In our research we consider the nine BioNLP’09 event types as acceptable values for this field.

This representation differs from the **mention** level representation of an event which concerns the syntactic attributes of the textual expression of the event such as the document and sentence it has appeared in, the exact terms used to describe the event and its participants, the terms used to assert or affirm the event, and the offset where these terms appear. An extensive list of the attributes that we have used in the mention level definition of an event is shown in Table 5.2 on page 207.

Although we include the trigger term as an attribute in the mention-level representation, to prevent the representation of biologically distinct events from being affected by the lexical variability of trigger terms and entity names, we do not include such “surface” features in the this representational definition.

3.1.4 Conflicting statements

As discussed in Chapter 2, there is no consensus in the research community on the definition of the concepts of contrasts and contradictions, so here we introduce our definition with relation to the aim of the conflict extraction task.

Contradictions. In this thesis, we are interested in contradictory statements expressed in possibly different documents, possibly by different authors. We distinguish between the following types and degrees of contradiction.

1. **Logical contradiction in biology**

   This type of contradiction would mean that a statement p and its opposite are true simultaneously. For example, if the preposition *“p53 is expressed”* is always true, then if a particular instance of p53 protein is not expressed we will have a contradiction of this type. We expect

this situation to never happen in nature in the same context.

2. **Contradiction in the literature**

   This type of contradiction happens between two statements from the literature reporting facts about the same subject. In other words, when sentence A states that *p* is true (e.g. an event happening) and sentence B states that $\neg p$ is true (e.g. an event not happening) we will have a contradiction of this type. For example, when author A states that *"p53 is expressed in mouse lung tissue"* and author B states that *"p53 is never expressed in mouse lung tissue"*.

3. **Contradiction in extracted data**

   This type of contradiction happens between statements that are generally conflicting, but that appear to be contradictory due to underspecification or incomplete context. For example *"p53 is expressed in mouse lung tissue at 36° C"* and *"p53 is not expressed in mouse lung tissue"*. Here, one sentence states that an event happens in a certain temperature, and the other states that the same event does not happen in a different temperature. Failing to extract the context of the event that is related to the temperature, one might argue that the two events are contradictory. However, adding the relevant context to the extracted events would reveal that the two events are only contrasting as they differ in a contextual feature.

This brings us to our definition of contradictory events. Amongst an aggregate number of events extracted from a large body of literature, we say two events are **contradictory** if they share features 1 (type), 2 (participants), and 3 (anatomical location) introduced in section 3.1.3, and are both assertive (feature 5), but differ in polarity (feature 4).

Example 3.6.

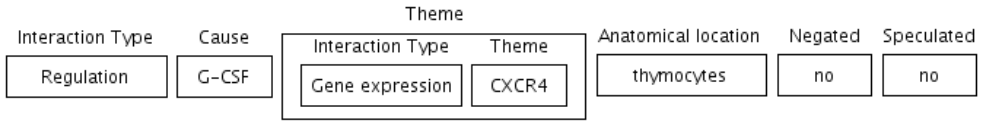
(a) *"Positive regulation of CXCR4 expression and signaling by*

interleukin-7 in CD4+ mature thymocytes correlates with their capacity to favor human immunodeficiency X4 virus replication." (From PMID 12719571, extracted by BioContext.)

(b) *"In contrast, in intermediate CD4(+) CD8(-) CD3(-) thymocytes, the other subpopulation known to allow virus replication, TEC or IL-7 has little or* ***no*** *effect on CXCR4 expression and signaling."* (From PMID 12719571, extracted by BioContext.)

The semantic representation of the two events in Example 3.6 are shown in Figure 3.4.

(a)



(b)

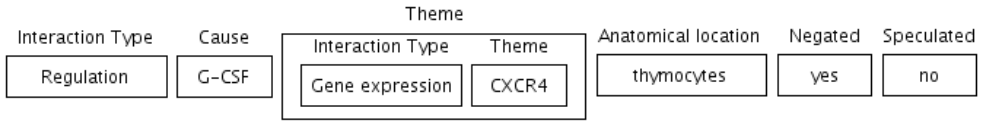


Figure 3.4: Semantic representation of the conflicting events in Example 3.6.

Note that the theme is an event, and is shown in the form of a recursive event. The complete recursive theme columns are not shown for simplification.

It can be seen in the figure that the two events have everything in common, except for their negation status. And as far as can be inferred from the sentences, they state contradictory claims.

Contrasts. We say two events are **contrasting** if they share all of the identifying features 1-5, except for either feature 2 (cause or theme) or feature 3 (anatomical location). Similar to contradictory events, they also need to differ

in their polarity (feature 4). More generally, two contrasting events have different polarities, but match in all but perhaps one contextual feature.

Example 3.7.

(a) *"In addition, cloning efficiencies were acceptable (over 30%) when IL 2 produced spontaneously from the leukaemic cell Jurkat (M-N) was used."*

(From PMID 6278580, extracted by BioContext)

(b) *"However, IL-2 is* ***not*** *normally synthesized by solid tumor cells."*

(From PMID 2190213, extracted by BioContext)

The two sentences in Example 3.7 refer to the expression of *IL-2*. Sentence ***(a)*** states that IL-2 is expressed in certain leukaemic cells, whereas sentence ***(b)*** says that it is not expressed in tumor cells. They differ in the anatomical entity (feature 3), but are the same in every other aspect. The semantic representations are shown in Figure 3.5.

(a)

Interaction Type	Theme	Anatomical location	Negated	Speculated
Gene expression	IL-2	Jurkat	no	no

(b)

Interaction Type	Theme	Anatomical location	Negated	Speculated
Gene expression	IL-2	tumor cells	yes	no

Figure 3.5: Semantic representation of the conflicting events in Example 3.7.

The two events match in all attributes except in their anatomical location (Jurcat vs. tumor cells).

It is difficult to determine whether this contrast is a contradiction or not. We need expert domain knowledge to know whether leukaemic cells are actually a type of tumor cells or not. Tumor cells concern cancer, and leukaemia is a type of cancer, but inferring anything further than than is not easy for non-experts.

Contradictory and contrasting events are special cases of a broader concept of **conflicting** statements: events that share *some* characteristics that would justify comparing them, but are not necessarily in agreement with each other.

We also introduce the notions of **strict** and **relaxed** conflicts. In strict conflict, we require every field of an event representation to be filled (known) before they can be compared. For example, if the cause of two regulatory events or their anatomical locations are not mentioned in the sentence and are therefore missing from the event representation, we cannot compare them in a strict way. On the other hand, in relaxed conflict, the missing or null fields are ignored when comparing two events. Obviously, two events that are strictly conflicting are more likely to represent an actual contradiction in the natural sense, due to the more complete context that is associated with it. On the other hand, relaxed conflicting events are less likely to be contradicting or contrasting.

In the following sections we introduce our methods for extracting molecular events and detecting their negation and speculation.

3.2 Semantic tokenisation

Sentences that represent molecular events are typically long and complex. For example, the sentences in the BioNLP'09 corpus are on average 26 words long, and there are outlier sentences that are more than 130 words long. These sentences contain many biomedical named entities, many of which consist of multiple words (e.g. *tumor necrosis factor-alpha* or *Homo Sapiens*). Some of these entities can even be parts of what would typically be considered as

tokens. For example, a phrase like *"NFkappa B/Rel"* refers to two entities: *NFkappa B* and *Rel*. A good tokeniser should recognise those entities as individual tokens, and not, for example, pick *B/Rel* as a token.

This is a challenging task, however, due to the biologically and linguistically complex nature of the named entities. In turn, this may confuse the parser and cause it to produce a sub-optimal parse tree. In the first stage of our work (the event extraction task) this was one of the major problems that arose when trying to align the extracted event components with the parse trees. The parse-tree-based rules mostly concerned individual nodes in the tree, whereas an entity could have been broken up across several nodes, and depending on which node was treated as the head, could result in different features. On the other hand, two entities could be grouped together in the same parse node, and end up having identical parse tree features, despite their different roles.

To address this issue, we delay tokenisation and parsing until after the named entities are extracted. Tokenisation then takes into account named entities as single tokens of the nominal type and automatic parsing is performed on the sentences with semantically separated tokens. We refer to this process as **semantic tokenisation**. We hypothesise that semantic tokenisation would increase the quality of the parses, and therefore the feature extraction process.

To demonstrate the process of semantic tokenisation, consider Example 3.8.

> **Example 3.8.** *"Tumor necrosis factor (TNF)-alpha-induced HIV-1 replication in OM10.1 or Ach2 cells was significantly inhibited by non-cytotoxic doses of AuTG [...]"*
> (PMID 10069412 from the BioNLP'09 corpus)

Figure 3.6 shows part of this sentence before and after semantic

tokenisation. Figure 3.6(a) shows the tokenisation as performed by GENIA and McClosky parsers and provided as part of the BioNLP'09 corpus. Figure 3.6(b) shows the same part of the sentence, but tokenised after the entity *"Tumor necrosis factor (TNF)-alpha"* was recognised by the gene name recognisers[17] and was treated as a single noun. At this stage, the recognised named entities are replaced by a generic noun to make sure that the parses treat them as nouns.

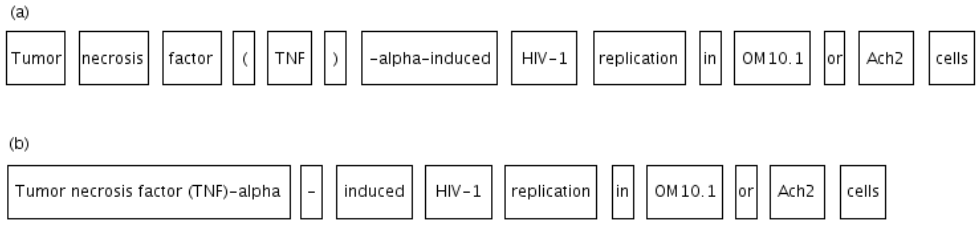


Figure 3.6: Example of semantic tokenisation.

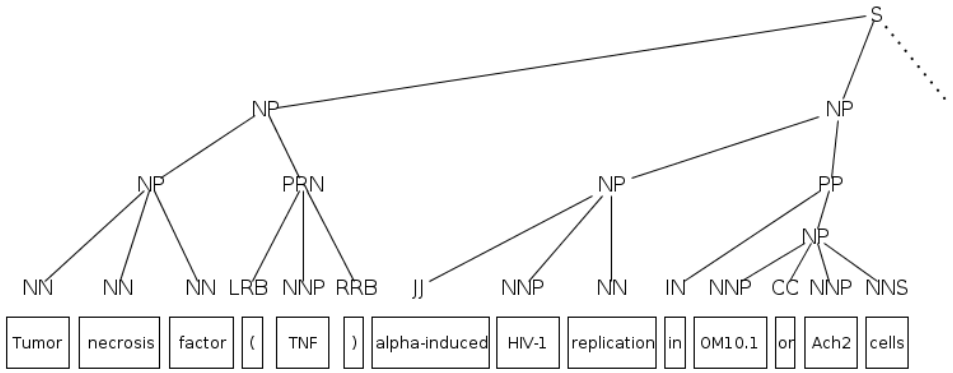
Part of the example sentence tokenised (a) in the default tokenisation of the BioNLP'09 corpus and (b) using semantic tokenisation.

Figure 3.7 shows the partial parse tree of the same sentence, as the named entity is treated as a single-token proper noun. As this figure demonstrates, the word *"alpha"* is part of the named entity *"Tumor necrosis factor (TNF)-alpha"* and should not be separated from the rest of the compound named entity. However, as can be seen in Figure 3.7(a), the parser has recognised *"Tumor necrosis factor (TNF)"* as a noun phrase, and has grouped *"alpha"* with another separated noun phrase. Consequently, if *"Tumor necrosis factor (TNF)-alpha"* is marked as a named entity, and *"alpha"* is treated as the head of this named entity (although it is not technically the head of the phrase), the features extracted for this token will be generally applied to the entire named entity. As this example shows, those features could be very different from the features of any of the other tokens of the named entity. This issue is addressed in Figure 3.7(b) in which parsing has been performed after the semantic tokenisation, and therefore the features extracted from the named entity will more likely be correct.

17 See section 5.1.3 for more details.

Note that in this example, the parser has tagged "*induced*" as a past participle form of a verb in Figure 3.7(b), whereas the correct grammatical structure of this phrase should recognise this word in the phrase "*Tumor necrosis factor (TNF)-alpha-induced*" as an adjective.

(a)



(b)

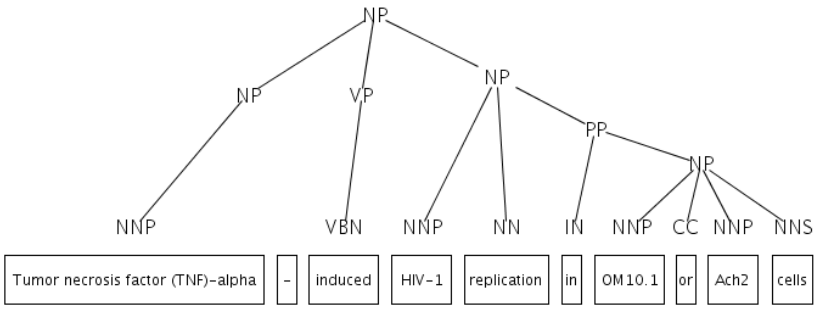


Figure 3.7: Semantic tokenisation on the parse tree of a sentence.

Syntactic parse tree of the example sentence, partially shown to demonstrate semantic tokenisation (a) without semantic tokenisation and (b) with semantic tokenization. The named entity "Tumor necrosis factor (TNF)-alpha" *is treated as a single-token proper noun in (b).*

Replacing multi-token entities with place-holder nouns could potentially

affect the performance of the other tools. For instance, sentence splitting is performed after named entity recognition, and capitalisation is one of the features that sentence splitters look for to determine the beginning of a sentence. To prevent sentence splitters from missing the split where the next sentence starts with a multi-token named entity, we make sure that we maintain the capitalisation when we replace the named entity with the place-holder noun.

Another issue is the overlapping entities. Named entities recognisers often recognise entities that have overlapping spans. This can sometimes be an error, for example in *"Human Immunodeficiency Virus"* at least three named entities can be recognised, namely *"Human"*, *"Virus"*, and *"Human Immunodeficiency Virus"*. Failing to recognise the longest string, the entity mention could be resolved into an incorrect type, e.g. *"Human"*. However, it does not always cause an error: often in the literature entities are actually overlapping, as one named entity could have a sub-component which is also a named entity of a similar type, with the longer entity merely being more specified. For instance, in *"fruit fly"* or *"Persian cat"*, recognising entities *"fly"* or *"cat"* would not resolve into incorrect entities, but into underspecified entities.

In such cases, considering the union of the overlapping named entities would solve both underspecification and incorrect problems. Therefore we take the union of the overlapping named entities as one named entity and assign the normalised properties of the longest named entity of the group to the resulting string.

3.3 Extracting molecular events

At the time of this study, no suitable molecular event detection software was publicly available, so we developed an event detection system called *Evemole* that uses a combination of machine learning and rule-based methods and makes use of dependency parse-tree-based features. It takes an input a sentence

with the gene and protein named entities already recognised.

The system consists of two main modules: (1) event trigger and type detection, and (2) event participant detection. Each module has a post-processing stage. Figure 3.8 shows an overview of the system.

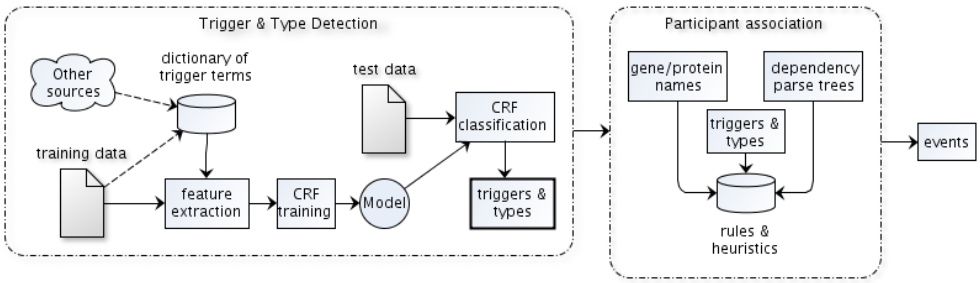


Figure 3.8: Overview of the event extraction system, Evemole.

The following sections explain the two main phases of the event extracting system, i.e. trigger and type detection, and participant association.

3.3.1 Event trigger and type detection

Our view of the event trigger and type detection modules was that each token in a sentence needed to be tagged either as a trigger for one of the nine event types, or as a non-trigger/event token. We therefore decided to identify event types and triggers in a single step by training a conditional random field (CRF) classifier that assigned one of ten (nine types plus non-trigger) tags to each token. CRFs have been shown to be particularly suitable for tagging sequential data such as natural language text, because they take into account features and tags of neighbouring tokens when evaluating the probability of a tag for a given token (see Section 2.3.4).

Tokens and their part-of-speech (POS) tags were recognised using the GENIA Tagger (Tsuruoka et al. 2005). Each stemmed token was represented using a feature vector consisting of the following features:

- A binary feature indicating whether the token is a protein (as identified by gene NER);

- A binary feature indicating whether the token is a known protein-protein interaction word—we used a pre-complied dictionary of such words collected from the training data and the previous studies (Fu et al. 2009); (Yang et al. 2008) (see Appendix B for the full list);
- The token's POS tag;
- The log-frequencies of the token being a trigger for each event type in the training data (nine features);
- The number of proteins in the given sentence.

Other features (e.g. separating the known interaction words according to the nine event types) were explored, but were not included in the final feature list since they increased the sparseness of the data and did not improve the overall results. Also, the high level of ambiguity among trigger words and event types meant that they could not be effectively used as features. For a full list of all trigger terms and the frequency with which they represent events of different type, see Appendix B.

In the following examples, a numeric type has been assigned by the CRF module to every token in the sentence. Type 0 indicates that the token is not a trigger word. Types with tags other than 0 are event triggers, and the numeric value indicates the type of the event. Table 3.1 shows tagging of the tokens in the phrase *"I kappa B/MAD3 masks the nuclear localization"* in which the word "*masks*" has been tagged as class 9, indicating that it is the trigger of an event of type Negative regulation.

tokens	I	kappa	B/MAD3	masks	the	nuclear	localization
tags	0	0	0	9	0	0	0

Table 3.1: Example tagging of a phrase by CRF

Tag 0 means that the token is not an event trigger. Tag 9 means that the token is the trigger of an event of type Negative regulation.

The performance of this phase was studied on the BioNLP'09 development dataset: we noted a number of false-positive and false-negative results that were mostly due to the incorrect identification of a set of recurring triggers. We therefore decided to perform a post-processing step to improve the identification of event triggers and event types.

For this purpose, the output of the CRF module was overridden in cases where the triggers appeared in a list of negatively discriminated trigger words which was collected after the manual analysis of the false positive results on the training and development data. Similarly, in cases where the CRF missed a highly indicative trigger from a manually collected set for a given event type, the trigger was added during the post-processing step (see Appendix B for a complete list).

Finally, since triggers could consist of more than one consecutive token, a set of simple rules were applied to remove typical false-negative constituents identified by the CRF as part of triggers, namely removing *'whereas'*, *'and'*, *'or'*, and *'but'* if recognised as part of a multi-token trigger.

In this task we used Monte (Memisevic 2007) which is a Python framework for building gradient based learning machines like neural networks and conditional random fields. It provides a range of kernel functions and trainers including linear and sigmoid functions.

3.3.2 Locating event participants

After detecting potential triggers and associated event types, the next task was to locate possible participants (i.e. 'themes' and 'causes') for each event.

We hypothesised that the linguistic structure of the sentence corresponds to biological semantics that are conveyed by it. It was obvious that participants did not have to be the nearest to the trigger on the surface level, so our approach was based on distances within the parse trees associated with the sentences containing candidate events. Parse tree distances have been studied previously in clustering and automatic translation tasks (Emms 2008), as well

as relation extraction tasks as discussed in Section 2.3.4. Therefore, we hypothesised that we could use their properties to identify the most likely participants.

The BioNLP'09 training data was analysed for the proximities between the triggers and the (correct) event participants in the dependency parse tree of the sentence. This analysis demonstrated that it was more likely for a theme to appear in the sub-tree of the corresponding trigger, with 70.5% of all single theme events (class I) having a theme which appeared in the sub-tree of the trigger. Figure 3.9 shows the proportions of event participants that are in the sub-tree vs non-sub-tree of the event trigger, for different classes of events.

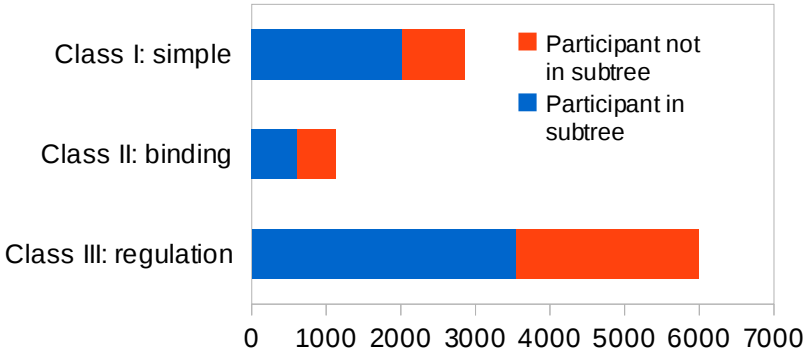
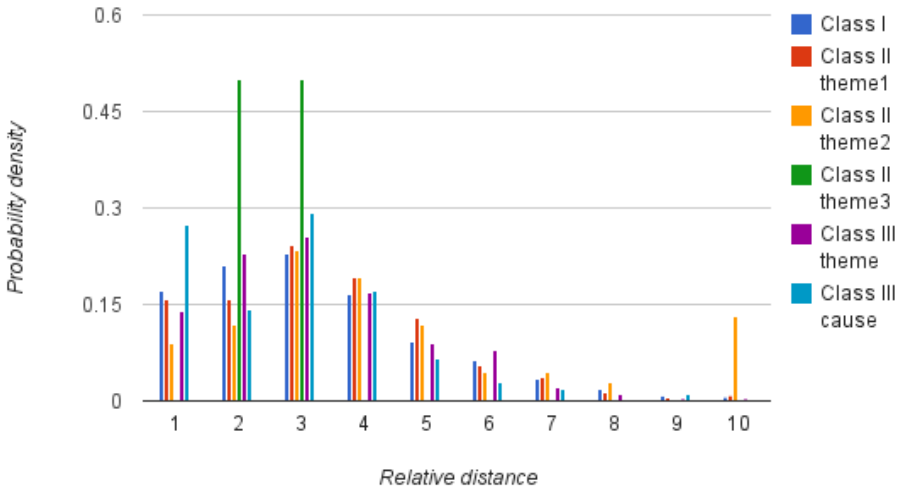


Figure 3.9: Sub-tree vs. non-sub-tree distribution of event participants

The proportions of event participants that are in the sub-tree vs non-sub-tree of the event trigger, for different classes of events.

Figures 3.10, 3.11, and 3.12 present detailed density functions of the dependency tree distances when the participants are or are not in the trigger's sub-tree, together with the cumulative distributions of these functions. (ignoring non-protein nodes). The analysis showed that the theme was usually amongst the nearest proteins to the trigger in terms of dependency parse tree distances: for example, in 60% of all class I events (i.e. single theme events e.g. *gene expression*, *localization*, etc.) the correct protein participant was the trigger's nearest or second nearest protein in the parse tree.

(a)



(b)

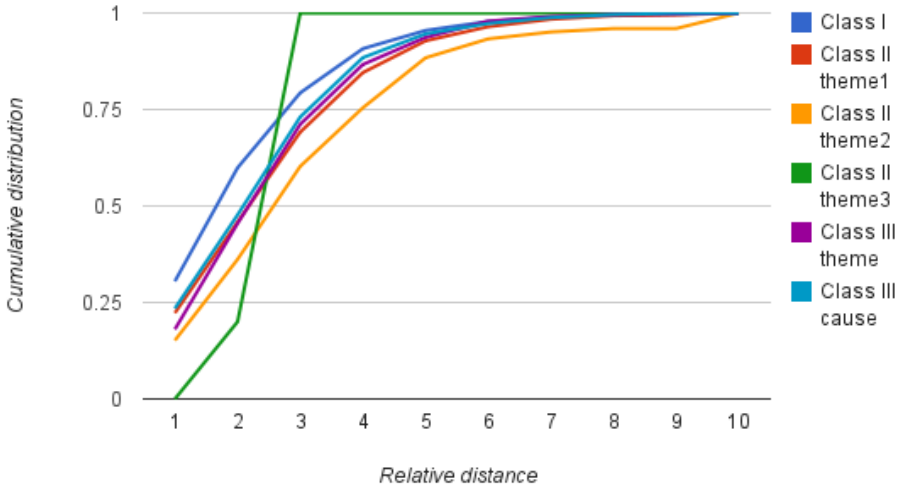
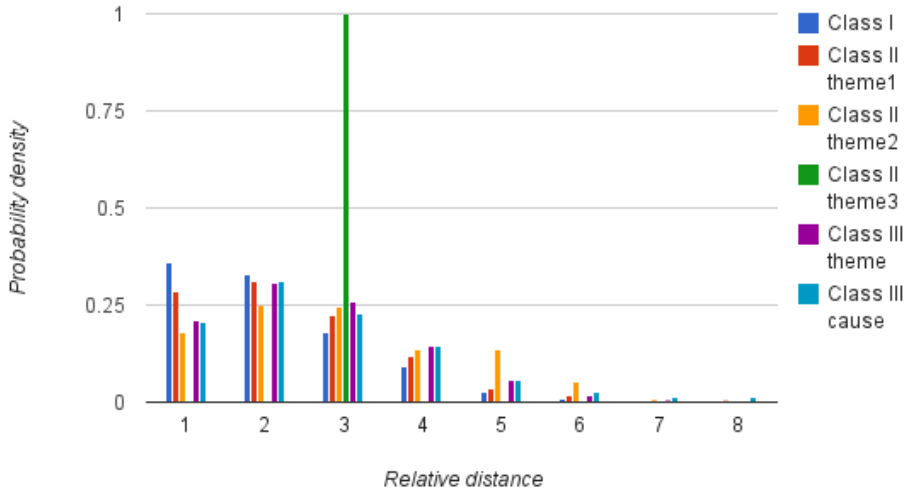


Figure 3.10: PDF and CD for participants in sub-tree of trigger

(a) Probability density function (PDF) and (b) cumulative distribution (CD) of the dependency distances between the trigger and the participant in the parse tree, when the participant is in the trigger's sub-tree.

(a)



(b)

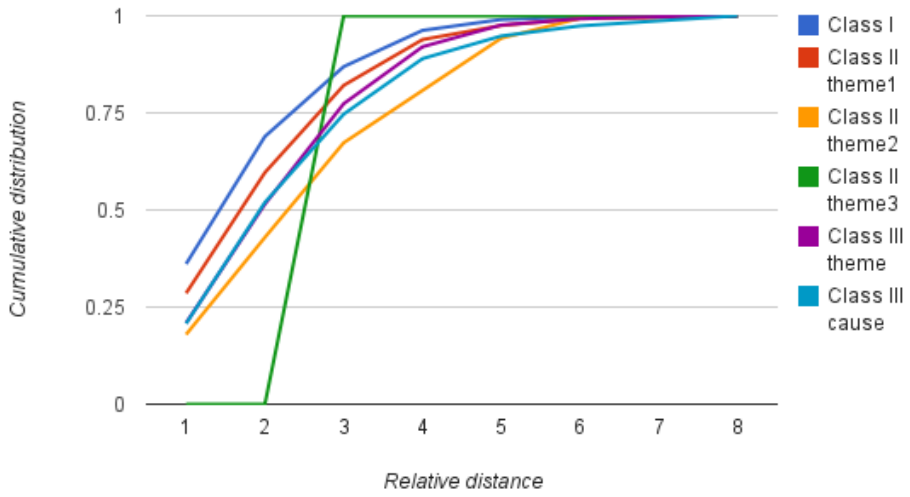
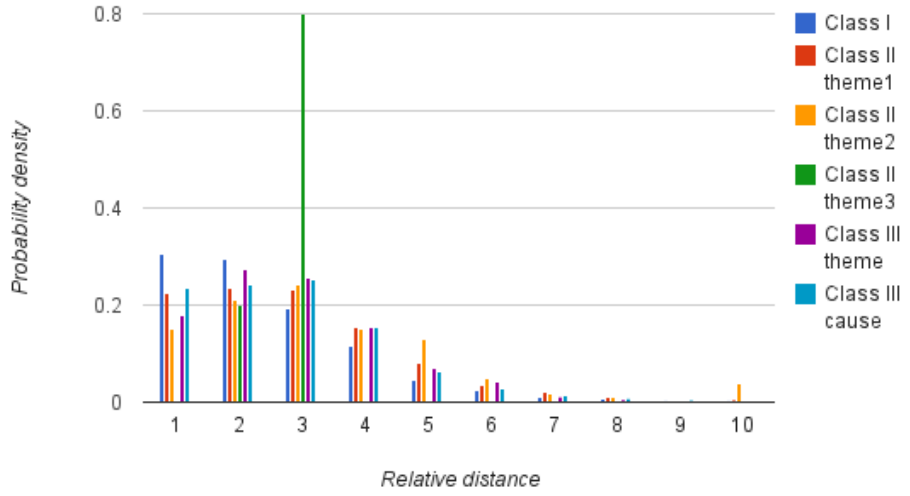


Figure 3.11: PDF and CD for participants not in the sub-tree of the trigger

(a) Probability density function (PDF) and (b) cumulative distribution (CD) of the dependency distances between the trigger and the participant in the parse tree, when the participant is ***not*** *in the trigger's sub-tree.*

(a)



(b)

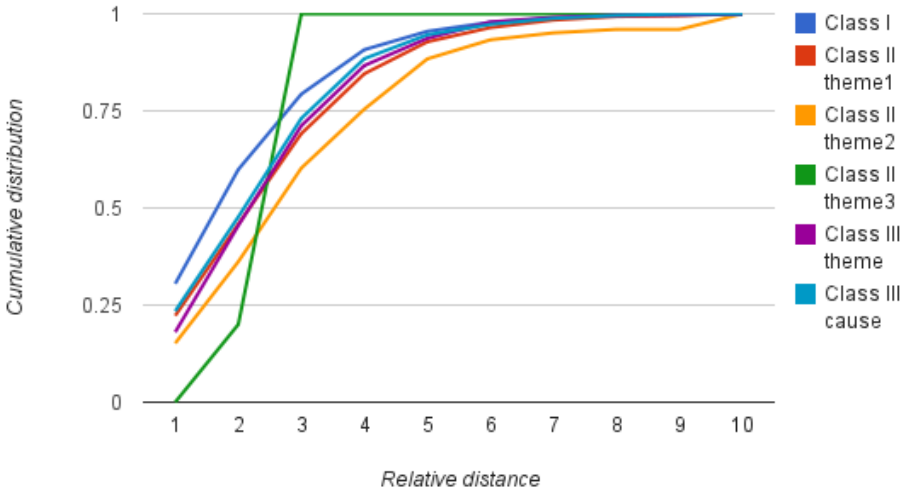


Figure 3.12: PDF and CD of the participant distances from trigger

Overall (a) probability density function (PDF) and (b) cumulative distribution (CD) of the dependency distance between the trigger and the participants in the parse tree.

The same pattern was observed in other event classes as well. Specific analyses of the parse trees associated with the class II (i.e. binding events which may have more than one theme) suggested a linear relationship between the parse tree distance and binding event participant number (the first participant is the nearest, the second participant is the second nearest, etc.). In Figures 3.10, 3.11, and 3.12 the third theme of binding events stands out as it most commonly appears in the third relative place.

We used this distributional analysis (derived from the BioNLP'9 training data) to design a rule-based method for the identification of participating themes. The rules were manually derived for each of the nine event classes, by defining:

- a threshold for the maximum distance to the trigger in the sub-tree for the given event type;
- a threshold for the difference between the maximum distance in the whole tree and the given sub-tree for the given event type;
- the number of nearest proteins to be reported for each trigger.

The thresholds were chosen experimentally to maximise performance, and are comparable to those used in previous studies.

Table 3.2 contains an algorithm demonstrating the rules that were used to assign participants to events of different types. All entities that satisfied a distance-based rule for a given trigger were selected as the corresponding theme(s) and/or cause. For example, if the event type was binding, then up to the second closest protein in the sub-tree, and the first closest protein in the rest of the tree are reported as themes.

0. The trigger is already found with CRF
1. Make a list of all the proteins in the sentence in which the trigger is found, sorted by the parse tree distance between the protein and the trigger
2. Depending on the type of the trigger apply one of these algorithms
   2.1. If the trigger is of types *Transcription, Localization, Phosphorylation,* or *Protein catabolism*
      - Iterate over the sorted list until the distance between the protein and the trigger minus the distance between the nearest protein to the trigger and the trigger is more than a threshold. Return all such proteins.
   2.2. If the trigger is of type *Gene Expression*
      - Iterate over the sorted list until some threshold on the distance, returning only the proteins in the sub-tree.
   2.3. If the trigger is of type *Binding* return the proteins which are
      - In the sub-tree of the trigger and have a distance less than a looser threshold; or
      - Have a distance less than a tighter threshold
   2.4. If the trigger is any of the *Regulation* types
      - If there is an event already detected in the same sentence, return the event
      - Otherwise return the nearest protein
3. Report all proteins/events found in step 3 as the participants of the event

Table 3.2: Algorithm to to associate entities with triggers

The algorithm used to construct events by associating entities as participants to triggers with already assigned types.

Engineering such rules for non-regulatory events was relatively straightforward. However, regulatory events could have different kinds of

participants (a protein or an event). Therefore, this would require a number of recursions in the application of the rules to represent nested regulatory processes. Still, the regulation events were specially complicated to detect, and particularly because the type of participant (theme/cause) had to be distinguished, so we aimed for a high recall and reported all the combinations of the entities in the sentence.

In the case of the participant being an event, we locate the nearest trigger for the event (being regulated) in the parse tree. For example, in Figure 3.13, the nearest option to the regulation trigger (*"secretion"*) was the trigger of the two localization events, and both events should be reported as the themes of two regulation events.

To further demonstrate the method we study the sentence shown in Example 3.9.

> **Example 3.9.** *"Monocyte tethering by P-selectin regulates monocyte chemotactic protein-1 and tumor necrosis factor-alpha secretion."*

Figure 3.13 shows the parse tree of Example 3.9 which contains multiple events.

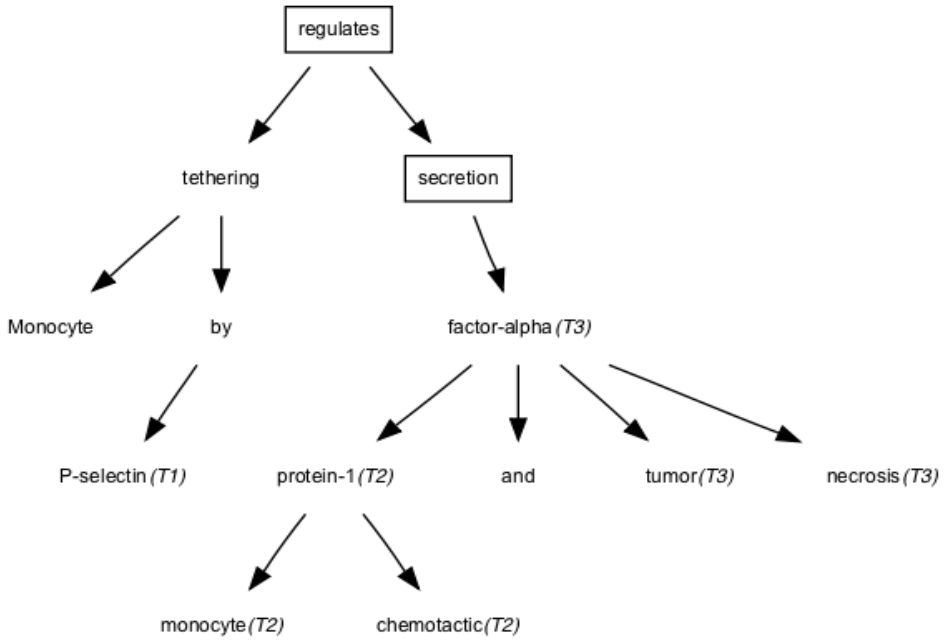


Figure 3.13: The parse tree of Example 3.9

The detected triggers are shown in boxes. Entities are numbered and are shown by italic T.

The words "*regulates*" and "*secretion*" are correctly identified as triggers for a *regulation* and a *localization* event in the first phase. Using the rules for localization, we would then correctly identify the themes for two localization events from the sentence parse tree as proteins *T2* and *T3* ("*monocyte chemotactic protein-1*" and "*tumor necrosis factor-alpha*".) It correctly ignores *T1* ("*P-selectin*") since it did not appear in the trigger's sub-tree.

The nearest option to the regulation trigger is "*secretion*", which is the trigger of the other events. Therefore both events would correctly be reported as the themes of the two regulation events. Figure 3.14 shows the representation of these four events.

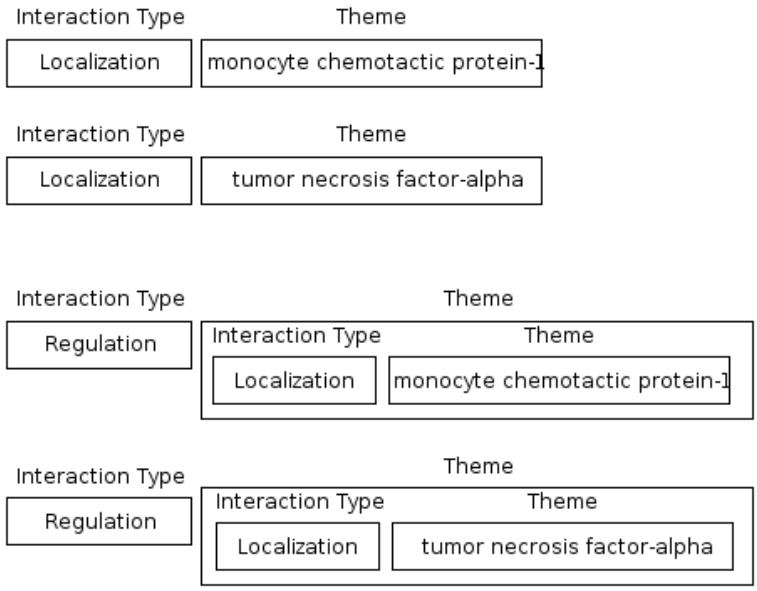


Figure 3.14: Representation of the events with participants

The events are extracted using parse-tree rules.

It can be seen that a number of recursions in the application of the rules would be required to represent higher-order regulatory events. For the purposes of this study, only regulations up to the second "order" were detected, allowing other events to act as themes and causes as well as proteins.

Similar to the previous stage, here we also performed post-processing based on studying the output of the system on the BioNLP'09 development data. In this phase, we forced highly indicative regulation triggers (if not previously identified) to be associated with an event by assigning proteins appearing in the sentence to them, even when no protein in the sentence satisfied the theme or cause criteria. This was aimed at improving the extremely low recall for regulatory events.

3.4 Extracting negation

Negation and speculation are defining features in extracting conflicting events.

We introduce a method, called *Negmole,* that uses machine learning to classify molecular events as negated or speculated. We further build a pipeline to integrate several text mining tools, including Negmole, to extract and contextualise molecular events from the available body of scientific literature. We use the event data that is extracted by this system to find conflicting events across the literature, and classify them as candidates for contrasts and contradictions.

We primarily focus on (and design our method for) negation detection, but also adjust the method for speculation detection with some minor modifications. At this stage, we assume that entity mentions, event triggers, types, and participants have already been extracted, so this data can be used as the input.

Initially, we implemented the NegEx algorithm (see Section 2.4.2) as a baseline and applied it on the union of training and development data, by considering the event triggers as the terms required by the algorithm. The P/R/F-score achieved by this method was 36%/37%/36%. Analysing NegEx's FP and FN predictions (after leaving the development data unseen) we identified the following patterns contributing to the errors.

It is a common characteristic of the biomedical literature abstracts describing protein interactions that more than one event or interaction is expressed in a sentence. Amongst those sentences that do express an event, an average of 2.6 event triggers appear, and the number of events actually described can be higher, as the event trigger does not repeat when the participants of different events are joined by a conjunction. Example 3.10 shows a sentence describing a number of events.

> **Example 3.10.** *"In this study, we demonstrate that Tax-stimulated nuclear expression of NF-kappa B in both HTLV-I-infected and Tax-transfected human T cells is associated with the phosphorylation and rapid proteolytic degradation of I kappa B alpha."*

(From PMID 7935451)

The NegEx algorithm implicitly assumes the occurrence of one idea per sentence. In the case where several events are expressed in the sentence—whether they are separated by conjunctions or not—NegEx fails to detect the correct scope. See Example 3.11, for instance.

> ***Example 3.11.*** *"We also demonstrate that the IKK complex, but not p90 (rsk), is responsible for the in vivo phosphorylation of I-kappa-B-alpha mediated by the co-activation of PKC and calcineurin."*
> (From PMID 10438457)

More specifically, when there is a contrast expressed in the sentence as in Example 3.11, NegEx fails to determine which one of the contrasting statements are negated and which one is affirmed. This could happen in cases where no conjunction appears, as in Example 3.12.

> ***Example 3.12.*** *"In contrast, NF-kappa B activity was not detected in the nucleus following long-term expression of Tax in Jurkat T lymphocytes."*
> (From PMID 1964088)

As there are more than two interaction words per sentence for sentences that describe an interaction, for any method to be able to effectively detect negations, it should be able to link the negation cue to the specific token/event trigger/entity name in question.

Figure 3.15 shows an overview of the negation detection system, Negmole, that uses detected entities and events as input. We construe the negation detection problem as a classification task where the aim is to classify the previously detected events as affirmative or negative. The same applies to

the speculation detection task (see Section 3.5). To extract negated and speculated events, we use machine learning with lexical, syntactic, and semantic features. Lexical features include negation cues, part-of-speech tags, and surface distances between key elements of an event. Syntactic features include the relationship between those elements within the constituency parse tree of the sentence. Semantic features involve the biological characteristics of the events, such as the types of the events or their participants. These will be explained in the following sections.

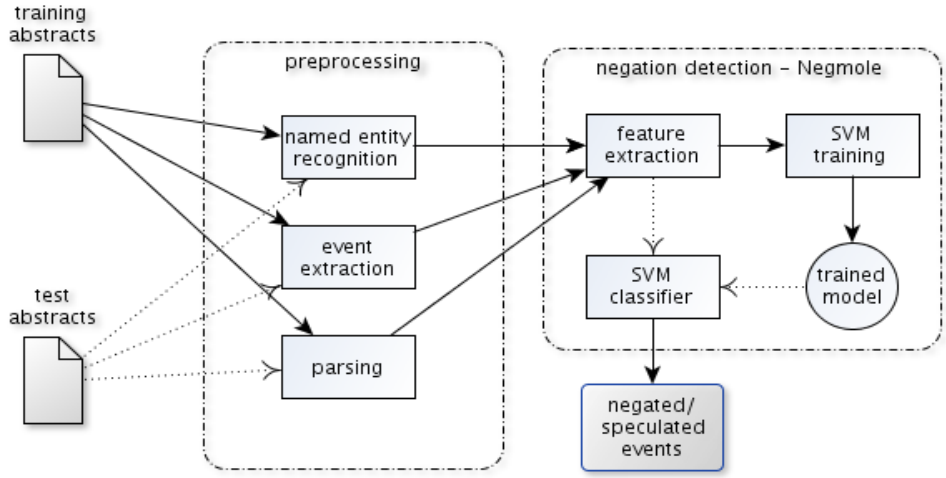


Figure 3.15: An overview of Negmole

The requirements of Negmole are specified.

About 95% of the annotated events in the BioNLP'09 corpus are encompassed in a single sentence (Björne et al. 2009), so we limit our attention to events that have their components (trigger and participants) within sentences. Negmole then classifies each event using the features engineered from an event-representing sentence. The following sections describe the details of the system.

3.4.1 Detecting negation and speculation cues

Detecting negation cues (see Section 2.4) can be generally construed as a

named entity recognition problem. Compared to other entities discussed so far, words and phrases indicating negation and speculation are relatively less ambiguous and have less term variability. Previous approaches that have used dictionary look-up methods report a reasonable performance (see Section 2.4), and thus here we also use a similar dictionary-based approach.

We built the cue dictionaries semi-automatically by analysing the BioNLP'09 training data. For both negation and speculation, we first considered exact matches between tokens and cues in the dictionaries. In later experiments, we used stemming to also detect inflected forms of the cues.

Negation cues

We used two different cue sets, a smaller, stricter set, and a larger set that included the first set. The small cue set was largely composed of general linguistic cues, whereas the larger one also contained domain-specific cues that would not necessarily have been considered a negation cue in a different domain. Words such as *"inhibit"*, *"unchanged"*, and *"block"* do not indicate a negation in natural language sentences *per se*. However, they commonly indicate the absence of the biomedical event that is being discussed. The two sets of negation cues as well as the negation cues with stemming are listed in Table 3.3.

The distribution of the most frequent of these cues in the BioNLP'09 training data is shown in Figure 3.16. This figure only shows the occurrence of these cue words in the corpus and not whether these words indicate any negated event. As the negation cues are not annotated in this corpus, it is not possible to infer with certainty what has been the exact clue for the annotators to mark a given event as negated. In fact, many of these occurrences, e.g. the occurrences of *"inhibit"* may indicate affirmative down regulation events. Others may refer to other negated concepts not related to molecular events. On the other hand, some negated events may have been marked as negated due to a cue that is not included in this list.

Small negation cue set without stemming	no, not, none, negative, without, absence, fail, fails, failed, failure, cannot, lack, lacking, lacked
Large negation cue set without stemming	no, not, none, negative, without, absence, fail, fails, failed, failure, cannot, lack, lacking, lacked, inactive, neither, nor, inhibit, unable, blocks, blocking, preventing, prevents, absent, never, unaffected, unchanged, impaired, little, independent, except, exception
Final negation cue set (stemmed)	absenc, absent, block, cannot, except, fail, failur, impair, inact, independ, inhibit, lack, littl, neg, neither, never, no, none, nor, not, prevent, unabl, unaffect, unchang, without

Table 3.3: The negation cue sets used in different experiments.

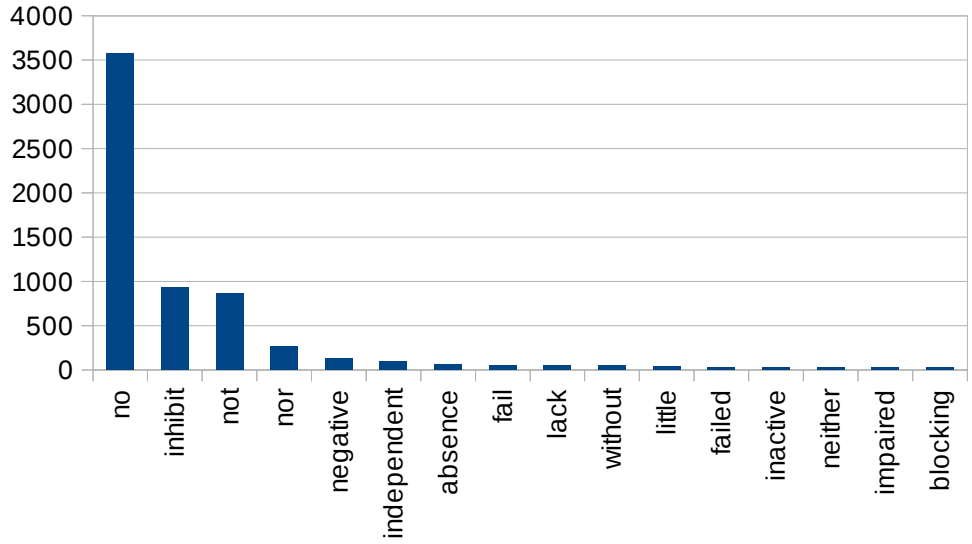


Figure 3.16: Distribution of negation cues in the BioNLP'09 training data

Showing the most frequent negation cue word occurrences, but not necessarily indicating a negated event.

The experiments using the different cue sets differed slightly in precision and recall, but overall the F-score was not affected. More detailed results and analysis will be presented in Chapter 4.

Speculation cues

Similarly to the negation cues, we used a set of stemmed speculation cues as the reference dictionary to detect speculation cues. After some experiments with the stems selected from the training data, we discovered that three of the stems, namely *"thought"*, *"confirm"*, and *"delin"* (stem for delineate) are having an adverse effect on the performance of the speculation detection task when tested on the development data. The initial and final speculation cue sets with stemming are listed in Table 3.4.

Initial set of speculation cues, composed from the training data (stemmed)	mai, can, could, might, mayb, thought, suggest, hypothes, investig, ask, found, find, possibl, confirm, seem, appear, examin, like, unclear, evid, must, probabl, undefin, clear, implic, observ, postul, determin, analys, analyz, partial, propos, assume, whether , delin
Final set of speculation cues, after removing some cues experimentally (stemmed)	mai, can, could, might, mayb, suggest, hypothes, investig, ask, found, find, possibl, seem, appear, examin, like, unclear, evid, must, probabl, undefin, clear, implic, observ, postul, determin, analys, analyz, partial, propos, assume, whether

Table 3.4: The set of speculation cues used in different experiments.

Note that both sets contain stemmed cues.

The distribution of the most frequent of these speculation cues in the BioNLP'09 training data is shown in Figure 3.17. The same considerations as the negation cue distribution apply here as well, as this figure can only be an approximation of the speculation cues that have in fact caused the event to be speculated. Note the difference in scale between Figure 3.16 and Figure 3.17. The top negation cue appeared more than 3500 times, whereas the most frequent speculation cue appeared only under 400 times.

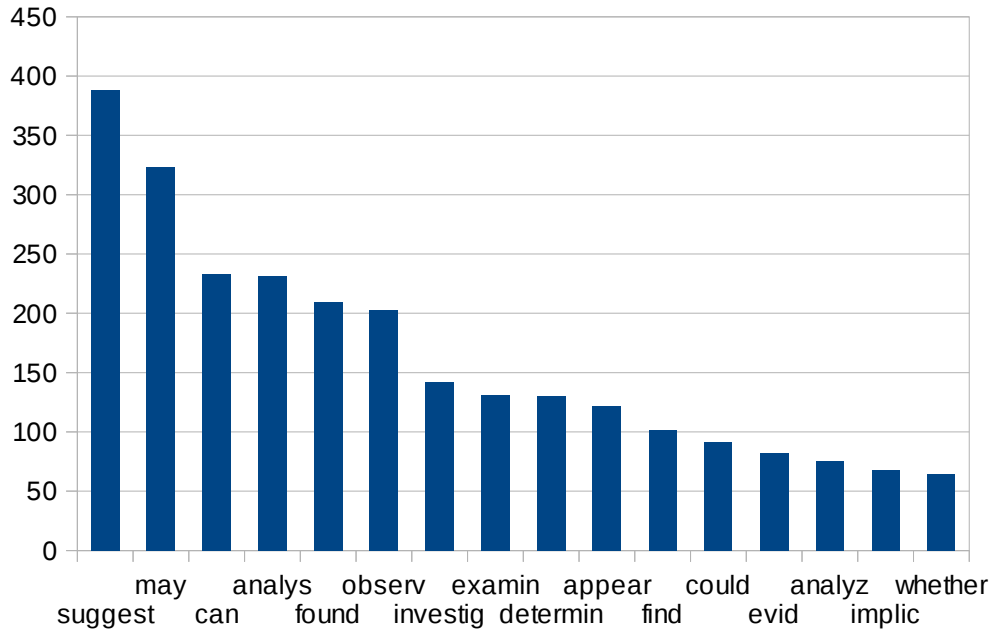


Figure 3.17: Distribution of speculation cues in the BioNLP'09 training data

Showing the most frequent speculation cue word occurrences, but not necessarily indicating a speculative event.

Handling multiple cues

It is common in general and specifically in the BioNLP'09 corpora for sentences to contain more than one negation or speculation cue. Analysing the training and development data sets shows that amongst the sentences that do contain a negation or speculation cue, on average they contain 1.44 cues. It was interesting that the distribution of the two cue types as well as the averages were quite similar, with the average number of negation cues per sentences with negation cues being 1.22 and the same statistic for speculation cues being 1.27.

Figure 3.18 shows the number of sentences in the combined BioNLP'09 training and development data sets for a given number of cues on a logarithmic scale. In this data set, there were two sentences with five speculation cues (as given in Example 3.13).

Example 3.13.

(a) *"We therefore investigated whether the activation of the IL-1RI-associated protein kinase could be a target for redox regulation and whether an altered activity of the kinase could influence IL-1-mediated NF-kappa B activation."*

(From PMID 9394832, speculation cues detected by Negmole.)

(b) *"Because this region of Stat3alpha is involved in transcriptional activation, our findings suggest the possibility that Stat3gamma may be transcriptionally inactive and may compete with Stat3alpha for Stat3 binding sites in these terminally differentiated myeloid cells."*

(From PMID 9823774, speculation cues detected by Negmole.)

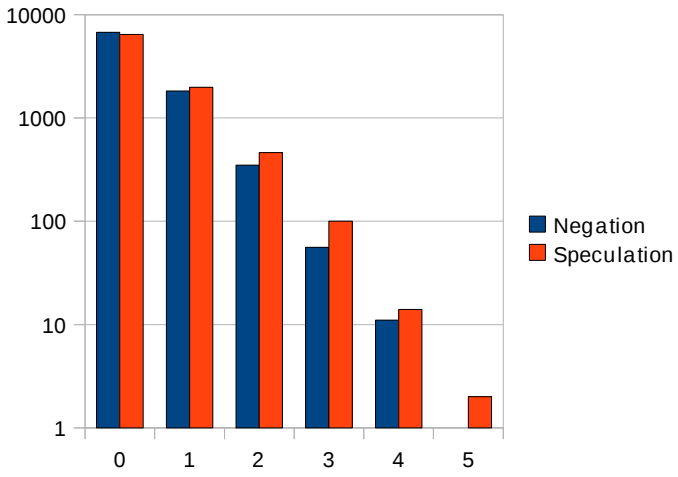


Figure 3.18: Distribution of sentences containing any cues

The number of sentences in the BioNLP'09 training and development data sets with a given number of negation or speculation cues, displayed on a logarithmic scale. Note that there was one sentence with five negation cues, which cannot be shown here. There are no sentences with six or more cues.

There was also one sentence in the corpus which contained five negation cues. This sentence and another example with four negation cues are shown in

Example 3.14.

Example 3.14.

(a) *"However, plasma membrane-proximal elements in these proinflammatory cytokine pathways are apparently not involved since dominant negative mutants of the TRAF2 and TRAF6 adaptors, which effectively block signaling through the cytoplasmic tails of the TNF-alpha and IL-1 receptors, respectively, do not inhibit Tax induction of NF-kappaB."*

(From PMID 9710600, negation cues detected by Negmole.)

(b) *"This inhibition was not mediated through Nef phosphorylation on Thr-15 or GTP-binding activity because mutations in critical sites did not alter this inhibition"*

(From PMID 7917514, negation cues detected by Negmole.)

Note that not all the cues in Example 3.14 indicate linguistic negation, as some of them refer to negative biological concepts, or down-regulation.

As we can see from these examples, the number of cues in a sentence can be an indication of how strongly the sentence expresses negation or speculation. Therefore, we use the number of cues in the sentence as a feature.

We choose one of the cues in the sentence as the main cue. Since our approach is event-oriented, we propose a way to identify the main cue for an event and not for the whole sentence. We hypothesise that the main cue is the cue which is responsible for the negation of an event, and we aim to extract relevant features from it. The main cue is selected for each event independently, rather than selecting a main cue for the whole sentence. For every event, the cue that has the shortest constituency parse tree path to the event trigger is selected as the **main cue**. Using this method, the main cue in a sentence can be different for every event.

To demonstrate how cue distances are calculated, consider Example 3.15.

Example 3.15.

"Structure and function analysis of the human myeloid cell nuclear differentiation antigen promoter: evidence for the role of Sp1 and ***not*** *of c-Myb or PU.1 in myelomonocytic lineage-specific expression. "*

(From PMID 9136080 annotated by the BioNLP'09 corpus annotators)

This sentence contains two molecular events: a regulation event (with trigger *"role"*) and a gene expression event (with trigger *"expression"*). The regulation event is negated, whereas the gene expression event is affirmative.

Figure 3.19 shows the partial parse tree of the sentence. It can be observed that the constituency parse tree distance between the negation cue *"not"* and the trigger of the regulation event *"role"* is equal to 4, and the constituency parse tree distance between the negation cue and the trigger of the gene expression event *"expression"* is equal to 5.

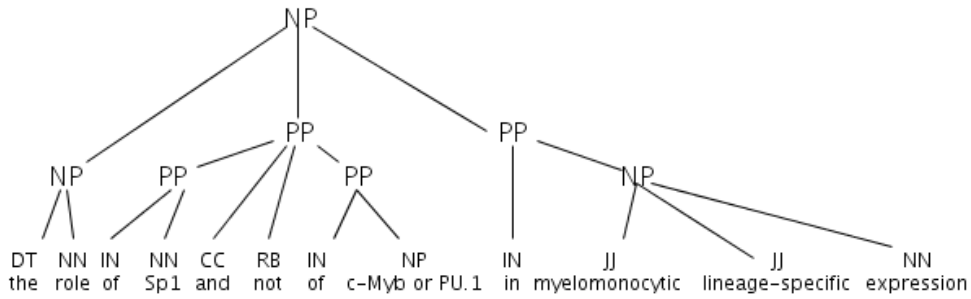


Figure 3.19: Partial constituency parse tree showing the trigger-cue distance

Partial constituency parse tree of the sentence in Example 3.15. The negation cue *"not"* has a distance of 4 with trigger *"role"*, and a distance of 5 with trigger *"expression"*.

There is only one negation cue in this sentence, therefore it will be considered as the main cue for both of the events. However, the features of the two events with regard to negation will be different, as their trigger-cue

distances are not the same.

Note that since negation and speculation detection is performed completely independently, it does not cause any complexities if a sentence contains a mixture of negation cues and speculation cues.

3.4.2 Negations with command rules

The command relation was introduced in Section 2.3.3. In this research, we only require condition 2 for command relation to hold, namely, node *a* 'commands' another node *b* if the *S*-node that most immediately dominates *a* also dominates *b*.

We hypothesised that if a negation cue has some command relationship with an event component, then the associated event could be negated. To test this hypothesis, we developed a rule-based system and experimented with three possible rules. An event is considered negated if either of the following conditions hold.

- the negation cue commands any event participant in the parse tree; or
- the negation cue commands the event trigger in the tree; or
- the negation cue commands both.

To determine whether token *a* X-commands token *b*, given the parse tree of a sentence, we use an algorithm introduced by (McCawley 1993), tracing up the branches of the constituency parse tree from *a* until a node that is labelled *X* is reached. If *b* is reachable by tracing down the branches of the tree from that node, then *a X*-commands *b*; otherwise, it does not.

Figure 3.20 displays a parse tree of Example 3.16.

> **Example 3.16.** *"We now show that a mutant motif that exchanges the terminal 3' C for a G fails to bind the p50 homodimer [...]"*
> (From PMID 9442380)

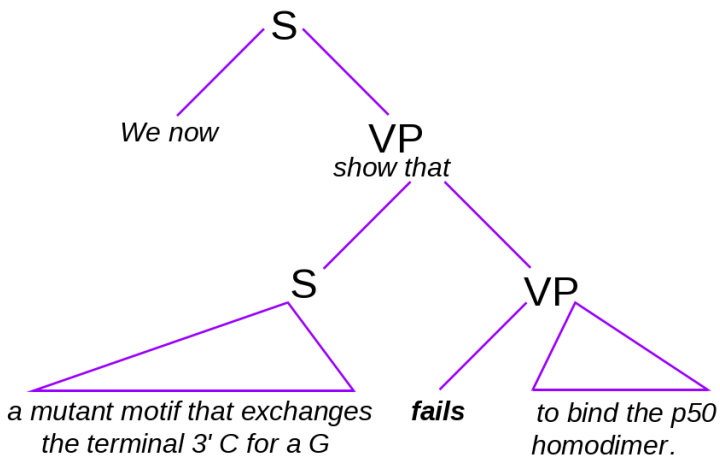


Figure 3.20: Command relation detecting a negated event

The simplified parse tree of the sentence "We now show that a mutant motif that exchanges the terminal 3' C for a G fails to bind the p50 homodimer." The negation cue "fails" VP-commands one of the event triggers, "bind" and therefore causes that event to be negated. There was no such command relation between the other trigger, "exchanges", and therefore the respective event is not affected.

This figure shows how the command relation can signal the affected part of the sentence. The negation cue, "*fails*", VP-commands the event trigger *bind*, and therefore indicates that the associated event is negated. However, another event trigger in the sentence, "*exchanges*", is not affected by this cue and therefore is not negated. The main verb of the sentence, *show*, is also not commanded by the cue and is therefore affirmative, as can be easily verified.

The results of this rule-based approach will be presented in Chapter 4.

3.4.3 Extracting negations—a machine learning approach

Here we explain another approach to negation detection, using machine learning to increase the performance of the rule-based method.

Given a sentence that describes an event, we further construe the negation detection problem as a classification task: the aim is to classify the event as affirmative or negative. For this purpose, we use an SVM (support

vector machine) classifier. In recent years, many machine learning algorithms have been implemented into efficient and customisable tools and have been made publicly available. These tools have been exploited for a wide range of applications in the areas of text mining and data mining.

The machine learning tools that have been helpful to our research are

- SVM light: Support Vector Machine implementation in C (Joachims 1999)
- SVM perf: an optimisation of the SVM, specifically for binary classification (Joachims 2006)

We explore two approaches: (1) using a single SVM classifier modelling negation for all events together; and (2) using three separate SVM classifiers, each one modelling negation for each of the event classes I, II, and III.

In the first experiment, the following features were engineered from an event-representing sentence. These features are common across all classes.

Lexical features:

1. Whether the sentence contains a negation cue from the cue list;
2. The negation cue stem (if present);
3. The part-of-speech (POS) tag of the negation cue;
4. The POS tag of the event trigger;
5. The POS tag of the theme of the event; if the theme is another event, the POS tag of the trigger of that event is used

Syntactic features:

6. The constituency parse node type of the lowest common ancestor of the trigger and the cue (i.e. the type of the smallest phrase that contains both the trigger and the cue, e.g. S, VP, PP, etc.);
7. Whether or not the negation cue commands any of the participants; nested events (for Class III) are treated as above (i.e. as being represented by their triggers);

8. Whether or not the negation cue commands the trigger;
9. The constituency parse-tree distance between the event trigger and the negation cue.

Semantic feature:

10. Event type (one of the nine types as defined in BioNLP'09);

Note that only one feature depends on the event type. We use a default value (null) where none of the other values apply (e.g. when there is no cue in feature 3 and 4). When there are more than one cue in the sentence, the main cue as described above (i.e. the cue that has the shortest constituency distance to the event trigger in the syntactic parse of the sentence) is considered. Therefore, different events in the same sentence might be affected by different negation cues.

In the first experiment, we train a single classifier on the whole training set, adding features incrementally and observing the effect of every added feature. Once the best set of features have been identified, we evaluate the trained model on each class separately.

In the second experiment, we train different models on the same common features for each class. Finally, we train three separate classifiers with class-specific feature sets.

Negation in regulation events

The structure of the regulatory events allows different number and types of participants (entities and events), as well as different participation type (theme and cause). They are the most common events in the literature and any improvement in the detection of negated and speculated regulatory events will have a considerable impact on the overall performance of the system. For these reasons, and to explore class-specific features in our approach, we decided to further analyse the regulation events and explore the effects of different

features on the detection of the negated events.

For this purpose, we designed the following experiments on the regulation events (class III).

Our target events are regulatory processes and causal relations between different biomedical entities and processes. Each regulatory event expressed in text is identified by:

- regulation type—we consider three regulation sub-types: positive regulation and negative regulation, in addition to regulation events where there is no indication if it is positive or negative;
- regulation theme represents an entity or event that is regulated;
- regulation cause—a protein or event that causes regulation;
- event trigger—a token(s) that indicates presence of the event in the associated sentence.

Lexical features are based on a list of negation cues and part-of-speech (POS) tagging of the associated sentence. We also consider the surface distance between the negation cue and trigger, theme and cause. More precisely, the lexical features include:

1. Whether the sentence contains a negation cue from the cue list;
2. The negation cue itself (if present);
3. The POS tag of the negation cue;
4. The POS tag of the trigger;
5. The POS tag of the theme; if the theme is another event, the POS tag of the trigger of that event is used;
6. The POS tag of the cause; if the cause is another event, the POS tag of the trigger of that event is used;
7. Surface distance between the trigger and cue;
8. Surface distance between the theme and cue;
9. Surface distance between the cause and cue;

Syntactic features are based on the results of constituency parsing of the associated sentence and the command relation. We explored various types of X-command, including S-command (for sentence or sub-clause), NP-command (noun phrase), VP-command (verb phrase), PP-command (prepositional phrase), etc. We also consider the distances of the event components within the tree. Specifically, the syntactic features include:

10. The type of the lowest common ancestor of the trigger and the cue (either S, VP, PP, NP, JJ or PP);
11. Whether or not the negation cue X-commands the trigger (X is S, VP, NP, JJ, PP)
12. Whether or not the negation cue X-commands the theme (X is S, VP, NP, JJ, PP)
13. Whether or not the negation cue X-commands the cause (X is S, VP, NP, JJ, PP)
14. The constituency parse-tree distance between the event trigger and the negation cue.
15. The constituency parse-tree distance between the theme and the negation cue.
16. The constituency parse-tree distance between the cause and the negation cue.

Semantic features introduce known characteristics of the regulation participants and the sub-type of regulation (if known):

17. Regulation sub-type (positive, negative, none);
18. Theme type, which can be either a protein or one of the nine event types as defined in BioNLP'09: gene expression, transcription, protein catabolism, localization, phosphorylation, binding, regulation, positive regulation, and negative regulation;
19. Cause type is defined analogously to the theme type.

The above features have been used to train a number of binary SVM (support vector machine) classifiers that aim to identify negated regulation events.

3.5 From negations to hedges

In the BioNLP'09 corpus, slightly more than 5% of the events are annotated as speculated. Two instances are given in Example 3.17. Note that the second sentence is both negated and speculated.

Example 3.17.

(a) *"However, it was not possible to ascertain whether Bcl-2 upregulation was a specific consequence of LMP1 expression."*

(From PMID 7520093 annotated by the BioNLP'09 corpus annotators)

(b) *"CD28-dependent elevation of c-jun mRNA does not appear to be mediated at the level of mRNA stability."*

(From PMID 7989745 annotated by the BioNLP'09 corpus annotators)

The question of whether an event is reported speculatively in text has several characteristics in common with that of negated events. They can both be construed as classification problems; they both have a cue, and the cue affects a part of the sentence (scope) within which the event may be expressed. If some part of an event is described in the part of the sentence that falls in the scope of the negation or speculation cue, the event would probably be negated or speculated.

Negmole was mainly developed and tested for the detection of negations of the molecular events expressed in text. However, we hypothesised that the negation detection methods were not specific only to this task, and could be expanded to detect similar semantic characteristics about these events, such as speculation.

To test this hypothesis, similarly to negation, we construed the

speculation detection problem as a classification problem that would classify extracted molecular events into speculative and assertive categories. We noted that an event can be independently speculated or negated, so the results of one classifier need not affect the other.

We chose the experimental setting with the best negation detection results and modified it for speculation detection. We used separate SVM models with class-specific feature sets to train on speculation data, as this setting resulted in the highest performance for negation detection (see Section 4.4.1).

We used the same syntactic and semantic features in the speculation detection task as the negation detection task. The lexical features were customised by adding a speculation cue list to replace the list of negation cues (See Table 3.4).

3.6 Summary

In this chapter we presented methods for event extraction (Evemole), a hybrid rule-based and machine learning approach for detecting molecular events.

We also presented Negmole to detect negated molecular events. The negation extraction system was expanded to detect information regarding statements and findings that are reported speculatively.

Chapter 4
Evaluation of event extraction and contextualisation

In this chapter we describe the evaluation approach used to evaluate the methods for molecular event extraction and negation and speculation detection. Subsequently, the results and evaluation of these methods are presented and discussed.

4.1 Evaluation method

4.1.1 Evaluation metrics and approach

The standard metrics precision, recall and F1-measure (introduced in Section 2.7.1) were used to evaluate the results of the methods which were presented in Chapter 3 . In the event extraction task, a true positive instance represents an event that is correctly identified. For this purpose, we need to determine whether the manually annotated event corresponds to the event extracted automatically. Due to the complex nature of the events, event equality as previously discussed in Section 3.1.3 is not trivially defined.

In the BioNLP'09 Shared Task, a number of definitions for event equality was used by the organisers to provide different levels of flexibility in the evaluation. In all cases it was required for the extracted event and the gold event to share the same type, trigger, participants, and participation type (i.e. theme/cause). If a participant is another event, those events should also match recursively. In the strictest evaluation case, for the two triggers to be the same, it was required that their textual boundaries exactly match. In a more relaxed evaluation methods, **approximate span matching**, the extracted triggers need only fall within an extension of the gold trigger span, by one word to either side of the trigger.

As regulation events can take other events as arguments and therefore

defining recursive equality can become very complicated, **approximate recursive matching** was defined to let the arguments of a regulation event to be only partially correct if they are event themselves. For partial matching, only theme arguments (as well as trigger and type) were considered. Cause arguments could be missing, incorrectly assigned, or redundantly extracted. Here, we also use approximate span matching, allowing the gold and extracted sub-strings to overlap. But we report exact boundary evaluation wherever appropriate.

In the evaluation of Evemole (introduced in Chapter 3) we consider an extracted event as a true positive instance if all of the following criteria hold:

1. The extracted trigger matches the gold trigger, approximately matching boundaries;
2. The extracted event type is the same as the gold event type;
3. The participation types match (theme or cause); and
4. If any of the participants is an event, it must also be a true positive, defined recursively.

Moreover, we require the matches to happen at the mention level. So, for example, if a sentence contains more than one mention of a certain entity, and the gold standard annotations consider one of these mentions as the participant of an event, we require the exact same mention of the entity (with correct start and end indices) to be assigned to the extracted event.

> **Example *4.1.*** *"Nuclear run-on assays and mRNA stability studies demonstrated that M-CSF regulates c-jun expression by both an increase in transcription rate and a prolongation in the half-life of c-jun transcripts."*
> (From PMID 1712226 annotated by the BioNLP'09 corpus annotators)

In Example 4.1, the gene "*c-jun*" is mentioned twice. Also, two events

involving *c-jun* are reported. One is the expression of *c-jun*, indicated by the trigger "*expression*", and the other is its transcription, indicated by the trigger "*transcription*". However, although these facts may be sufficient information extraction for a biologist, in this NLP evaluation we require the *correct mention* of *c-jun* to be associated with the extracted events. Specifically, for the events to be considered true positive, the first mention of *c-jun* must be detected as the participant of the gene expression event, and the second mention as that of the transcription event.

4.1.2 Evaluation corpora

The corpus used for the evaluation of event extraction, negation, and speculation is the BioNLP'09 gold annotated development corpus, described in Section 2.6. The corpus is derived from the GENIA corpus, with some modifications to restrict the data to molecular events between gene or protein entities (see Section 2.6).

However, as part of the BioNLP'09 Shared Task, the organisers reported the results of system evaluations on another "test" data set. The BioNLP'09 Shared Task test data set was a set of 260 abstracts from a subset of the GENIA corpus without publicly available gold annotations for events. The test data set was used to test the performance of the participants of the Shared Task.

The BioNLP'09 assessment was based on the output of the system when applied to this test dataset of 260 previously unseen abstracts. This data set, similarly to the training and development data sets, had manual annotations for gene and protein entities.

We use this corpus, along with the other two BioNLP'09 corpora (training and development data sets) in the evaluation of event extraction methods described in Chapter 3.

4.2 Evaluation of event extraction

The overall F-score of Evemole on the unseen test data was 30.35% with 48.61% precision using approximate boundary matching for the triggers (see

Table 4.1). The best performing event types were phosphorylation (the best F-score and the best recall) and gene expression (the best precision with a reasonably good F-measure).

Event Class	#Gold	R	P	F-score
Localisation	174	44.83	53.06	48.60
Binding	347	12.68	40.37	19.30
Gene expression	722	52.63	69.34	59.84
Transcription	137	15.33	67.74	25.00
Protein catabolism	14	42.86	50.00	46.15
Phosphorylation	135	78.52	53.81	63.86
Non-regulatory total	**1529**	**41.53**	**60.82**	**49.36**
Regulation	291	3.09	19.15	5.33
Positive regulation	983	1.12	8.87	1.99
Neg. regulation	379	12.4	20.52	15.46
Regulatory total	**1653**	**4.05**	**16.75**	**6.53**
All total	**3182**	**22.06**	**48.61**	**30.35**

Table 4.1: Evaluation of Evemole on the BioNLP'09 test data

The evaluation is reported on 260 abstracts, using approximate boundary matching criteria. #Gold refers to the number of instances in the gold standard data set.

An analysis of the results was performed on the development data, which had around 5% higher overall F-score than the test data (9% for non-regulation events, see Table 4.2 for details).

The CRF parameters were adjusted for maximum performance on the development corpus, including the choice of training algorithms (chain CRF linear, conjugate gradients, back propagation and other neural network models, etc.), the number of training steps, the size of the window within which the tokens can affect any given token, and the number of training abstracts used in each training step. It was interesting to observe that there were no significant

improvements in the performance after training on 100, 400 or 800 abstracts from the training set, suggesting that the model is already stable after training on the first 100 examples (data not shown).

Event Class	#Gold	R	P	F-score
Localization	53	67.92	46.75	55.38
Binding	312	21.47	63.81	32.13
Gene expression	356	64.61	76.33	69.98
Transcription	82	53.66	89.80	67.18
Protein catabolism	21	90.48	67.86	77.55
Phosphorylation	47	91.49	53.09	67.19
Non-reg total	**871**	**50.4**	**68.44**	**58.05**
Regulation	172	5.23	33.33	9.05
Positive regulation	632	3.48	21.36	5.99
Neg. regulation	201	9.45	15.08	11.62
Regulatory total	1005	4.98	19.53	7.93
All total	**1876**	**26.07**	**54.46**	**35.26**

Table 4.2: Evaluation of Evemole the BioNLP'09 development data

The evaluation is reported on 150 abstracts using approximate boundary matching criteria. #Gold refers to the number of instances in the gold standard data set.

4.3 Event extraction discussion

The overall F-score for Evemole was 30.35% with 48.61% precision on the previously unseen test data (see Table 4.1 for details). The best performing event types were *phosphorylation* (the best F-score and the best recall) and *gene expression* (the best precision with a reasonably good F-measure).

While the results for non-regulatory events (classes I and II) were encouraging, they were low for regulatory events (class III). Among the 24 teams submitting the test results, our results were ranked 12th for the overall F-

score and 8th for the F-score of non-regulation events, suggesting that improvement in the detection of class III events could result in the overall improvement of the system. A summary of the results of all the participating teams can be found in Table 4.3.

Team	R	P	F
UTurku	46.73	58.48	51.95
JULIELab	45.82	47.52	46.66
ConcordU	34.98	61.59	44.62
UT+DBCLS	36.9	55.59	44.35
VIBGhent	33.41	51.55	40.54
Utokyo	28.13	53.56	36.88
UNSW	28.22	45.78	34.92
Uzurich	27.75	46.6	34.78
ASU+HU+BU	21.62	62.21	32.09
Cam	21.12	56.9	30.8
Uantwerp	22.5	47.7	30.58
UNIMAN (Evemole)	**22.06**	**48.61**	**30.35**
SCAI	25.96	36.26	30.26
Uaveiro	20.93	49.3	29.38
Team 24	22.69	40.55	29.1
Uszeged	21.53	36.99	27.21
NICTA	17.44	39.99	24.29
CNBMadrid	28.63	20.88	24.15
CCP-BTMG	13.45	71.81	22.66
CIPS-ASU	22.78	19.03	20.74
Umich	30.42	14.11	19.28
PIKB	11.25	66.54	19.25
Team 09	11.69	31.42	17.04
KoreaU	9.4	61.65	16.31

Table 4.3: The results of all the teams in the BioNLP'09 Shared Task

An analysis of our results was performed on the development data, which had around 5% higher overall F-score than the test data (9% for events of classes I and II, see Table 4.2 for details).

To further explore the factors affecting the results, we define the **lexical**

variability of each event type to be the number of different word stems used to describe that event divided by the total number of mentions in the text that trigger that event. This measure is akin to the well-know measure of **lexical variation** of a document which refers to the number of word types divided by the number of word tokens in a document.

For example, there are a total number of 47 phosphorylation events in the development data, triggered by 40 different mentions. The reason for the disparity is that sometimes, as in Example 4.2, one mention triggers several events. The sentence in Example 4.2 is counted three times in the total number of events, but only once in the total number of triggers.

> **Example 4.2.** *"Phosphorylation of the IkappaB cytoplasmic inhibitors, IkappaBalpha, IkappaBbeta, and IkappaBepsilon, by these kinases triggers proteolytic degradation and the release of NF-kappaB/Rel proteins into the nucleus."*
> (From PMID 9804806, annotated by the BioNLP'09 corpus annotators)

Phosphorylation events have been described in the development data using the following forms: *phosphorylation*, *phosphorylate*, *phosphorylates*, *phosphorylation sites*, *underphosphorylated form*, and the capitalised *Phosphorylation*. There are 5 different terms—ignoring the capitalised variation—triggering the event phosphorylation. If we bundle up all of those mentions which have the same stem, we will end up with only the following 3 stems: *phosphoryl*, *underphosphoryl form*, and *phosphoryl site*.

To calculate the lexical variability of an event type, we divide the number of different stems of the triggers by the total number of mentions for that event type.

$$V[type] = \frac{number\ of\ different\ trigger\ stems}{number\ of\ trigger\ mentions}$$

The lower this number is, the less lexically variable the triggers for that event type are. The lexical variability of different types are calculated and presented in Table 4.4.

Type	# Events	# Trigger mentions	# Distinct triggers	# Distinct trigger stems	Lexical variability	Confusion
Gene expression	356	282	49	37	0.13	0.122833
Transcription	82	68	22	18	0.26	0.259023
Protein catabolism	21	19	4	3	0.16	0.020202
Localisation	53	40	15	13	0.33	0.112195
Phosphorylation	47	40	5	3	0.08	0
Binding	249	180	51	33	0.18	0.01037
Regulation	173	138	63	43	0.31	0.097902
Positive regulation	618	462	164	111	0.24	0.059021
Negative regulation	196	153	86	62	0.41	0.008011

Table 4.4: The lexical variability of the triggers with respect to interaction type

In order to measure how characterisable an event type is with regard to the triggers that are used to express events of that type, we consider the different types a specific trigger can refer to as different *senses* of the trigger, and use a weighted sum of the word sense entropy of the trigger to define the **characterisability** of the type.

Shannon's entropy H of a discrete random variable X with possible values $\{ x_{1,}\cdots,x_n \}$ and probability mass function p is defined as the following.

$$H[X]=\sum_{i=1}^{n} p(X_i)\log_2(\frac{1}{p(X_i)})$$

We define the word sense entropy of a given trigger based on Shannon's

entropy as the following.

$$H[trigger] = -\sum_{all\,types} \left(\frac{\#(trigger, type)}{\#(trigger, *)} . \log_2\left(\frac{\#(trigger, type)}{\#(trigger, *)}\right)\right)$$

In this equation, $\frac{\#(trigger, type)}{\#(trigger, *)}$ refers to the number of times that a trigger is used to refer to an event of a certain *type* divided by the total number of times it is used for all different event types. The asterisk represents all different event types.

Moving on from the above definition of the entropy for every trigger, we define the entropy of a type as a weighted sum over the triggers of that type.

$$H[type] = \sum_{trigger \in type} H[trigger] . \frac{\#(trigger, type)}{\#(*, type)}$$

And finally, we introduce the following, with a similar idea as tf-idf, as a measure of **confusion**, or how un-characterisable a class is with regard to trigger term.

$$C[type] = \sum_{trigger \in type} \left(\frac{\#(trigger, *-type)}{\#(trigger, *)} . \frac{\#(trigger, type)}{\#(*, type)}\right)$$

The confusion measures as described above for the 9 event types are displayed in the last column of Table 4.4. As can be seen in Table 4.4, phosphorylation triggers have the lowest lexical variability, and zero confusion. This observation explains the high quality of trigger detection for this type despite the relatively small number of training instances. Low lexical variability means that there is a relatively low chance of a word with a previously unseen stem act as the trigger of a phosphorylation event and therefore be missed, resulting in a false negative instance. Low confusion

means that there is a relatively low chance that a word that could potentially be the trigger for a different event type (or a non-trigger), be mistakenly recognised as a phosphorylation trigger, resulting in a false positive instance.

Gene expression is the most frequent event type. Despite the relative high confusion measure of the gene expression trigger terms, the low lexical variability together with the high frequency of instances could have been responsible for the higher accuracy of the trigger identification by the CRF module.

The type-specific performance could be inversely related to the lexical variability and confusion of the trigger terms as well as the low frequency of the type. Specifically, analysing our trigger recognition results suggests that recall is negatively correlated with the lexical variability and precision is negatively correlated with confusion. Figures 4.1 and 4.2 show such correlations in our trigger evaluation results.

An analysis of the overall results of the other teams[18] showed some correlation between the average recall and lexical variability ($R^2 = 0.68$) but no such correlation between confusion and the average precision ($R^2 = 0$). The coefficient of determination, R^2, is the square of the correlation coefficient, and is used as a measure of correlation (Mendenhall et al. 2009).

We do not have access to the trigger-only evaluation of the other systems, and therefore cannot measure the exact correlation. Future work is needed to generalise these finding, and find more accurate measures involving other factors such as term frequencies.

18 A summary of the results of all the teams can be accessed at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/results/results-master.html

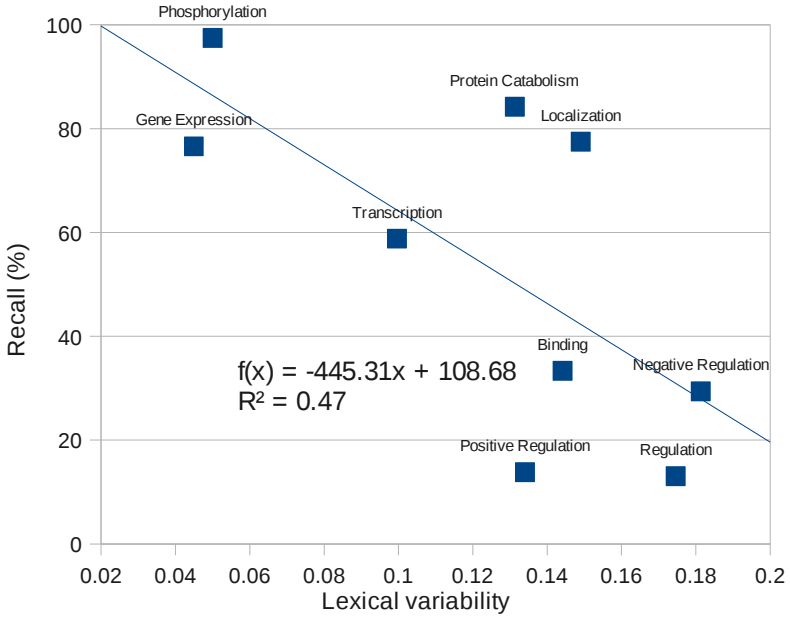


Figure 4.1: Correlation between recall and lexical variability for event types

The linear regression equation and the correlation coefficient are shown as f(x) and R.

These correlations can potentially be used to predict the results of an information extraction task before accomplishing it, and set theoretical upper and lower bounds on the results of a tested system on new data, without having to apply the system to the data, and only by looking at the distribution and the properties of the data.

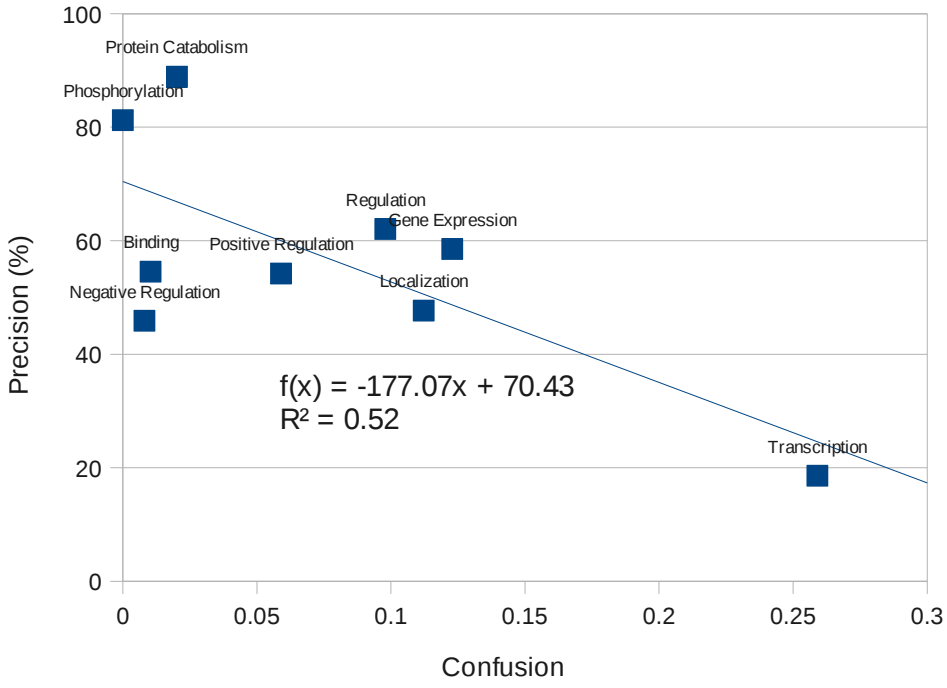


Figure 4.2: Correlation between precision and confusion for event types

The linear regression equation and the correlation coefficient are shown as f(x) and R.

We will now show that the accuracy of the trigger detection is directly related with the overall performance of the event extraction.

In order to assess the effects of different steps in our approach, we evaluated the performance of the event trigger and event participant detection steps separately. The results presented in Table 4.5 show the trigger-only evaluation before the participants are associated to form the events. These results indicated that the performance of the trigger detection (CRF) module was not much better than the overall performance of the system (an F-score of 43% vs. 35%), suggesting that the CRF module for trigger detection was mostly responsible for the errors, by both missing triggers and falsely reporting them. This was particularly the case with class I and even class II events, but less so for class III events.

Conversely, when considering only those events whose triggers were correctly identified, their participants were also correctly recognised in most

cases. Overall, the analysis suggested that the parse tree distance method performed reasonably well, despite a reduction in recall of approximately 12%.

Event Class	#Gold	R	P	F-score
Localisation	40	77.50	47.69	59.05
Binding	180	33.33	54.55	41.38
Gene expression	282	76.60	58.54	66.36
Transcription	68	58.82	18.60	28.27
Protein catabolism	19	84.21	88.89	86.49
Phosphorylation	40	97.50	81.25	88.64
Non-reg total	**629**	**63.91**	**48.73**	**55.30**
Regulation	138	13.04	62.07	21.56
Positive regulation	462	13.85	54.24	22.07
Negative regulation	153	29.41	45.92	35.86
All total	**1382**	**38.28**	**49.44**	**43.15**

Table 4.5: Trigger-only evaluation on the BioNLP'09 development data

The performance of only trigger and type detection on the development data. #Gold refers to the number of instances in the gold standard data set.

There are a number of possibilities for improvements. We believe applying the CRF model for trigger detection in two stages would be a better approach to detect events: first identify triggers (binary classification) and then classify them into different types. In addition, the rules employed for determining themes need to be more specific to reflect both event type and grammatical structure.

In the case of class III events, however, significantly better results were noticed in the trigger detection part when compared to the overall scores, indicating that it was difficult to identify regulatory participants, as any of those participants could be either a protein or another event, and our rules did not clearly discriminate between the participation type (theme and cause)

which resulted in incorrect output.

Overall, the results achieved by Evemole suggest that combining parse tree results, rules and CRFs is a promising approach for the identification of non-regulatory events in the literature, while more work would be needed for regulatory events.

4.4 Evaluation of negation and speculation detection

Our method for negation and speculation detection was initially developed and analysed for negation extraction only, and later was adopted to also detect speculations by training a new model on speculation data and using a new dictionary of speculation cues (see Section 3.4.1). In this section we present detailed evaluation and analysis on the negation detection task, and also provide a performance report on the speculation detection task.

4.4.1 Evaluation of negation detection

We use the BioNLP'09 training and development corpora for the evaluation of Negmole. To be able to evaluate the negation detection system as a separate, stand-alone system, we use the gold annotations for entity mentions (genes and proteins) and gold annotated molecular events. Sentences that report molecular events are annotated with the corresponding event type, textual trigger and participants. Every event in both training and development data sets has been tagged as either affirmative (reporting a specific interaction) or negative (reporting that a specific interaction has not been observed). We use this information to train and test our system.

Table 4.6 provides an overview of the two BioNLP'09 data sets with regard to negated and speculated events.

Event class	Training data			Development data		
	total	negated	speculated	total	negated	speculated
Class I	2,858	131	106	559	26	15
Class II	887	44	29	249	15	8
Class III	4,870	440	320	987	66	71
Total	8,615	615	455	1,795	107	95

Table 4.6: Negated and speculated events in BioNLP'09 corpus

Overview of the composition of the negated and speculated events in the training and development datasets of the BioNLP'09 corpus

Baseline methods

To compare Negmole, we considered two baseline methods and calculated the precision, recall, F-score and specificity (see Section 2.7.1) for them. The analysis was done using the gold-annotated events on the BioNLP'09 development data set.

As can be inferred from Table 4.6, only around 6% of the gold annotated events are negated. Therefore, if no negation detection at all was performed, a specificity of 94% would be achieved. We considered the case where any event described in a sentence with a negation cue is marked as negated. In addition, we implemented the NegEx algorithm (see Section 2.4.2) using event triggers as the list of terms. The results of the two baseline methods are shown in Table 4.7.

Approach	P	R	F1	Specificity
Any negation cue present	20%	**78%**	32%	81%
NegEx	**36%**	37%	**36%**	**93%**

Table 4.7: Baseline measures

Two different baseline measures evaluated on the BioNLP'09 development data.

The first baseline method (bag-of-words) has a high recall, indicating

that 78% of the negated events have a negation cue somewhere in the sentence. However, as expected, the precision of such a method is low, as the presence of a negation cue does not necessarily indicate that every event in that sentence is negated. NegEx has a lower recall, which is due to missing the instances where the trigger has a longer surface distance from the negation cue. The F-score of both baseline methods are in the range of 30%.

Rule-based method

Here we present the results of our rule-based experiments with the command relation. First we only considered the *S*-command relation as it was Langacker's original definition of the command relation. We marked as negated any event where the negation cue in the sentence S-commanded any of the participants. Then we marked as negated any event where the negation cue S-commanded the event trigger. Finally, we required both conditions to hold in order for an event to be marked as negated. The results are shown in Table 4.8. The highest F-score was achieved in the third case, but the differences were small.

Approach	**P**	**R**	**F1**	**Specificity**
Negation cue commands any participant	23%	**76%**	35%	84%
Negation cue commands the trigger	23%	68%	34%	85%
Negation cue commands both	23%	68%	**35%**	**86%**

Table 4.8: Evaluation of negation rules on the BioNLP'09 data

Performance is reported on the BioNLP'09 development data set when only the S-command relation is used. The numbers are rounded.

These results show that using the command relation as a rule dramatically increases the recall compared to NegEx, suggesting that the command relation can successfully reach beyond the scope of NegEx. However, the relatively lower precision means that the F-score is at a similar level (or lower) compared to NegEx. The precision is lower than that of

NegEx, and stays the same over the applications of different rules. This shows that there are many cases in which a command relation between the negation cue and components of the event exists, but other factors make the event not affected by the cue.

These observations suggest that although the command relation is not very affective as a stand-alone rule, it could work as a predictor of negated events if combined with other features. Furthermore, they suggest that participants of an event play as important a role as the trigger in the negation of the event as a whole. This bring us to the following experiments where we used these findings to design a series of machine learning experiments using the command relation as a feature along with other features.

Machine learning experiments for negation detection

As explained in Section 3.4.3, initially we combined all the features that were not class-specific and trained a single classifier on the whole training dataset. Secondly, we used the common features, but trained different classifiers for the three different classes of events and acquired three different models (one for each class). Finally, we added class-specific features to the classifiers and trained three different models with different number and types of features. We report the results of these experiments, as well as the effects of each group of features on the regulatory events (class III). All the evaluations are reported on the BioNLP'09 development data set. For a summary of the experiments and the features used in each one, see Table 4.9.

Experiment	Classifier	Features
Experiment 1	Single classifier	1. Whether the sentence contains a negation cue from the cue list; 2. The negation cue stem (if present); 3. The part-of-speech (POS) tag of the negation cue; 4. The POS tag of the event trigger; 5. The POS tag of the theme of the event; if the theme is another event, the POS tag of the trigger of that event is used 6. The parse node type of the lowest common ancestor of the trigger and the cue (i.e. the type of the smallest phrase that contains both the trigger and the cue, e.g. S, VP, PP, etc.); 7. Whether or not the negation cue commands any of the participants; nested events (for Class III) are treated as above (i.e. as being represented by their triggers); 8. Whether or not the negation cue commands the trigger; 9. The parse-tree distance between the event trigger and the negation cue; 10. Event type (one of the nine types as defined in BioNLP'09).
Experiment 2	3 class-specific classifiers on common features	Same as Experiment 1
Experiment 3	Class-specific classifiers on class-specific features: Classes I and II use features 1-8, 10-12, 14, 15. Class III uses all the features.	1. Whether the sentence contains a negation cue from the cue list; 2. The negation cue itself (if present); 3. The POS tag of the negation cue; 4. The POS tag of the trigger; 5. The POS tag of the theme; if the theme is another event, the POS tag of the trigger of that event is used; 6. The POS tag of the cause; if the cause is another event, the POS tag of the trigger of that event is used; 7. Surface distance between the trigger and cue; 8. Surface distance between the theme and cue; 9. Surface distance between the cause and cue; 10. The type of the lowest common ancestor of the trigger and the cue (either S, VP, PP, NP, JJ or PP);

		11. Whether or not the negation cue X-commands the trigger (X is S, VP, NP, JJ, PP); 12. Whether or not the negation cue X-commands the theme (X is S, VP, NP, JJ, PP); 13. Whether or not the negation cue X-commands the cause (X is S, VP, NP, JJ, PP); 14. The parse-tree distance between the event trigger and the negation cue; 15. The parse-tree distance between the theme and the negation cue; 16. The parse-tree distance between the cause and the negation cue; 17. Regulation sub-type (positive, negative, none); 18. Theme type, which can be either a protein or one of the nine event types; 19. Cause type is defined analogously to the theme type.

Table 4.9: Summary of the experiments and the features used

A single classifier for all classes of the events (Experiment 1)

Table 4.10 shows the performance of a single classifier trained on the entire data, with common features added incrementally.

Feature set	**P**	**R**	**F1**	**Specificity**
Features 1-6 and 10	43%	8%	14%	99.2%
Features 1-7 and 10	73%	19%	30%	**99.3%**
Features 1-8 and 10	71%	38%	49%	99.2%
Features 1-10	**76%**	**38%**	**51%**	99.2%

Table 4.10: Evaluation of Experiment 1; the single SVM classifier method for negation detection on BioNLP'09

In this table, features 1-7 are lexical and POS-tag-based features. Feature 7 models whether the cue S-commands any of the participants. Feature 8 is related to the cue S-commanding the trigger. Feature 9 is the parse-tree distance between the cue and trigger. Feature 10 is the semantic feature related to the event type.

Table 4.11 shows the results of this method applied on each class, together with the micro-average across the entire development corpus.

Event class	Number of instances	P	R	F1	Specificity
Class I	26	82%	**54.00%**	**65%**	**97%**
Class II	15	**100%**	7%	13%	94%
Class III	66	73%	41%	52%	95%
Micro Average	**107**	79%	39%	50%	96%

Table 4.11: Class-specific evaluation of a single classifier for negation detection on BioNLP'09

The results of training a single classifier for negation detection on all classes using common features, but evaluated on individual classes. The evaluation is reported on the BioNLP'09 development data set.

Different models for each class, common features (Experiment 2)

Table 4.12 shows the results of training three classifiers with the same features as above, but on different classes separately. We note an increase in both precision (88%) and recall (49%) over the single-classifier approach.

Event class	Number of instances	P	R	F1	Specificity
Class I	26	94%	**65%**	**77%**	99.8%
Class II	15	**100%**	33%	50%	**100%**
Class III	66	81%	44%	57%	99.2%
Micro Average	**107**	88%	49%	63%	99.4%

Table 4.12: Evaluating separate classifiers trained on each class for negation detection on BioNLP'09

Here, again, common features have been used.

Class-specific features in different classifiers (Experiment 3)

Finally, we trained three classifiers with class-specific feature sets. Tables 4.13 and 4.14 show the results without and with semantic tokenisation. We note that there was some drop in the performance, both in terms of precision and recall.

Event class	Number of instances	P	R	F1	Specificity
Class I	26	**80%**	**62%**	**71%**	**97%**
Class II	15	75%	43%	57%	96%
Class III	66	79%	39%	53%	95%
Micro Average	**107**	**79%**	**45%**	**58%**	**96%**

Table 4.13: Evaluating separate classifiers without semantic tokenisation for negation detection on BioNLP'09

The results of training different negation classifiers with class-specific features on each class; ***without*** *semantic tokenisation.*

Event class	Number of instances	P	R	F1	Specificity
Class I	26	**88%**	**54%**	**67%**	**97%**
Class II	15	57%	29%	38%	95%
Class III	66	67%	46%	55%	95%
Micro Average	**107**	70%	46%	55%	94%

Table 4.14: Evaluating separate classifiers with semantic tokenisation on BioNLP'09

The results of training different negation classifiers with class-specific features on each class; ***with*** *semantic tokenisation.*

Amongst the three classes, the highest F-score was achieved on class I which contains the simplest event structures. The data set of class II events, with only 44 training instances and 15 test instances, was not large enough for making any specific conclusions.

Class III, however, was the largest and most complex of all three classes, and showed relatively lower results. To further explore the feature space and investigate the effect of each feature, we analysed several class-specific features on the regulation events. We chose that class due to its complex nature as well as frequency of mention and its biological prominence.

Evaluation on the regulation events

The training set contained a total of 4,870 regulation events, 440 of which are reported as negated. The test set contained 987 regulation events, of which 66 are negated. The training data was used for modelling and all the results refer to the methods applied on the development dataset using 10-cross validation.

Impact of lexical and other shallow features

The results of using lexical features only are presented in Table 4.15. Features 1-6 concern word forms and POS tags, whereas features 7-9 are surface distance features. See Table 4.9 and Section 3.4.3 for an extensive list.

As expected, surface distances to the negation cue are not good indicators, and do not improve the performance of standard lexical and POS features—on the contrary, they reduce precision. Overall, precision is relatively high but recall is low.

Lexical features	Precision	Recall	F1
Features 1-6 (no surface distances)	75.00	22.73	34.88
All lexical features	71.43	22.73	34.48

Table 4.15: Evaluation of negation detection on regulatory events using lexical features only

Impact of syntactic features

The results of using syntactic features only are presented in Table 4.16.

Features 10-13 are command-related features, and features 14-16 are parse-tree distance features. See Table 4.9 and Section 3.4.3 for an extensive list.

As opposed to surface distances, parse-tree distances are more suitable features, improving the overall performance significantly (F1 improving from 11% to 36%). There were no significant differences in performance when different types of X-command relations are used. Focusing only on S- and VP-command provides the same levels of accuracy as using all the other X-command features, with no statistically significant differences.

Syntactic features	Precision	Recall	F1
Features 10-13 (no parse-tree distances)	**80.00**	6.06	11.27
All syntactic features	60.71	**25.76**	**36.17**

Table 4.16: Evaluation of negation detection on regulatory events using syntactic features only

Impact of semantic features

The performance of the models based on the lexical and syntactic features were approximately the same, with no significant differences between the best performing feature subsets of each category. However, semantic features on their own resulted in very low performances, virtually missing all negated regulatory events (data not shown). This was comparable to the baseline model in which no negation detection was performed.

Combining features

Table 4.17 shows the results when features of various types are combined. Combining several feature types (lexical, syntactic and semantic) proved to be beneficial. Surface distances still reduce the overall precision, but overall improve recall. It is interesting that adding semantic features (which characterise the participants involved in the regulation) significantly improves precision (by 20% when compared to the lexical and syntactic feature sets). On the other hand, command relations improve recall (by almost 20%).

Features	Precision	Recall	F1
Lexical + syntactic	66.67	39.39	49.52
Lexical + semantic	50.00	15.15	23.26
Syntactic + semantic	72.22	19.70	30.95
All with no surface distances	73.68	**42.42**	**53.85**
All with no X-command on theme and cause	78.12	37.88	51.02
All features	**78.79**	39.39	52.53

Table 4.17: Evaluation of negation detection on regulatory events combining different features

We note that some feature subsets (e.g. features 10-13, Table 4.16) do not provide a balance between precision and recall; depending on the application, the classification threshold could be adjusted to produce higher recall or precision.

To conclude, training separate classifiers on different classes showed the best results. Although exploring the feature space provides insight into the effects of each type of feature, these experiments performed slightly worse than the previous ones. This could be due to implementational differences, the modest data size, or the use of semantic tokenisation.

4.4.2 Evaluating speculation detection

We adopted the method used for negation detection to classify speculated events. As using class-specific features and separate classifiers for each event class showed the best performance in negation detection, we applied this method to extract speculations. The results are shown in Table 4.18, and with semantic tokenisation in Table 4.19.

Event class	Number of instances	P	R	F1	Specificity
Class I	15	50%	29%	36%	**97%**
Class II	8	14%	14%	14%	95%
Class III	72	**73%**	**43%**	**54%**	95%
Micro Average	**95**	64%	38%	48%	95%

Table 4.18: Evaluating separate speculation classifiers without semantic tokenisation on BioNLP'09

The results of training different speculation classifiers with class-specific features on each class; ***without*** *semantic tokenisation*

Event class	Number of instances	P	R	F1	Specificity
Class I	15	**70%**	**50%**	**58%**	**98%**
Class II	8	17%	14%	15%	95%
Class III	72	64%	41%	50%	94%
Micro Average	**95**	61%	40%	48%	95%

Table 4.19: Evaluating separate speculation classifiers with semantic tokenisation on BioNLP'09

The results of training different speculation classifiers with class-specific features on each class; ***with*** *semantic tokenisation*

4.5 Negation and speculation detection discussion

Negation and speculation detection are challenging problems. There are numerous ways to express negation and speculation in language, both grammatically and lexically. Typically a given sentence bears several concepts, any number of which can be negated or speculated. Not all of these concepts are of interest or relevant to a given IE task, and some only serve as a figure of

speech, without claiming any negative or speculated claims. Still, failing to detect negated and speculated facts in IE could affect the quality of the extracted information.

We introduced a negation detection system, Negmole, for the biomedical domain. The method performs at the event level, addressing the issues that other approaches with higher granularity are faced with. The event-level approach assigns negation as an attribute to the smallest unit of meaningful information that can be extracted as a statement or a fact. It is a necessary step for performing reasoning (e.g. conflict detection) on the extracted data.

Negmole made use of an underlying linguistic phenomenon, the *command* relation, that had previously been suggested to be related to negation. We experimented with variations of this relation used in a rule-based setting and observed that, when regarded as a rule, the command relation performs at least as well as previously existing methods. This observation suggested that the command relation could serve as a highly indicative feature in a machine learning setting.

4.5.1 Cue detection

The BioNLP'09 data does not include annotations for negation and speculation cues. There are other data sets available that have such cues annotated, but they include all negation words, independently of what they are negating. For example, in Example 4.3, the word *not* indicates some negative concept (a property not being restricted to a cell type) but not a negated molecular event. It is correctly identified by Negmole as a negation cue that does not indicate a negated event.

> **Example 4.3.** "*This shows that transcription of both IL4 forms is not restricted to T cells and can be induced in other cell types as well.*"
> (From PMID 8603435, event and context extracted by BioContext)

Figure 4.3 shows the parse tree of Example 4.3. As can be seen in this figure, the word *"not"* is situated in the VP phrase *"is not restricted to T cells"* and only VP-commands the tokens inside that phrase. Therefore, as long as the relation VP-command is concerned, this word cannot affect any of the other parts of the sentence, including the event trigger *"transcription"* and event participant *"IL-4"*.

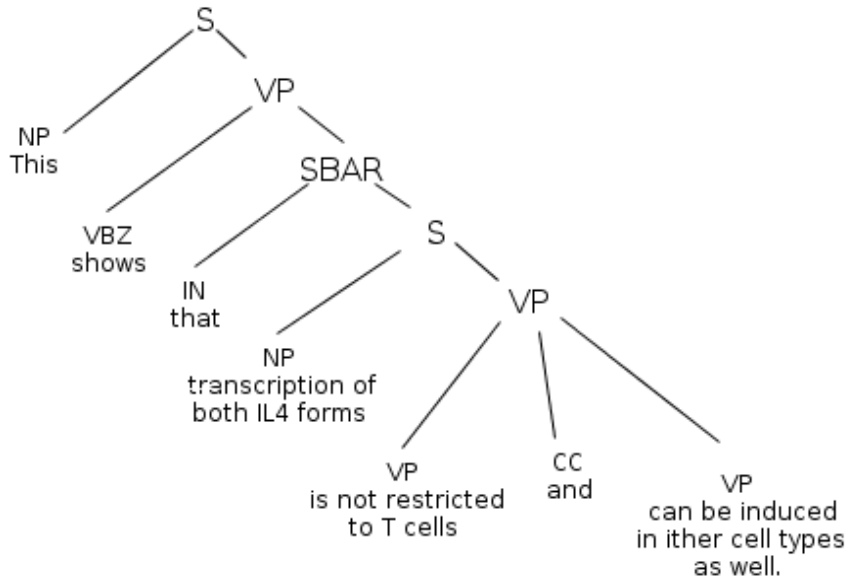


Figure 4.3: The parse tree of the example sentence

This parse tree shows that the negation cue word "not" is situated in the verb phrase "is not restricted to T cells" and therefore does not VP-command any other part of the sentence outside this verb phrase.

Tagging only negation words is a relatively easier task, as it can be construed as an NER task with relatively low levels of ambiguity and variation, and dictionary-based methods have previously shown reasonably high performance.

In the current task, our approach involves detecting only the events that are negated, rather than any negation within the sentence, and therefore not all negation cues are of interest. Cues could affect any concept expressed in text, a

small proportion of which are bio-molecular events.

Despite not being able to formally evaluate the performance of cue detection, manual examination of some examples shows that the method used to find the "main cue" amongst the group of cues in the sentence with more than one cue (see Section 3.4.1) performs as desired.

Example 4.4.

(a) *"However, **neither** induction of p53 in MCF-7 cells nor induction of p21 in either cell line was detected, suggesting that tamoxifen-induced RB dephosphorylation and apoptosis are independent of the p53/p21 pathway."*

(b) *"However, neither induction of p53 in MCF-7 cells **nor** induction of p21 in either cell line was detected [...]"*

(From PMID 9751262, extracted automatically by BioContext)

The event in Example 4.4*(a)* is the positive regulation (trigger: *"induction"*) of p53 in MCF-7 cell line, and the event in Example 4.4*(b)* is the positive regulation (trigger: second mention of *"induction"*) of p21 in the same cell line. They are both reported as negated and detected by Negmole as negated. Note that the negation cues that have affected each of the two events are not the same. The sentence has two negation cues: *"neither"* and *"nor"*. The first one, i.e. *"neither"*, has the shortest parse tree distance to the trigger of the first event (i.e. the first mention of *"induction"*), so is assigned as the main cue in the feature vector of that event. Similarly, *"nor"* is assigned to the second event.

Figure 4.4 shows the partial parse tree of Example 4.4, showing the single noun phrase containing *"neither induction of p53 in MCF-7 cells nor induction of p21 in either cell line"*. The parse tree distances of the two negation cues *"neither"* and *"nor"* with the two event triggers *"induction"* (two mentions) are shown in Table 4.20.

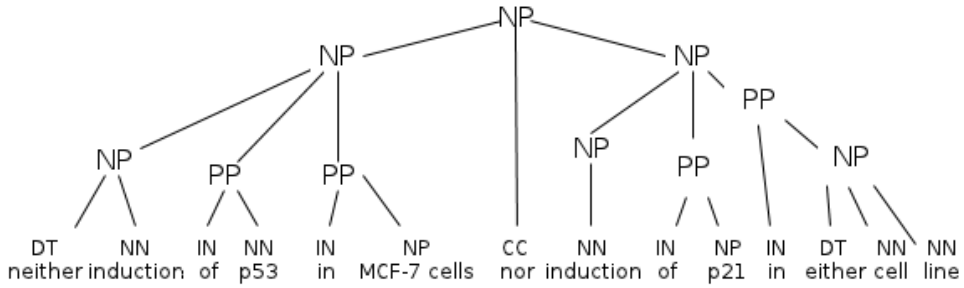


Figure 4.4: Parse tree of a phrase with two negated events.

The phrase sub-string appears in a single noun phrase: " neither induction of p53 in MCF-7 cells nor induction of p21 in either cell line"

	"neither"	*"nor"*
First mention of *"induction"*	**2**	4
Second mention of *"induction"*	6	**4**

Table 4.20: An example of parse tree distances between multiple negation cues and triggers.

The negation cue "neither" is associated with the first mention of the trigger "induction" and therefore with the first event (the positive regulation of "p53"). Similarity, the negation cue "nor" is associated with the second mention of the trigger "induction" and therefore with the second event (the positive regulation of "p21").

The detection of the negation and speculation cues were performed using a dictionary of cue stems derived from the training data. However, the distinction between a semantically negative event (e.g. negative regulation) and a negated event (e.g. a negated regulation event or a negated positive regulation event) is not always clear. Expert annotators have not been entirely consistent in differentiating between the two, and words such as "*block*" have served both as a trigger for a negative regulation event, and as an indication of a non-existent or negated event.

Although we have included such domain-specific ambiguous terms as

"block" and *"inhibit"* amongst the negation cues, comparing the negation cue distributions (Figures 3.16 and 6.7) shows that there was little confusion between the two senses of these words. Figure 3.16 shows that the word *"inhibit"*, for instance, is the second most common negation cue in the BioNLP'09 data. Figure 6.7 (which summarises negation cue distribution on the entire MEDLINE and PMC corpora) shows that not very often this presumably common word has been detected as the negation cue and caused an event to be negated. On the other hand, the type-specific event extraction results show that detecting negative regulation events are not particularly more challenging than other event types in class III.

Although it is very common for a sentence to have more than one negation or speculation cue, previous event-based approaches have not explicitly addressed the issue of handling multiple cues in a sentence. Negmole detects the specific cue mention responsible for the negation or speculation of an event, based on the parse-tree distances. Although no gold annotated evaluation data existed to evaluate this approach, Example 4.4 showed cases in which the correct negation cue has been associated with the event. Another such example where the negation cue is correctly identified amongst several cues in the sentence is shown in Example 4.5.

> **Example 4.5.** *"In contrast, there was no significant difference in force generation between old striae fibroblasts and normal fibroblasts with cells expressing* ***no*** *alpha-smooth muscle actin."*
> (From PMID 15883849, events and negation extracted by BioContext)

There are, of course, other cases in which an incorrect cue was assigned to the event, but the negation was still detected correctly. One such example is shown in Example 4.6.

> **Example 4.6.** *"CD4 and CD8, gamma/delta TCR bearing T cells and*

CD45R0 on CD4+ T cells as a marker for memory cells, on TL no differences could be detected between patients with or ***without*** *anti-TPO. ”*

(From PMID 8750571, events and negation extracted by BioContext)

In Example 4.6, the word *“no”* appearing before the gene extraction event trigger *“detected”* is causing the event to be negated. But the word *“without”* has been incorrectly marked as negation cue. This was an example of an error that does not lead to a false results.

Some of the errors were due to incorrectly identified negation cues. In Example 4.7, we see one such example that has contributed towards a false result.

Example 4.7. *“Taken together, it may be concluded that* ***NO*** *down-regulates IFN-gamma production mainly by inhibiting T-cell proliferation.”*

(From PMID 8806814, events and negation extracted by BioContext)

Here the word *“NO”* which is an abbreviated form of the chemical Nitrous Oxide has been incorrectly marked as negation cue due to its lexical similarity with *“no”*, specially after normalisation and stemming. Errors of this type could be addressed with the application of post processing rules, but as any other word sense disambiguation task, the solution will not be perfect.

We observed examples of events correctly reported as affirmative, despite the existence of a negation cue (in some cases several) in the same sentence. See Example 4.8.

Example 4.8. *“None of these changes were associated with any visible redistribution of actin, intermediate filaments or microtubules, and no nuclear involvement was detected.”*

(From PMID 6318692, events and negation extracted by BioContext)

Example 4.8 shows a sentence containing at least two negation cues: *"no"* and *"none"*. Some consider the word *"any"* as a highly indicative negation feature as well. However, although there are negated concepts expressed, the main event in question (i.e. the transcription of actin in filaments triggered by *"redistribution"*) is not reported negatively. This is an example which other approaches would have failed to detect correctly. Depending on the exact rules used, many bag-of-words, sentence-level, or surface distance approaches could have reported this event as negated.

4.5.2 Error analysis

We analysed the false positive and false negative results reported by Negmole on both negated and speculative events. The relatively small number of instances in the evaluation sets makes both interpretation of the results and error analysis difficult and limited. However, after analysing FP and FN results, we were still able to identify the following categories of errors.

One of the major sources of FNs was the issue of identification of contrasting patterns, usually causing an affirmative and a negative event to appear in close proximity in the same sentence. This causes problems with identifying the boundaries of the negation scope. Example 4.9 illustrates the common issues.

Example 4.9.

(a) Negated, FN: *"T cells lack active NF-kappa B but express Sp1 as expected."*

(b) Negated, FN: *"Unlike TNFR1, LMP1 can interact directly with receptor-interacting protein (RIP) and stably associates with RIP in EBV-transformed lymphoblastoid cell lines."*

(c) Negated, FN: *"Nmi interacts with all STATs except Stat2."*

In instance ***(b)*** of Example 4.9, a negated interaction is expressed, but there is no sign of a negation cue or negative sentence structure. Still, we can infer that TNFR1 cannot interact directly with RIP; it may also imply that TNFR1 does not stably associate with RIP in certain cell lines. The negation therefore can only be inferred by taking the following steps:

1. Recognising the presence of a contrasting pattern in the sentence;
2. Identifying the contrasting entities (in this example TNFR1 and LMP1);
3. Extracting the explicitly stated event (LMP1 interacts with RIP in this case);
4. Identifying the scope of contrast; this can be ambiguous, as in Example 4.9 it is not clear whether the two entities also contrast in "*stably associates with RIP*", or only in "*interact directly with RIP*".

Contrasting patterns are not uncommon. There are 125 phrases expressing contrast in the training data (in 800 abstracts) and 32 in the development data (150 abstracts) using only the patterns "unlike A, B", "B, unlike A", and "A; in contrast B". In these cases, the negation is usually not linguistically explicit, and has to be inferred by analysing the contrasts. Future work could explore a rule-based framework that would identify contrasting patterns and entities, and treat such expressions separately from explicit negations, for which a ML approach could still be useful.

At present, if a sentence contains more than one negation or speculation cue, we only extract the features concerning the "main cue", i.e. the one with the smallest parse-tree distance to the trigger. This causes the production of wrong results in some of the more complicated sentences containing double negation in particular.

Furthermore, a number of these double negation cases contain one linguistic negation and one biological negation. In Example 4.10, the word "suppress" indicates a biological negation, but paired with the negation cue

"nor", makes the overall sentence affirmative.

Example 4.10.

Negated, FN: *"In contrast, neither the RA-stimulated, RARE-mediated transcription nor the induced RAR-beta expression was suppressed by VitD3."*

A more correct annotation would have been to interpret "*suppress*" as negative regulation, which is negated by the cue "*nor*"; this negative regulation event was instead annotated as a negation by the annotators. Confusion between negated regulation and negative regulation—even by human annotators—has also resulted in a number of errors.

A number of errors were due to negation/speculation cues that were missing from the cue lists which were created semi-automatically from the training data. For example, *"potential"*, *"unknown"*, and *"possibility"* did not appear in the training data and were missing from the speculation cue lists.

Finally, a number of errors originate from potentially subjective, inconsistent, or simply incorrect annotation by the human annotators. See Example 4.11.

Example 4.11.

(a) Negated, FP: Negmole (correctly) found this event as a negated localization event, whereas the annotators have reported it as an affirmative localization event:

"[...] failure of p65 translocation [...]"

(b) Speculated, FP: The following event was (correctly) classified by Negmole as speculation, whereas the annotators did not consider it speculative:

"Proliferation, as measured by the percentage of cells in cycle appeared normal, as did rearrangement and expression of the TCR

beta-chain."

(c) Speculation, FN: The following event was (correctly) not caught by Negmole as speculation:

"Analysis of the regulation of the p40 gene promoter revealed that ASA inhibited NF-kappaB activation and binding to the p40-kappaB"

Some examples suggest that classification may not always be a correct construction of the negation and speculation detection problem. In Example 4.12, an affirmative and a negative event are stated, each observed in a different population. If the population context is not extracted, the two events would be recognised as a single event, and therefore assigning a single polarity value to them would be incorrect.

Example 4.12. *"Interleukin-2 production was diminished in the patient but not in the healthy twin."*
(From PMID 6239872)

In this research we addressed the problems of negation detection and speculation detection independently. It will be interesting to investigate whether one could help as a feature in the detection of the other, or whether the combination of the two could add richer context to the extracted information. With less than 0.1% of the events in the evaluation corpus being both negated and speculative, we did not have enough data at this stage to further investigate these questions.

The same observation can be made about the inferred events. The occurrence of the event components in a conjunctive structures can be used as an additional feature in the detection of negations and speculations.

While in minority, the number of negated and speculative events in the BioNLP'09 corpus is still significant (7% and 5%, respectively). Using these results, we extrapolate that by applying Negmole to a large-scale corpus,

around 1.56 million negated events and 1.24 million speculated events could be identified. This data could provide a very useful resource both for academics searching for previously reported results that are related to or conflict with their own, for data miners aiming to detect interesting patterns, and for bioinformaticians who wish to perform large-scale *in silico* experiments.

Although Negmole extensively uses semantic features related to the molecular events, it does not rely on any characteristic exclusive to biomedical events *per se*. We have demonstrated the extensibility of the method by applying it with minimal modifications to the similar task of speculation detection. Whilst not evaluated, the lexical, syntactic, and semantic features are generic enough to be applied to other types of relations extracted from domains other than the domain of the biomedical literature.

4.5.3 Further discussion

As expected, approaches that focus only on event triggers and their surface distances from negation cues proved inadequate for biomedical scientific articles. Low recall was mainly caused by many event triggers being too far from the negation cue on the sentence level to be detected as within the scope.

Furthermore, compared to clinical notes for instance, sentences that describe molecular events are significantly more complex. This is partly demonstrated by the occurrence of on average 2.6 event triggers in the event-describing sentences in the training data, and higher number of events per sentence, sometimes with opposite polarities.

Consider for example the sentence shown in Example 4.13.

> **Example 4.13.** *"We also demonstrate that the IKK complex, but not p90 (rsk), is responsible for the in vivo phosphorylation of I-kappa-B-alpha mediated by the co-activation of PKC and calcineurin."*
> (From PMID 10438457, BioNLP'09 corpus annotations expanded by adding protein complexes.)

Here, the trigger (*phosphorylation*) is linked with one affirmative and one negative regulatory events with two different entities (as well as participate as the theme of two regulatory events) hence triggering two events of opposite negations.

These findings, together with previous work, suggested that for any method to effectively detect negations, it should be able to link the negation cue to the specific token, event trigger or entity name in question. Therefore, more complex models are needed to capture the specific structure of the sentence as well as the composition of the interaction and the arrangement of its trigger and participants.

By combining several feature types (lexical, syntactic and semantic), the machine learning approach proved to provide significantly better results. In the incremental feature addition exploration process, adding the cue-commands-participant feature had the greatest effect on the F-score, suggesting the significance of treating event participants. We note, however, that many of the previous attempts focus on event triggers only, despite the fact that participants do play an important role in the detection of negations in biomedical events and thus should be used as negation targets instead of or in addition to triggers. It is interesting that adding the feature concerning the parse-tree distance between the trigger and negation cue improves precision by 5% (see Table 4.10).

Differences in event classes (in the number and type of participants) proved to be important. Significant improvement in performance was observed when individual classifiers were trained for the three event classes, suggesting that events with different numbers or types of participants are expressed differently in text, at least when negations are considered. Class I events are the simplest (one participant only), so it was expected that negated events in this class would be the easiest to detect (F-score of 77%). Class II negated events (which can have multiple participants), demonstrated the lowest recall (33%).

It is surprising that negated regulation events (Class III) were not the most difficult to identify, given their complexity.

We applied the negation detection on the type, trigger and participants of pre-identified events in order to explore the complexity of negations, unaffected by automatic named entity recognition, event trigger detection, participant identification, etc. As these steps are typically performed before further contextualisation of events, this assumption is not superficial and such information can be used as input to the negation detection module.

The best F-score for negation and speculation detection in BioNLP'09 were in the region of 23-25%, with a reported recall of up to 15%, but with overall event detection sensitivity of 33% (Kilicoglu et al. 2009) on the test dataset (different from that used in our evaluation). These systems did not use gold-standard event data, and this makes it difficult to directly compare their results to our work.

Using their precision and recall values for event extraction, it is however possible to provide some rough estimates of what their results would have been if applied on gold-standard event data. Had all events been correctly recognised, their negation detection approach could have reached 45% recall (compared to 49% in our case). With precision of around 50%, their projected F-score, again assuming perfect event identification, could have been in the region of 50% (compared to 63% in our case).

The approaches that focus on sentence-level modality annotation (e.g. (Shatkay et al. 2008)) have reported F-measures above 70%, but a meaningful comparison of the results between such systems and the current study is not possible. We also note that, for both negations and speculations, the major issue is improving recall by highlighting the variability of the negation/speculation expressions.

The experiments with rules that were based on the command relations have proven to be generic, providing high recall (76%) but with poor precision (23%). Although only the results with *S*-command relations have been reported

here (see Table 4.8), we examined other types of command relation, namely NP-, PP-, SBAR-, and VP-command. The only variation able to improve prediction accuracy was whether the cue VP-commands any of the participants, with an F-score of 42%, which is higher than the results achieved by the S-command (F-score of 35%).

In the machine learning approach, applying the method on all the classes shows that the best micro-averaged results for negated event detection (F-measure of 63%) have been achieved when separate classifiers were trained with the identical set of shared features. Sparsity of data is a likely reason for the drop in performance when additional class-specific data was used for training.

(MacKinlay et al. 2009) used gold annotations as input for negation detection, and reported an (estimated) precision, recall, and F-score of 68%, 24%, and 36% respectively on the same dataset (compared to 88%, 49% and 63% in our case) by using an ML with features comprising complex deep parse features.

As expected, negated events from class I (only one participant) were the easiest to detect (F-measure 67% to 77%). On the other hand, class III negated events, although the most complex between all event types, were easier to detect than class II negated events (possibly multiple participants). However, we note that the testing data had very few negated events of class II (only 15). When the same model was applied to speculation detection, there was a significant drop in the quality of results (F-measure of 48%). Still, it is interesting that precision between 64-73% was achieved on class III speculated events, which are the most complex and also most frequent in the training set. With semantic tokenisation, the precision of class I speculated event detection reached 70%. Class II events proved to be challenging, although any conclusions are limited by the small number of testing examples (only 8).

Using semantic tokenisation was beneficial in avoiding common errors while aligning multi-token and sub-token entities to the nodes of the parse tree

of the sentence in order to extract syntactic features. However, the results of the two pairs of experiments that only differ in the use of semantic tokenisation (Tables 4.13 and 4.14, and also Tables 4.18 and 4.19), do not provide enough evidence that it affected the results significantly. The effects of semantic tokenisation on the quality of information extraction will need to be further investigated.

4.6 Summary and conclusion

In this chapter we evaluated the methods proposed and developed for molecular event extraction (Evemole) and negation and speculation detection (Negmole).

At the time of developing Evemole, no other reliable event extraction system was available. We showed that Evemole can detect events with simple structures (events of classes I and II), but has room for improvement on regulation events. Although Evemole was later outperformed by other state-of-the-art purpose-built systems, its performance is comparable with these.

Negmole, on the other hand, detects the negated events with competitive performance, and was successfully expanded to detect speculative events as well. This suggests that it can serve as a key component in a larger text mining pipeline to detect conflicting statements.

Chapter 5

Large-scale consolidation of molecular event data

In this chapter we describe an approach to facilitate data consolidation in the domain of molecular events by finding conflicting claims of facts in the literature. For this purpose, it is useful to integrate all possible outputs from all types of tools. In particular, since different tools report many non-overlapping sets of entities, events, or other information, by integrating them we can aim for higher recall or precision and use them for data mining.

We propose a way to aggregate, analyse, and consolidate the extracted data to discover potential conflicts, contrasts, and contradictions. To achieve this, we merge events from different gene and protein named entity recognisers and normalisers. Moreover, we modify and merge the outputs of two state-of-the-art event recognition tools, namely TEES and EventMiner which became publicly available in the later stages of this research.

At this stage, we decided not to include the output from Evemole in the final integrated results, given that other reliable and high-performing tools were now available, and that event extraction was not the focus of this research, but merely a means of meeting our goal of mining conflicts. However, the modular nature of the integration pipeline means that it is possible to merge the results of any other tool, or replace any of the existing tools with new ones. We provide a wrapper for Evemole, which makes it possible to integrate its output with the other tools.

The rest of this chapter is structured as follows. Technical details on the implementation of the text mining framework will be presented in Section 5.1. Section 5.2 introduces the strategy of representing event mentions in text as well as representing biologically distinct events. As a way of valuating the events, we assign confidence values to the extracted events. The method to derive and assign these values are described in Section 5.3.

The method of mining the extracted data to find conflicts is described in Section 5.4. Finally, Section 5.5 explains how the data and code can be accessed for download or to browse through a web interface.

The text mining framework and the large-scale experiments were parts of a larger joint project with Martin Gerner (Faculty of Life Sciences, University of Manchester).

5.1 Framework for TM result integration and consolidation

In order to construct a unified system for consolidation of text mining results, it was necessary to integrate a number of different components that perform event extraction and contextualisation. For this purpose, we designed and implemented an integrated text mining system, called **BioContext**, that extracts, expands and integrates mentions of molecular events from the literature and facilitates data analysis and consolidation. The system relies on **TextPipe**, a framework for text mining result integration and consolidation (see also (Gerner 2011)).

5.1.1 TextPipe

In order to facilitate the integration of tools and merging of data, we constructed a lightweight framework called TextPipe. While other text mining frameworks like UIMA[19] and GATE (Cunningham et al. 2011) are available, we designed a system which was more light-weight and, more importantly, could be easily modified and optimized for any stability or performance problems we might (and did) encounter.

TextPipe makes extensive use of modularisation, parallel processing, database optimisation, error handling and recovery to address various practical challenges when applying a diverse set of tools to large sets of documents, and in our case, abstracts and full-text articles. It is written in Java, but allows tools written in any language to be integrated as components.

Any tool can become a component in a system deploying TextPipe in

19 http://uima.apache.org/

order to benefit from the functionalities that it provides. Tools are wrapped as TextPipe components (treated as black boxes internally) by implementing two simple methods: one to specify the output fields of the tool, and another to call the main method of the tool. Data is communicated in the form of lists of key-value pairs, similar to the model used in Google's MapReduce (Dean et al. 2008).

TextPipe components are either applied directly to documents or run as services. They do not need to provide a list of dependencies. Instead, during run-time they connect directly to other components, providing the document (or documents, if run in batch mode) that need to be processed, and fetching the output of those components to use as their required input. Computed results can be stored in databases for later re-use to avoid multiple processing of the same documents.

To summarise, TextPipe offers the following features as a tool integration framework.

- Capability to input and process a diverse range of textual formats;
- Ability to be deployed as a web service;
- Incorporation of any tool with a simple wrapper, and providing an interface between tools;
- Provision of methods for large-scale processing of documents;
- Support for concurrency;
- Use of databases for storage and retrieval of computed data;
- Use of caching.

5.1.2 BioContext overview and components

Figure 5.1 shows an overview of the integrated system. Processing is performed in four stages: named entity recognition and normalisation, grammatical parsing, event extraction, and context extraction. Each stage is composed of several components. In some cases, outputs from multiple components are merged prior to use by other components. These processing

stages and their components are described below.

The numbered circles in Figure 5.1 show the places where data integration is performed. The output of gene NER modules are merged in the circle numbered 1. The combined identified entities with the additional entities related to the anatomical locations and species names are replaced by place-holder nouns in circle numbered 2 and semantic tokenisation is performed as a preprocessing step for parsers.

The event extraction tools use the output of these parsers as their input, and produce independent sets of events, which is later merged in circle numbered 3. The merged data is further processed by adding context, i.e. information regarding negation, speculation, and anatomical location associated with every event.

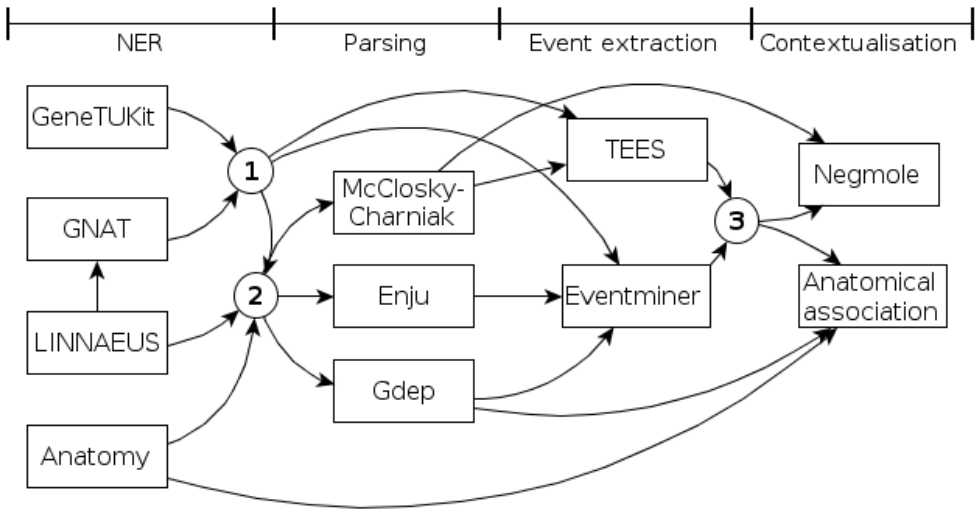


Figure 5.1: An overview of BioContext

The diagram shows how the different components of the system are connected. Circles represent merging and post-processing of data. Different stages are shown on the bar above. The numbered circles show the places at which data integration or merging happens.

The following sections will describe the components we used in the construction of BioContext in more detail.

5.1.3 NER

In the first stage we perform recognition and normalisation of any entities that are needed in later stages of the system.

Named entity recognition for genes and proteins are performed by GeneTUKit (Huang et al. 2011) and GNAT (Hakenberg et al. 2011); (Solt et al. 2010). To the best of our knowledge, these tools are the only tools available that are capable of normalisation and are applicable to large-scale datasets from a practical point of view. GeneTUKit normalises genes and proteins from any species, while GNAT can only normalise genes and proteins from 30 of the most frequently discussed organisms). GNAT uses species NER as input, which was performed by LINNAEUS (Gerner et al. 2010b). Both tools were configured to use BANNER (Leaman et al. 2008) for recognition to improve coverage.

We used a modified version of GNAT that reports not only the mentions that could be normalized to database identifiers, but also any mentions that were recognized by BANNER but could not be normalized. GNAT relies on species NER, which was performed by LINNAEUS (Gerner et al. 2010b). Data extracted using these non-normalized entities will have limited context, and will not be as reliably assigned to their correct distinct group. However, leaving them out would have caused errors in the other components, specifically the event extractors, as they rely on entities to be marked in the sentence. In the absence of recognised entities, they will miss the event altogether, or report an unrelated entity appearing elsewhere in the sentence as the participant. The first case lowers the recall, and the second case affects the precision.

The output from both gene/protein NER systems are merged after production. If the two tools have identified overlapping spans, then we create a new span with the union of their coordinates. If the tools have assigned different Entrez Gene identifiers in the original overlapped spans, then priority is given to the GeneTUKit normalisation.

NER of anatomical locations (e.g. *“brain”*, *“T cells”*) and cell-lines

(acting as proxies for anatomical locations, e.g. *"HeLa"* for cervical cells) were performed by the anatomical NER system from GETM (Gerner et al. 2010a) which relies on a comprehensive dictionary of anatomical locations collected from 13 OBO ontologies.

5.1.4 Grammatical parsing

In the second stage, a number of deep and shallow grammatical parsers process the texts. In order to increase the accuracy of the parsers when applied to sentences with long and complex entity names, we performed semantic tokenisation (see Section 3.2). Using this, we ensured that multi-word entities were not tokenized into multiple tokens. This was performed by replacing any entities recognized in the first stage with place-holders (generic terms tagged as nouns) prior to parsing. After parsing, the place-holders were replaced with the original strings again.

We used the McClosky-Charniak constituency parser (McClosky et al. 2006), the Gdep dependency parser (Sagae et al. 2007b), and the Enju parser (Sagae et al. 2007a) to parse every sentence in the corpus.

In addition to the McClosky-Charniak parser, we also experimented with the constituency parse trees automatically produced by the parser reported in (Bikel 2004). No significant differences were observed in the results of Negmole, one of the components requiring constituency parse trees. Therefore all the results and processes are reported using the McCloslky-Charniak parser.

5.1.5 Event extraction and integration

For the extraction of events, we chose to use two systems: the Turku event extraction system (TEES) (Björne et al. 2009), and EventMiner from the University of Tokyo (provided by Makoto Miwa, currently unpublished). TEES was the highest-scoring system in the BioNLP'09 event extraction challenge (see Section 2.3.4), and evaluation results for EventMiner presented as a keynote talk at BioCreative III showed it as having higher accuracy (unpublished). To the best of our knowledge, these are the only systems that

are both accessible and maintained. Both systems use gene/protein NER results from the first stage. In addition, TEES also uses output from the McClosky-Charniak parser and EventMiner uses results from Enju and Gdep.

In order to take advantage of these tools we have designed a method to merge several event extraction outputs. The integrated results can be useful when deciding the balance between precision and recall, depending on how the data will be used (see Section 5.2).

The output from the two systems, which is merged after production, consists of information about the event type, the event trigger and the event participants. The results are reported in the BioNLP'09 format, with each event referencing the entities or the other events by their IDs. So, each extracted event is stored in the database with its components spread over several rows, each referencing the others.

The events extracted from the two tools are compared to determine whether they refer to the same mention of an event. Two events extracted from a given sentence match if their type and participants match (we used approximate boundary matching conditions, allowing overlap for the participant mentions). If the event involves other events, the matching criteria is examined recursively. Note that here we do not require the triggers to match, as they do not convey any biological information.

After studying a sample of the merged output from the large-scale MEDLINE event extraction, we noticed recurring patterns that contributed towards many incorrectly extracted events.

To increase the precision, we designed post-processing methods that negatively discriminated (i.e. removed probable false positives) against those events that followed these patterns. The rules were based on the event trigger and the event structure, as explained below. We also consider improving coverage by inferring additional events (see Section 5.1.7).

Negative discrimination based on the event trigger

Events whose triggers indicate that the events are wrong are removed. Very short triggers (one or two characters, mostly consisting of punctuation, single letters or abbreviations) were removed. We also compiled a white list of 11 short words (*"/"*, *"-"*, *"is"*, *"by"*, *"as"*, *"on"*, *"up"*, *"at"*, *"be"*, *"do"*, and *"if"*) that could be triggers, and a blacklist of 15 longer words which were common English stop words (*"the"*, *"and"*, *"in"*, *"of"*, *"cells"*, *"to"*, *"when"*, *"patients"*, *"are"*, *"mice"*, *"from"*, *"both"*, *"that"*, *"mouse"*, and *"what"*) and were often recognised incorrectly as event triggers. Events that had a trigger from the white list were not removed, and events that had a trigger from the blacklist were removed.

In addition, events with capitalised triggers which were not situated at the beginning of a sentence were removed, as many of these capitalised words turned out to be proper nouns, and seldom functioned as an interaction trigger. For example, an event with the trigger *"Expression"* from the sentence *"Expression of the argA gene carried by a defective lambda bacteriophage of Escherichia coli."* (extracted from PMID 130376) would be retained since it was in the beginning of the sentence, but an event with the (incorrect) trigger *"E.P."* from the sentence *"[...] primary visual E.P. in Medial Laternal gyrus [...]"* (PMID 142561) would be removed, since *"E.P."* was not in the beginning of the sentence.

Negative discrimination based on the event structure

The event extractor components identify nested regulatory events in which either or both of the participants may be other events. However, they are likely to report events that are circularly nested (e.g. E1 causes E2, which itself causes E1) or in a very long chain (e.g. E1 causing E2 causing E3 and so on.) In one instance TEES found a chain of 211,769 connected events. We noticed that there are very few instances in the training data where events are nested deeper than two levels, and there are no circularly nested events. In addition to

making little biological sense, none of the cases of circular or long chain events that were manually examined were correct. Therefore, we categorically removed all the events that were nested any deeper than two levels.

5.1.6 Adding context

In the final stage, further context is extracted and is associated with the events that were extracted in the previous stage. This information includes associated anatomical locations and whether extracted processes have been reported as speculative or negated.

Negation and speculation association

We use Negmole to determine whether the events are described negatively or speculatively. The input of Negmole, in addition to the extracted events, contains constituency parse trees from the McClosky-Charniak parser.

The methodology and design principles of Negmole have been described in detail in Sections 3.4 and 3.5. It takes as input text, named entities marked with offsets, parse trees of the sentences and the extracted events (trigger, type, and participants.) Negmole classifies each event as negated/affirmative and speculated/asserted. The flowchart in Figure 3.15 on page 138 summarises the operation and requirements of this system.

The input formats are those of the BioNLP'09 Shared Task. The Java implementation comes with a wrapper to function as a module in the TextPipe framework and uses the same unified input and output format as any other TextPipe module: a map of string to string.

We used the features extracted from the BioNLP'09 training set to train an SVM model which was used in the SVM classification algorithm to classify each of the test instances. Negmole extracts features from the inputs and creates feature files to be used as the input to the SVM machine learning system. We used SVMperf as the choice of the SVM engine.

Species and anatomical association

Anatomical locations are associated with events using an expanded version of the method described in (Gerner et al. 2010a). It relies on Gdep dependency trees to link events and associated anatomical entities. The tool was integrated as a TextPipe component with a wrapper interfacing the framework.

We used LINNAEUS (Gerner et al. 2010b) to extract mentions of species names (e.g. *"human"*, *"dog"*, *"mus musculus"*, etc.) and mentions of anatomical locations (e.g. *"blood"*, *"vein"*, *"epithelium"*, or cellular locations such as *"lymphoid tissue"* or *"nucleus"*).

The anatomical entity IDs come from 13 different ontologies from The Open Biological and Biomedical Ontologies (OBO) foundry[20], some of which are species-specific, and others refer to higher taxonomic orders such as genus, class, etc. The mentions were normalised based on their string equality as well as the LINNAEUS's native dictionary matching method. Moreover, the anatomical entities were associated with a certain species whenever possible. The anatomical locations were assigned unique internal identifiers that reflected the anatomical location as well as the species it exists in.

5.1.7 Inferring additional events from enumerated entity mentions

Mining conflicting events in the literature is a task that requires large-scale extracted data. In order to increase the number of events that have been extracted, we infer further events from enumerated entity mentions.

We noted that a relatively large number of gene/protein and anatomical entities in MEDLINE are part of "enumerations", i.e. lists of more than one entity connected within a conjunctive phrase (for example, three anatomical entities are enumerated in the phrase *"amniserosa, dorsal ectoderm and dorsal mesoderm"*). We hypothesize that wherever event extractors or the anatomical association method associate an event with a gene/protein or anatomical entity which is part of such an enumerated group, we can infer additional events,

20 http://www.obofoundry.org/

where the original entity is replaced with each of the other entities in the enumeration.

For example, in Example 5.1 gene expression events should be extracted for all three *Dorsocross* genes, and each of those events should be associated with each of the anatomical locations mentioned. If any of these nine events are not extracted, the event inference based on enumeration will be able to infer the event(s) that were missed by the event extractors.

> **Example 5.1.** *"In the present study, we describe three novel genes, Dorsocross1, Dorsocross2 and Dorsocross3, which are expressed downstream of Dpp in the presumptive and definitive amnioserosa, dorsal ectoderm and dorsal mesoderm."*
> (From PMID 12783790)

Since each event is initially only associated with a single anatomical location, this method makes it possible to infer additional events for the anatomical locations that were not associated to the original events.

In order to implement this method, we used regular expression patterns (see Table 5.1) to detect groups of enumerated entities. The regular expressions are applied recursively, causing any number of entities to match. For example, in the phrase "T1, T2, and T3" the first two entities (T1 and T2) will be in the same group by applying the first regular expression, and T2 and T3 will be in the same group by applying the second rule. Finally, the groups are merged to form a perfect partitioning of the entities, and therefore, T1, T2, and T3 would belong to the same entity group. If T1 and T2 belong to the same entity group and an event E1 is extracted with T1 as the participant, we construct a new event E2 with the entity T2. Except for T1, all other properties of E1 will be duplicated in E2.

Patterns to match	Regular expression to match the sub-string between T1 and T2
T1, T2 T1/T2	`"^[,/] ?$"`
T1, and T2 T1 and T2	`"^,? and $"`

Table 5.1: Regular expressions used to enumerate named entities

The regular expressions are applied recursively, causing any number of entities to match.

5.2 Event representation

5.2.1 Event mention representation

We are mainly concerned with the event-level information about biomedical processes. Therefore we would like to extract, represent, modify, and analyse the data on the level of events. For that purpose, we designed a model that is represented as a denormalised table where every record is one instance of an event *mention* in a document. We populated the table with the extensive mention-level information about each extracted event. Cross references to the other events in nested events were expanded and added as attributes to the parent event record. The columns of the denormalised table and a brief description of each column are listed in Table 5.2.

Attribute Type	Name	Values	Description
G	document ID	PMID or PMC unique identifier	The MEDLINE or PMC document in which the event was mentioned
G	sentence	string	The sentence that mentions the event
G	sentence offset	integer	The character offset of the sentence in the document

E	confidence	real	The confidence of event extraction
E	type	enum (nine possible values)	The biological type of the event; one of the nine types used by the BioNLP'09 corpus
E	level	integer	Indicates whether the event is simple (i.e. has entity participants) or nested (has other events as participants)
E	trigger term	string	The textual trigger of the event
E	trigger start, trigger end	integer	The character offsets of the start and end of the textual trigger
E	TEES	boolean	Whether the event was extracted by TEES
E	Tokyo	boolean	Whether the event was extracted by EventMiner
E	inferred gene	boolean	Whether the event was inferred based on enumerated gene/proteins
E	inferred anatomy	boolean	Whether the event was inferred based on enumerated anatomical entity
E	negated	boolean	Whether the event is negated
E	negation cue	string	The textual cue for negation
E	negation cue start, end	integer	The character offsets of the start and end of the negation cue
E	speculated	boolean	Whether the event is speculated
E	speculation cue	string	The textual cue for speculation
E	speculation cue start, end	integer	The character offsets of the start and end of the speculation cue
E	anatomical location	string	The anatomical entity which has been assigned as the location of the molecular event
E	anatomical entity ID	Internal ID	Internal ID linked to OBO Foundry ontology identifiers
E	anatomical entity start, end	integer	The character offsets of the start and end of the anatomical entity mention
R	participants	multiple columns	The participants of the event. This spans over several columns as each participant may be an entity or another event. There are separate columns for 'theme' participants and 'cause' participants.
R	participant type	boolean	Whether the participant is an event or a gene/protein entity
R	entity term	string	The actual text that refers to the entity that participates in the event
R	entity ID	Entity ID in	The normalised ID of the gene/protein entity in

		NCBI	the NCBI Entrez Gene reference database. This could be null if the entity cannot be normalised.
R	entity start, end	integer	The character offsets of the start and end of the protein/gene mention
R	GNAT + BANNER	boolean	Whether the entity was found by GNAT + BANNER
R	GeneTUKit	boolean	Whether the entity was found by GeneTUKit
R	shallow match	boolean	Set if both GeneTUKit and GNAT overlap, but map the entity to different IDs.[21]
R	gene confidence	real	The confidence of gene extraction

Table 5.2: The attributes of mention level event representation

Each row of this table shows a column of the denormalised table identifying each event mention and its attributes. The attributes of type G are general attributes, type E are event level attributes, and type R are recursive attributes and are repeated as needed.

The left-most column in Table 5.2 shows the type of each of the columns of the denormalised database table. The G columns are the general attributes and every record has exactly one of each one of them. The E columns are event-specific attributes. Since every record represents an event, each row will have all the E records that refer to that event. However, some of the participants of the event can be other events. Therefore, R (recursive) attributes can be repeated or recursively repeated as required. For example, if a binding event involves two genes, we will have two sets of all the R attributes, one for every gene.

If a regulation event has a theme that is itself an event, we will repeat all the E attributes in that record, once for the event itself, and a second time for its theme that is also an event. The same principle applies to other types of nested events or events with multiple participants.

An example of an event represented in this manner can be seen in Table

21 GeneTUKit is prioritised over GNAT, since it was the best performing tool in the BioCreative III challenge.

5.5.

5.2.2 Distinct event representation

The denormalised table contains every mention of an event anywhere in the corpus, and may contain many similar events that are reported or discussed in the literature. In order to analyse the *distinct* events contained in the literature, and specifically as a way of normalising event mentions into equivalence classes, we use the concept of distinct events introduced in Section 3.1.3.

We chose the columns that are essential in defining an event and use them to construct a hash function that assigns an integer (hash) to a combination of attributes, ignoring extrinsic attributes such as all G attributes, index and offset attributes, terms (e.g. trigger term, entity terms, and cues). Instead, we include attributes that are intrinsic to the event such as event type, participating entity IDs, and negation information.

Specifically, the attributes that are used to identify distinct events are displayed in Table 5.3. Note that these are a subset of the attributes in Table 5.2.

Attribute Type	**Name**
E	type
E	negated
E	speculated
E	Anatomical entity ID
R	Participants
R	Participant type
R	Entity ID

Table 5.3: The attributes of every distinct event in the collapsed table.

The R columns are recursive, and are repeated for every participant, and therefore are in fact multiple columns.

The hash is calculated from the string containing these attributes (in the

most possible normalised form) and also added to every record in the denormalised table. The importance of the hashes lies in their definition of the "identity" of an event. The information incorporated in the hash is the essential information about an event, and corresponds with the distinct representation of an event introduced before.

We use the hash to collapse the denormalised table by grouping together the events that are likely referring to the same biological process, regardless of the document in which they have been mentioned, and of the words with which they have been described. The columns of the collapsed table are shown in Table 5.3. Similar to the denormalised table, the R columns are recursively repeated, for every participant. They include a copy of the E columns for each of the participants that is an event.

5.3 Ranking the events by text mining confidence

Not all the extracted events have the same quality. In a pipeline comprised of many modules, the quality of the final output is affected by the precision of every stage that affects the input. For example, the precision of the gene and protein NER stage which is one of the earlier stages of the pipeline is propagated through all the other stages that use the NER output as one of their inputs.

We use a method of confidence assignment to calculate the confidence level of every event as it is extracted by the system and stored in the database. We identify the precision level of every stage that the extraction of an event involves and use it as a factor to determine the confidence of the event.

For example, if two gene NER tools agree on an entity, the confidence of the extraction of that entity will be higher than when only one of them has detected the entity. The confidence will be proportionate to the precision of each tool.

NER		BANNER + GNAT	0.8
		GeneTUKit	0.72
		Intersection	**0.82**
Event extraction	EventMiner	Binding	0.27
		Gene expression	0.47
		Localization	0.36
		Negative regulation	0.32
		Phosphorylation	0.61
		Positive regulation	0.35
		Protein catabolism	**0.8**
		Regulation	0.22
		Transcription	0.48
	TEES	Binding	0.34
		Gene expression	0.58
		Localization	0.67
		Negative regulation	0.41
		Phosphorylation	0.6
		Positive regulation	0.44
		Protein catabolism	**0.62**
		Regulation	0.25
		Transcription	0.51
	Intersection	Binding	0.49
		Gene expression	0.7
		Localization	0.76
		Negative regulation	0.6
		Phosphorylation	0.7
		Positive regulation	0.61
		Protein catabolism	**0.92**
		Regulation	0.45
		Transcription	0.73
Inference		Gene enumeration	0.44
		Anatomical entity enumeration	0.34

Table 5.4: Coefficients that determine the confidence

The numbers in bold are the maximum in each category and are used to normalise the coefficients in every stage.

In every stage, we normalise the confidence coefficient by dividing all of the coefficients belonging to that stage by the highest coefficient for that stage. Table 5.4 shows the coefficients of every stage, and how they are normalised. The numbers are derived from the precision evaluation results that will be presented in Chapter 6 .

Table 5.5 shows an example from the extracted events. Note that the confidence is not very high (0.00136), which is expected for an event of level 1. It is interesting to observe that, apart from the event-participant association which is incorrect, the other attributes of the main event and the nested event, including the negation of the main event, the negation of the nested event, and the anatomical association by inference to the main event have been identified correctly.

Attribute Type	**Name**	**Values**
G	document ID	PMC2727658
G	sentence	For example in the liver and skin, there was no activation of toll-like receptors (which could play a role in pathogen recognition), no change in known antimicrobial peptide genes (although not all AMPs were represented on our chip because some sequences were shorter than our 60-mer probes), and no change in MHC class I or II genes, or genes involved in antigen presentation (e.g., LMP7, TAP1 and 2, cathepsines).
G	sentence offset	27983
E	confidence	0.001360
E	type	Regulation
E	level	1
E	trigger term	“change”
E	trigger start, trigger end	28283, 28289
E	TEES	FALSE
E	Tokyo	TRUE
E	inferred gene	FALSE
E	inferred anatomy	TRUE
E	negated	TRUE

E	negation cue	“no”
E	negation cue start, end	28280, 28282
E	speculated	-
E	speculation cue	-
E	speculation cue start, end	-
E	anatomical location	“liver”
E	anatomical entity ID	anat:184
E	anatomical entity start, end	28002, 28007
R0	participant type	Event
R0	participation type	Theme
R0	type	Positive regulation
R0	level	0
R0	trigger term	“activation”
R0	trigger start, trigger end	28031, 28041
R0	TEES	FALSE
R0	Tokyo	TRUE
R0	inferred gene	FALSE
R0	inferred anatomy	FALSE
R0	negated	TRUE
R0	negation cue	“no”
R0	negation cue start, end	28028, 28030
R0	speculated	FALSE
R0	speculation cue	FALSE
R0	speculation cue start, end	-
R0	anatomical location	“skin”
R0	anatomical entity ID	anat:115
R0	anatomical entity start, end	28012, 28016
R1	participant type	Entity
R1	entity term	“toll-like”

R1	entity ID	37272
R1	entity start, end	28045
R1	GNAT + BANNER	TRUE
R1	GeneTUKit	TRUE
R1	shallow match	FALSE
R1	confidence	0.014760

Table 5.5: Example event representation

The main event represented here is a Regulation event whose only participant is a positive regulation even with a theme role. It is extracted automatically using the event extraction pipeline, BioContext.

5.4 Finding conflicting statements

We focus on strict contrasts, as allowing some of the fields to be empty results in events that have less context extracted and therefore are likely to have less implicit context in common. We select a subset of the events that satisfy the following criteria:

1. The events have an associated anatomical location;
2. If they are binding events, two themes are present;
3. If they are regulatory events, they have causes;
4. The entity participants are normalised to standard entries;
5. The events are not speculative.

For every unique event satisfying the above criteria, we calculate the hash for a hypothetical event that matches it in every aspect, but has the opposite negation attribute. We then search the database for any event with this given hash.

This method allows us to find pairs of events that are common in all aspects (type, participants, and anatomical locations), and their only difference is the fact that one is affirmative and the other is negated.

We assign a **score** to every pair to indicate how prominent that pair is. To compute this score, we start by calculating the **cumulative confidence** of each one of the two hashes in the pair. Cumulative confidence is equal to the

number of different documents in which the distinct event corresponding to that hash (which we refer to as the **supporting event** for the hash) appears, regardless of the number of times it has appeared in a single document. The aim is to remove the bias caused by the repeated appearance of the same event in a document. In other words, the **cumulative confidence** of a hash is defined as:

$$cum(h) = \sum_{d_i \in Documents} max_j(c_{ij}(h))$$

where $c_{ij}(h)$ is the confidence of the *j*th occurrence of an event with hash *h* in document *i*, and the max function runs on different values of *j*. Assume that a number of mentions of a given event have been extracted from document *i*. Not all of these events are of equal confidence, and it is possible that some of them are false positive instances. However, regardless of those lower quality mentions, we consider the maximum score, i.e. $max_j(c_{ij}(h))$ in the above equation. Considering this event to denote the appearance of this particular event in document *i* prevents the lower quality of the other (possibly more complex) mentions from adversely affecting the notion of document-level confidence. By adding these document-level confidences, we take into account how commonly a distinct event is reported. We only use this cumulative confidence score for ranking purposes, and therefore we do not consider applying a logarithmic function on the sum.

Using this measure for how commonly and confidently a distinct event is reported, we define the score of a pair of events as the minimum of the cumulative confidences of the two corresponding hashes.

$$score(h_1, h_2) = min(cum(h_1), cum(h_2))$$

This is a measure that favours pairs that have a combined high frequency and confidence. Here we chose to use the minimum as the way to combine the

scores of the two events in a pair. Other ways of combining two scores such as various averages would have caused the more prominent event to dominate the overall score for the pair. We expect the pairs that are not in fact conflicting to have a very low score on one of the events of the pairs. For example, if it is a well-known fact that p53 is expressed in lung, we expect the cases of p53 not expressing in lung to be rare or with low confidence (likely to be false positives) and therefore to have a low score. On the other hand, the event "expression of p53 in lung" would have a very high score, and could dominate the score of the pair, despite the fact that the pair is not very likely to represent a conflict.

To address this issue, we would like the score of the pair to depend solely on the score of the event with the lower score. This will guarantee that a pair will be scored highly if both components have at least a minimum confidence score. The score will be used in ranking and evaluation of conflicting pairs, and in estimating the confidence of a pair. Using this ranking, we can present to a user all conflicting statements that have a minimum confidence or satisfy a certain filter.

5.5 Exploring the data

We applied BioContext, our integrated system, to extract events and their context, to MEDLINE (2011 baseline files, containing 10.9 million abstracts) and to the open-access subset of PMC (downloaded May 2011, containing 235,000 full-text articles). In this section we describe access to the data as well as the source code of BioContext and its components.

5.5.1 Browsing the data

We provide a web interface for browsing the data, and for performing web searches[22]. It is implemented in Python, and runs server calls to the database containing the results.

The interface is simple, allowing the user to enter any of the two

22 The web interface can be accessed at http://www.boicontext.org/

participants, and the anatomical location they are interested in, and select any number of the nine event types. The input strings are processed by entity recognisers and normalised. This will allow all the events involving the same entity to be fetched, regardless of the term used to refer to the entity.

When the query is submitted, the events matching the criteria will be returned in tabular form, with full context and citation information concisely displayed. The results can also be downloaded in structured text files from this page.

The following screen shots demonstrate the design and features of the web interface. On the first page (Figure 5.2) the user can enter the theme, cause, and the anatomical location they are interested in, or leave any of them blank. If any of the fields are left empty, events with all possible values for those attributes will be returned. We use the word “Target” for theme and “Regulator” for cause, as these terms are more comprehensible and appear more natural to biologists. The event types that are of interest can be selected with check boxes. By default all the check boxes are selected.

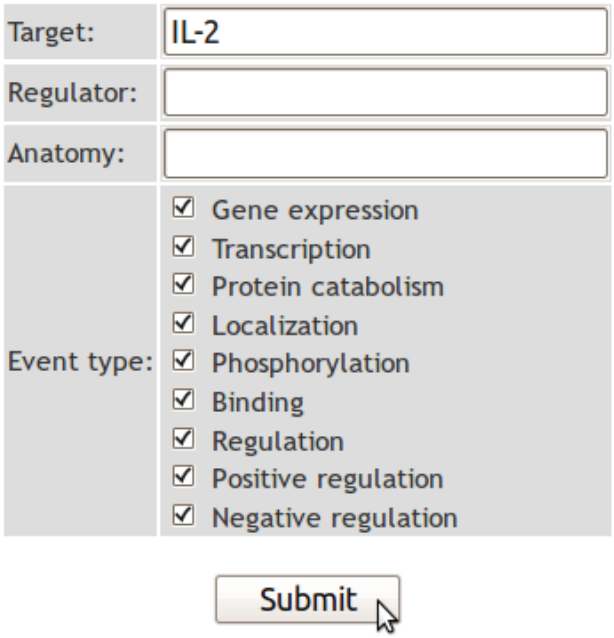


Figure 5.2: BioContext web interface: the first page

The query returns events of any type with IL-2 as the target (theme).

The queries are parsed by entity recognisers and are normalised to the relevant identifiers, possibly belonging to a certain species. For example, if one enters “rat IL-2” into the target box, the results related to IL-2 in rats will be returned.

After filling in the Target field with the query “IL-2”, we see a summary of the affirmative and negated cases of each type of event involving this gene. Clicking on any of these numbers will show the details of the extracted events (Figure 5.3).

Click here to choose a different target, regulator, or anatomical location for your search queries.

Process	Positive cases	Negative cases
Gene expression	29942	1022
Transcription	1928	95
Protein catabolism	142	1
Localization	5962	133
Phosphorylation	353	13
Binding	6594	204
Regulation	2846	347
Positive regulation	8982	436
Negative regulation	3648	74

Figure 5.3: BioContext web interface: summary of the query results

The list of events involving the human version of the gene/protein after submitting the query with IL-2 in the "Target" field. The count for affirmative and negated mentions are displayed separately. Clicking on the link indicated with 'here' will show the list of homologs of the entity in other species.

By default, if the species is not specified, human results are displayed. However, we could also choose other homologs (same entities in other species) by clicking on the appropriate link. A list of resolved entities for the IL-2 gene in other species are displayed, and will display the specific events occurring in those species (Figure 5.4).

Figure 5.4: BioContext web interface: list of homologs

If the entity from the search query (IL-2 in our case) exists in species other than human, a list of homologs can be accessed. The links will direct the user to the list of events involving the specified entity.

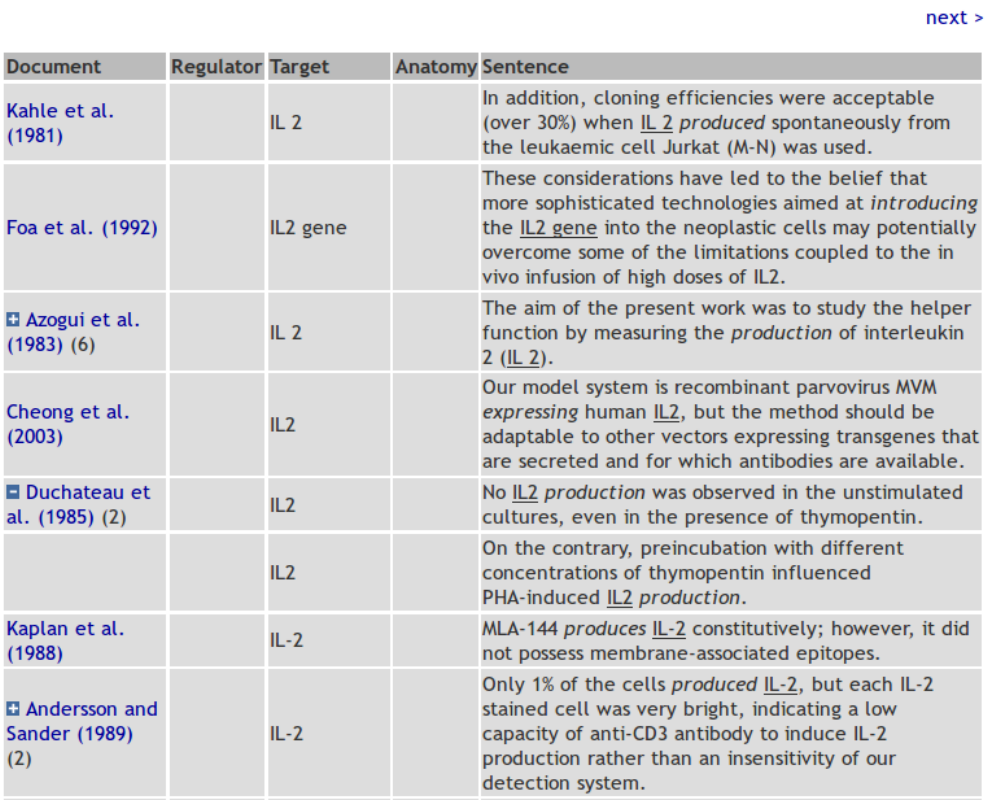
On the same page, we also see a list of all the distinct events involving the queried gene. Here, a distinct event can be selected for the viewing of the individual mentions (Figure 5.5).

Process	Target(s)	Regulator	Anatomy	Positive cases	Negative cases
Gene expression	IL2 (H. sapiens)			12179	419
Gene expression	IL2 (H. sapiens)		T cells	7364	249
Positive regulation	IL2 (H. sapiens)			3891	198
Gene expression	IL2 (H. sapiens)		lymphocytes	2480	57
Localization	IL2 (H. sapiens)			2412	67
Binding				1930	57
Negative regulation	IL2 (H. sapiens)			1862	43
Localization	IL2 (H. sapiens)		T cells	1592	30
Positive regulation	IL2 (H. sapiens)		T cells	1445	61
Gene expression	IL2 (H. sapiens)		mononuclear cells	1287	29
Regulation	IL2 (H. sapiens)			1128	161
Transcription	IL2 (H. sapiens)			919	57
Gene expression	IL2 (H. sapiens)		blood	816	17

Figure 5.5: BioContext web interface: list of the distinct events

List of distinct events, regardless of the document or sentence they have appeared in, which involve the queried entity in the specified role.

The affirmative or negated cases of a distinct event can be selected for display. The empty columns indicate incomplete context. The citations are extracted and formatted, and are linked to the original document. Figure 5.6 shows the list of affirmative mentions of this particular event. The sentences from the same document are grouped together, and can be viewed individually by selecting the plus sign next to the document citation information (compare, for example, the third row with with rows 5 and 6 in Figure 5.6.) The open access PMC articles are indicated by a lock sign next to the reference.



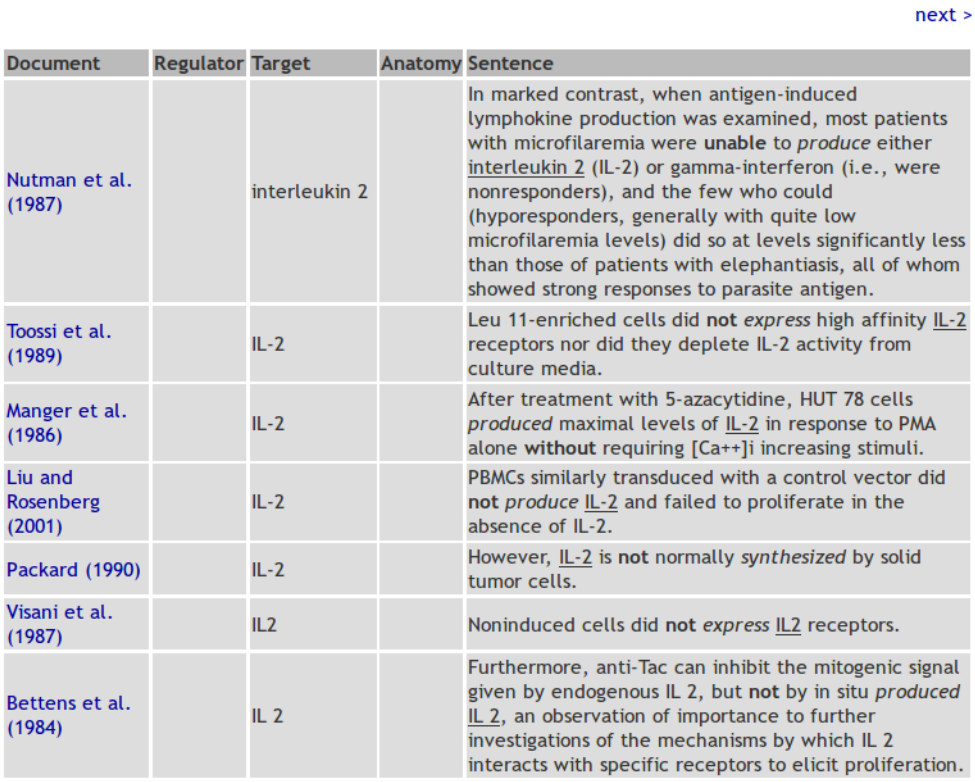
next >

Document	Regulator	Target	Anatomy	Sentence
Kahle et al. (1981)		IL 2		In addition, cloning efficiencies were acceptable (over 30%) when IL 2 *produced* spontaneously from the leukaemic cell Jurkat (M-N) was used.
Foa et al. (1992)		IL2 gene		These considerations have led to the belief that more sophisticated technologies aimed at *introducing* the IL2 gene into the neoplastic cells may potentially overcome some of the limitations coupled to the in vivo infusion of high doses of IL2.
⊞ Azogui et al. (1983) (6)		IL 2		The aim of the present work was to study the helper function by measuring the *production* of interleukin 2 (IL 2).
Cheong et al. (2003)		IL2		Our model system is recombinant parvovirus MVM *expressing* human IL2, but the method should be adaptable to other vectors expressing transgenes that are secreted and for which antibodies are available.
⊟ Duchateau et al. (1985) (2)		IL2		No IL2 *production* was observed in the unstimulated cultures, even in the presence of thymopentin.
		IL2		On the contrary, preincubation with different concentrations of thymopentin influenced PHA-induced IL2 *production*.
Kaplan et al. (1988)		IL-2		MLA-144 *produces* IL-2 constitutively; however, it did not possess membrane-associated epitopes.
⊞ Andersson and Sander (1989) (2)		IL-2		Only 1% of the cells *produced* IL-2, but each IL-2 stained cell was very bright, indicating a low capacity of anti-CD3 antibody to induce IL-2 production rather than an insensitivity of our detection system.

Figure 5.6: BioContext web interface: list of affirmative cases of the given event

Individual affirmative mentions of the IL-2 expression, with highlighting on the original sentence and link to the source document.

Similarly, negative cases can be viewed in a list with references to the original documents in MEDLINE or PMC (Figure 5.7). The negation cues are highlighted, as well as the speculation cues wherever they cue an affected event. The cues that do not affect events are left without highlighting.



next >

Document	Regulator	Target	Anatomy	Sentence
Nutman et al. (1987)		interleukin 2		In marked contrast, when antigen-induced lymphokine production was examined, most patients with microfilaremia were **unable** to *produce* either interleukin 2 (IL-2) or gamma-interferon (i.e., were nonresponders), and the few who could (hyporesponders, generally with quite low microfilaremia levels) did so at levels significantly less than those of patients with elephantiasis, all of whom showed strong responses to parasite antigen.
Toossi et al. (1989)		IL-2		Leu 11-enriched cells did **not** *express* high affinity IL-2 receptors nor did they deplete IL-2 activity from culture media.
Manger et al. (1986)		IL-2		After treatment with 5-azacytidine, HUT 78 cells *produced* maximal levels of IL-2 in response to PMA alone **without** requiring [Ca++]i increasing stimuli.
Liu and Rosenberg (2001)		IL-2		PBMCs similarly transduced with a control vector did **not** *produce* IL-2 and failed to proliferate in the absence of IL-2.
Packard (1990)		IL-2		However, IL-2 is **not** normally *synthesized* by solid tumor cells.
Visani et al. (1987)		IL2		Noninduced cells did **not** *express* IL2 receptors.
Bettens et al. (1984)		IL 2		Furthermore, anti-Tac can inhibit the mitogenic signal given by endogenous IL 2, but **not** by in situ *produced* IL 2, an observation of importance to further investigations of the mechanisms by which IL 2 interacts with specific receptors to elicit proliferation.

Figure 5.7: BioContext web interface: list of negated cases of the given event

Individual negated mentions of the IL-2 expression, with highlighting on the original sentence and link to the source document similar to the affirmative cases. The negation cue responsible for negating the even is highlighted in each case.

5.5.2 Availability of data and the code

In addition to the data that can be accessed for browsing and downloading through the web interface, we also provide the data produced as the output of this system, as well as all the intermediary data freely available. It is accessible on the web, and also available through the supplementary materials of this thesis[23].

All the code written for the BioContext, the wrappers for several tools, and the tools that we developed will be available at http://www.biocontext.org/.

23 Available at www.cs.man.ac.uk/~sarafraf/thesis-supplementary.html

Appendix D provides a list of data and code from each stage. For more details about the size of each data set, see Chapter 6 .

Chapter 6
Large-scale event extraction: data and evaluation

We applied BioContext, our integrated system, to extract events and their context, to MEDLINE (2011 baseline files, containing 10.9 million abstracts) and to the open-access subset of PMC (downloaded May 2011, containing 235,000 full-text articles). In this chapter we present, evaluate, and discuss the data resulted from this experiment.

To evaluate the pipeline of text mining tools, we also evaluate the performance of each of these components individually, in order to measure the impact that each of the more complex components have on the data as it moves through the pipeline.

6.1 Evaluation method

6.1.1 Evaluation metrics and approach

From an NLP perspective, it is important whether the textual elements indicating the event are correctly identified. There are several ways to assess the equality of these elements. The textual boundaries of the extracted event triggers and participants could match those of the gold standard ones either approximately or exactly. The nested events could be evaluated recursively, or only based on the highest level event.

From a biological perspective, mention level evaluation might not be very useful, as it is only interesting to know whether a particular event is described in a document, regardless of the exact phrase used to describe it. Textual triggers may be of little importance, and the terms used to refer to the biological entities involved in an event are variable.

To compromise for these considerations, we use slightly different methods for evaluating the event extraction task than for the aggregate

analysis. In the event extraction task (Chapter 3), we count an extracted event as a true positive if its type, trigger and all participants are correctly identified. Similarly, in the negation and speculation extraction task, a true positive represents a correctly identified negated (or speculated) event and a false negative is a negated (or speculated) event reported incorrectly as affirmative (or asserted). Here we present an aggregate analysis of results obtained as described in Chapter 5. We focus on the biological significance of the data, and allow the textual triggers of the event to vary.

6.1.2 Evaluation corpora

In the evaluation of BioContext using automatically extracted entities as opposed to gold standard entities, we noticed that many false positive results are created by entities missing in the BioNLP'09 corpus. The obvious thing to do was to consider these entities as false positive instances from the NER stage, but on closer examination we found that many of these entities are also present in the original GENIA corpus (see Section 2.6 for the differences between the two corpora).

The removal of those entities in the construction of the BioNLP'09 corpus was based on the argument that they do not strictly meet the "gene or gene product" definition. Many of these entities are protein complexes (such as *NF kappa B*) or other entities that behave (both biologically and linguistically) similarly to ordinary genes and proteins. Furthermore, these entities are typically detected by gene recognition systems, and participate in molecular events which are described in the biomedical literature using similar linguistic expressions, and are potentially interesting to biologists.

We therefore constructed a new corpus, referred to as the **B+G corpus** hereafter, combining the BioNLP'09 gold annotations with the subset of the GENIA corpus that most closely resembles the BioNLP'09 construction criteria, but also accounts for the omitted protein complex entities. The B+G corpus includes all the entities from the BioNLP'09 corpus plus the entities

from the GENIA corpus with the *protein molecule* and *protein complex* tags (see Section 2.6).

The events included in the B+G corpus are all the events in the BioNLP'09 corpus, in addition to a subset of the GENIA events. To construct this subset, we select any GENIA event whose participants are already included in the B+G corpus (be it an entity or an event) and whose type is one of the following GENIA event types: *Positive regulation*, *Negative regulation*, *Regulation*, *Gene expression*, *Binding*, *Transcription*, *Localization*, *Protein catabolism*, *Protein amino acid phosphorylation*, and *Protein amino acid dephosphorylation*. The last two event types, *Protein amino acid phosphorylation*, and *Protein amino acid dephosphorylation*, together construct the event class Phosphorylation in the BioNLP'09 corpus.

There are a number of cases where mapping an event from the GENIA corpus to its derived event in the BioNLP'09 corpus is not straightforward. For example, in a number of cases, the original and the derived events differ in their trigger mention and therefore could be two different events. So, strictly speaking, both events should appear in the B+G corpus. However, upon further manual investigation of all the respective sentences, we realised that the two seemingly different events are in fact the same, and the BioNLP'09 corpus creators have moved the trigger of a GENIA event to a different word. In such cases, only one of the events are included in the B+G corpus, with priority given to the BioNLP'09 corpus.

Every event in the GENIA corpus has "assertion" and "uncertainty" attributes. Assertion is a binary attribute with possible values of "exist" and "non-exist" which corresponds to our definition of negation, whereas uncertainty is a tertiary attribute corresponding to speculation, with the three possible values of "certain", "probable", and "doubtful". In the BioNLP'09 corpus, an event can be "negated", corresponding to the "non-exist" attribute in the GENIA corpus. Independently, it can be "speculated", corresponding to the union of probable and doubtful events. We stay with the binary classification

of the BioNLP'09 corpus, grouping the probable and doubtful events.

Table 6.1 shows the distribution of event types in the B+G corpus. Compared to the statistics for the combined BioNLP'09 training and development corpora, of the 14,781 events in the corpus, a total of 2,607 belong to the set of abstracts appearing in the development set of the BioNLP'09 corpus.

Event type	**Number of events in the BioNLP'09 training + development data sets**	**Number of events in the B+G corpus**
Gene expression	2,094	2,399
Localization	318	497
Transcription	658	683
Protein catabolism	131	146
Phosphorylation	216	252
Binding	1,136	1,711
Regulation	1,134	1,608
Positive regulation	3,465	5,457
Negative regulation	1,258	2,028
Total	**10,410**	**14,781**

Table 6.1: Summary of the events in the B+G corpus

To evaluate and compare the event extraction tools, we use the BioNLP'09 corpus wherever the gold annotated entities are used as input. But when the tools are using automatically extracted entities as part of a pipeline, the B+G corpus is used for evaluation. In the following sections we report the results of the different tasks on the above corpora.

As no gold annotated corpus exists for anatomical and species named entities on a mention level, we randomly selected 100 events that are associated with entity names by methods described in Section 5.1.6 for post-hoc evaluation. Similarly, we selected 100 events that were inferred from the extracted events (see Section 5.1.7) for post-hoc evaluation of event inference methods.

6.2 NER

Table 6.2 shows the number of gene/protein entities (both entity mentions and distinct entities) extracted from the MEDLINE and PMC data sets. We also consider entities recognised by both (intersection) or either (union) of the two recognisers. The two corpora (MEDLINE and PMC) have an overlap, consisting of the articles whose abstracts are listed in the MEDLINE corpus, and whose full text is present in the PMC corpus. Throughout the results reported in this chapter, wherever joint MEDLINE + PMC results are reported, this overlap has been taken into account, and only reported once.

GNAT was additionally adapted to also return non-normalized entities whenever those were detected by BANNER but could not be linked to identifiers. In both the GeneTUKit and the intersecting data, all entries were normalised (since GeneTUKit only reports normalised mentions). Of the 80,003,072 extracted gene mentions in the union set, 10,261,208 (12.8%) were not normalised, all of which were produced by GNAT. Both the GeneTUKit data and the intersecting data contain only normalised entities, linking a mention to its database identifier.

We also report the number of distinct gene/protein names that were recognised. For this purpose, the gene mentions that could not be normalised were grouped together based on surface string equality.

Tool	Gene entity mentions			Distinct entities		
	MEDLINE	PMC	MEDLINE + PMC	MEDLINE	PMC	MEDLINE + PMC
GNAT	35,910,779	12,729,471	48,050,830	227,809	129,244	253,929
GeneTUKit	47,989,353	19,217,778	66,431,789	258,765	143,706	287,218
Intersection	26,281,266	8,638,823	34,479,547	224,604	125,763	249,932
Union	57,618,866	23,308,426	80,003,072	261,412	146,552	290,557

Table 6.2: Gene and gene product recognition counts in MEDLINE and PMC

The number of gene mentions and distinct genes recognized by GNAT and GeneTUKit in MEDLINE and PMC. In the MEDLINE + PMC columns, the overlap is only reflected once.

The evaluation results for the gene/protein named entity recognition systems on the B+G corpus are shown in Table 6.3. Both precision and recall are in the same range as what has previously been reported for common recognition tools (BANNER: 85% P, 79% R; ABNER: 83% P, 74% R (Leaman et al. 2008)). Studying the FP and FN errors suggested that some of the more common categories of errors include incorrect dictionary matches due to acronym ambiguity, incomplete dictionaries, and incomplete or incorrect manual annotations of the gold-standard data.

We note that there is currently no gold-standard corpora available for mention-level gene normalisation.

	P	R	F1
GNAT	79.8%	83.7%	81.7%
GeneTUKit	72.2%	79.1%	75.5%
Intersection	82.8%	70.4%	76.1%
Union	71.4%	92.0%	80.4%

Table 6.3: Entity recognition performance on the B+G corpus

Gene/protein named entity evaluation results on exactly 3,000 instances in the B+G corpus.

We used LINNAEUS to recognise anatomical location mentions and

species name mentions from the articles. A summary of the number of entities extracted from MEDLINE and PMC can be found in Table 6.4. These results have not been evaluated. For the evaluation of LINNAEUS see (Gerner et al. 2010b).

Tool	MEDLINE	PMC	MEDLINE + PMC
Anatomical location	47,002,254	9,656,994	56,659,248
Species names	33,187,566	3,771,333	36,958,899

Table 6.4: Species and anatomical entity recognition counts in MEDLINE and PMC

The number of species and anatomical location mentions recognized by LINNAEUS in MEDLINE and PMC.

6.3 Event extraction

In this section we summarise the results and evaluations of TEES and EventMiner as well as the union and intersection of their outputs. The evaluation on a corpus with gold-standard gene and protein entities are reported in (Kim et al. 2009), where TEES achieved precision/recall/F-score of 58%/47%/52%, and these measures for EventMiner were 54%/28%/37%. Here, we evaluate these systems in a real-world situation using automatically extracted genes/proteins as input. For this purpose, we define a true positive as before (as presented in Section 4.1.1), but with one additional condition:

5. The entity participants are true positives, and approximately match boundaries with the gold participants.

6.3.1 TEES

We executed the Turku Event Extraction System (TEES) on the B+G corpus (introduced in section 4.1.2), using genes that were extracted by the integrated

gene NER system (union). The results of that evaluation can be seen in the Table 6.5.

Number of true positives (TP)	655
Number of false positives (FP)	1142
Number of false negatives (FN)	1949
Precision	36.4%
Recall	25.1%
F-score	30.0%

Table 6.5: Evaluation of TEES as deployed locally on the B+G corpus

Table 6.6 shows the event-type specific evaluation of the TEES data filtered by our automatically extracted gene and protein entities.

Type	**TP**	**FP**	**FN**	**p (%)**	**r (%)**	**F (%)**
Gene expression	340	188	78	64.3	81.3	71.8
Localization	58	23	23	71.6	71.6	71.6
Phosphorylation	49	14	11	77.7	81.6	79.6
Transcription	60	31	25	65.9	70.5	68.1
Protein catabolism	22	4	4	84.6	84.6	84.6
Class I total	**529**	**260**	**141**	**67.0**	**78.9**	**72.5**
Binding **(Class II)**	136	246	249	35.6	35.3	35.4
Regulation	98	129	149	43.1	39.6	41.3
Positive regulation	511	554	499	47.9	50.5	49.2
Negative regulation	135	196	184	40.7	42.3	41.5
Class III total	**744**	**879**	**832**	**45.8**	**47.2**	**46.5**
All	**1409**	**1385**	**1222**	**50.4**	**53.5**	**51.9**

Table 6.6: Type-specific evaluation of the TEES data on B+G

This event extraction evaluation corresponds to the second column of Table 6.7, showing the break-down of the TEES performance using automatically extracted gene and protein entities as part of this research.

The Department of Information Technology in the University of Turku have also provided a database of the events (Björne et al. 2009). extracted from

the MEDLINE abstracts. From this database, we selected only the abstracts that are included in the B+G corpus. We evaluated the data, once filtering those events that reference any gene/protein entities that are not in our extracted genes, and a second time just evaluating the data as presented. The results of both evaluations can be found in Table 6.7.

Note that by filtering out the events referring to the entities that are missing from our extractions, we are missing a small number of events. Because of this, recall is reduced slightly. However, it does not have a significant effect on the precision or the F-score.

	TEES data on extracted genes	**Original TEES data**
Number of true positives (TP)	1409	1433
Number of false positives (FP)	1385	1427
Number of false negatives (FN)	1222	1192
Precision	50.4%	50.1%
Recall	53.6%	54.6%
F-score	51.9%	52.2%

Table 6.7: Evaluating the data released by TEES developers

The evaluation was performed on the B+G corpus. The second column shows the evaluation statistics when the original data was filtered based on the automatically extracted genes. The third column shows the evaluation of the TEES data as presented in the original database.

6.3.2 Evaluation of EventMiner

The source code of EventMiner has been provided by the developers for the purpose of the development of this project. Providing automatically extracted gene and protein entities for the software and running it on the B+G corpus we achieved the results summarised in Table 6.8. Table 6.9 shows the event-type

specific evaluation of this data. Despite using automatically extracted entities, the results of this improved version of the system are higher than those reported as part of the BioNLP'09 Shared Task (overall F-score 46% vs. 34.6%). But it should be noted that the evaluation datasets are different: our results are evaluated on the B+G corpus that are effectively the same abstracts and BioNLP'09 development corpus which was also used for training and tuning of the systems; the Shared Task evaluation was performed on the BioNLP'09 test corpus, which is not publicly available.

Number of true positives (TP)	1201
Number of false positives (FP)	1428
Number of false negatives (FN)	1438
Precision	46%
Recall	45%
F-score	46%

Table 6.8: Evaluation of EventMiner on the B+G corpus

The evaluation is performed on the B+G corpus using automatically extracted genes.

Type	TP	FP	FN	p (%)	r (%)	F (%)
Gene expression	326	240	89	57.5	78.5	66.4
Localization	62	30	21	67.3	74.6	70.8
Phosphorylation	47	31	13	60.2	78.3	68.1
Transcription	54	52	32	50.9	62.7	56.2
Protein catabolism	22	13	4	62.8	84.6	72.1
Class I total	**511**	**366**	**159**	**58.2**	**76.2**	**66.0**
Binding **(Class II)**	140	269	258	34.2	35.1	34.6
Regulation	56	160	193	25.9	22.4	24
Positive regulation	390	485	619	44.5	38.6	41.4
Negative regulation	104	148	209	41.2	33.2	36.8
Class III total	**550**	**793**	**1021**	**40.9**	**35.0**	**37.7**
All	**1201**	**1428**	**1438**	**45.7**	**45.5**	**45.6**

Table 6.9: Type-specific evaluation results for the EventMiner data on the B+G corpus

This event extraction evaluation shows the break-down of the EventMiner performance using automatically extracted gene and protein entities as part of this research.

6.3.3 Merging the outputs

The evaluation scores achieved by the two event extractors (see Table 6.10) show the best precision of 66% (for intersection) and the best recall of 62% (for union).

TEES still provides the best balance between precision and recall (52%). Still, these results differ from previously reported precision levels for TEES at 64% (Björne et al. 2010). However, the evaluation methods were different, making comparisons difficult: in the evaluation of Björne et al., 100 events were selected randomly for post-hoc manual verification, rather than being compared to a gold-standard corpus. Their definition of “entity” was also slightly different, allowing “cells, cellular components, or molecules involved in biochemical interactions” to be counted as true positive, but not necessarily contributing to false negative, as recall is not considered in post-hoc

evaluation. Indeed, many FP and FN errors by the event extractors were due to incorrect entity recognition that propagated to event FPs or FNs, sentences that were particularly complex linguistically or semantically, and incomplete manual annotation of the corpora.

	P	**R**	**F1**
TEES	50.4%	53.6%	51.9%
EventMiner	45.6%	45.5%	45.5%
Intersection	66.2%	36.6%	47.1%
Union	41.3%	62.0%	49.6%

Table 6.10: Overall event extraction evaluation

Evaluation on the B+G corpus with 2,607 instances.

Type	**TP**	**FP**	**FN**	**p (%)**	**r (%)**	**F (%)**
Gene expression	295	103	114	74.1	72.1	73.1
Localization	53	16	28	76.8	65.4	70.6
Phosphorylation	44	12	15	78.5	74.5	76.5
Transcription	48	22	37	68.5	56.4	61.9
Protein catabolism	20	1	6	95.2	76.9	85.1
Binding	81	96	302	45.7	21.1	28.9
Regulation	44	30	204	59.4	17.7	27.3
Positive regulation	294	149	697	66.3	29.6	41.0
Negative regulation	70	55	241	56	22.5	32.1
All	**949**	**484**	**1644**	**66.2**	**36.6**	**47.1**

Table 6.11: Evaluating the intersection of event extraction outputs

Type-specific event extraction evaluation results on the intersection data.

Type	TP	FP	FN	p (%)	r (%)	F (%)
Gene expression	370	340	54	52.1	87.2	65.2
Localization	67	39	16	63.2	80.7	70.8
Phosphorylation	52	33	9	61.1	85.2	71.2
Transcription	66	61	20	51.9	76.7	61.9
Protein catabolism	24	15	2	61.5	92.3	73.8
Binding	195	419	205	31.7	48.7	38.4
Regulation	110	264	138	29.4	44.3	35.3
Positive regulation	606	895	422	40.3	58.9	47.9
Negative regulation	171	292	150	36.9	53.2	43.6
All	**1661**	**2358**	**1016**	**41.3**	**62.0**	**49.6**

Table 6.12: Evaluating the union of event extraction outputs

Type-specific event extraction evaluation results on the union data.

Table 6.13 presents the number of events extracted from the corpora. In addition to the number of event mentions, we provide an estimate for the number of distinct events. For this purpose we define two events to be the same if:

- The events are of the same type.
- They involve the same normalised gene entities. If non-normalised genes are involved, the gene mention strings must match. If more than one entity is involved, all pairs must match.
- Either no anatomical entity is associated with either of the two events, or if one event is associated with an anatomical entity, the other event should also be associated with an entity normalised to the same anatomical location. In the case of non-normalised anatomical entities, the entity mention strings must match.
- They are both affirmative, or both negated.
- They are both asserted, or both speculative.
- If any of the participants of the events is another event, those nested

events should also match recursively.

Tool	Event mentions			Distinct events		
	MEDLINE	PMC	MEDLINE + PMC	MEDLINE	PMC	MEDLINE + PMC
TEES	19,406,453	4,719,648	23,856,554	6,570,824	1,804,846	7,797,604
Eventminer	18,988,271	4,010,945	22,737,258	6,502,371	1,588,178	7,539,364
Intersection	9,243,903	1,331,456	10,455,678	3,080,900	2,676,257	3,424,372
Union	29,150,821	7,399,137	36,138,134	9,635,566	573,903	11,442,462

Table 6.13: Literature-scale event extraction counts

The number of event mentions and distinct events extracted by TEES and Eventminer in MEDLINE and PMC . In the MEDLINE + PMC columns, the overlap is only reflected once.

Each distinct event represents a number of event mentions in the literature, referred to as **supporting mentions** for that distinct event (see also Section 5.2). The sentence in which this event occurs is called the **supporting sentence**. For example, for the distinct event of "negated positive regulation of the expression of IFN-gamma caused by IL-2", the sentences shown in Example 6.1 will be a supporting sentence each.

Example *6.1*.

(a) *"Neither IL 1* ***nor*** *IL 2 alone induced IFN-gamma production in purified T lymphocyte cultures."*

(From PMID 3086435, events extracted by BioContext)

(b) *"IL 2 had the ability to restore lytic activity to PMA-treated cells but did* ***not*** *induce IFN gamma production."*

(From PMID 3930891, events extracted by BioContext)

Some events are commonly reported and will have a high number of supporting mentions and supporting sentences, whereas others are only reported a few times across the literature. Table 6.14 shows the maximum

number of supporting mentions as well as the average for each event type.

Type	Total event mentions	Percentage of total events	Distinct events	Maximum count of the supporting mentions for a single event	Average number of supporting mentions per event
Gene expression	9,636,642	26.4%	1,785,161	**25,561**	5.40
Localization	2,051,035	5.6%	488,738	18,721	4.20
Phosphorylation	747,083	2.0%	141,436	8,069	5.28
Protein catabolism	348,031	1.0%	105,945	2,011	3.29
Transcription	732,827	2.0%	247,994	2,322	2.96
Binding	5,392,795	14.8%	1,900,223	6,868	2.84
Regulation	3,686,616	10.1%	749,889	8,658	4.92
Positive regulation	8,948,707	24.5%	1,293,502	18,011	6.92
Negative regulation	5,006,222	13.7%	754,930	12,014	6.63
Total	**36,549,958**	**100%**	**7,467,819**	**-**	**4.89**

Table 6.14: Supporting mention counts extracted by BioContext

Number of event mentions and distinct events, and the number of supporting mentions for each distinct event in the extracted data (the union set).

We compared the composition of the automatically extracted data from the entire available literature with that of the B+G corpus with regard to the distribution of the event types. Figure 6.1 shows this comparison broken down by type, and shows that the distribution of different types in the B+G corpus is reflected in the large-scale event extraction results.

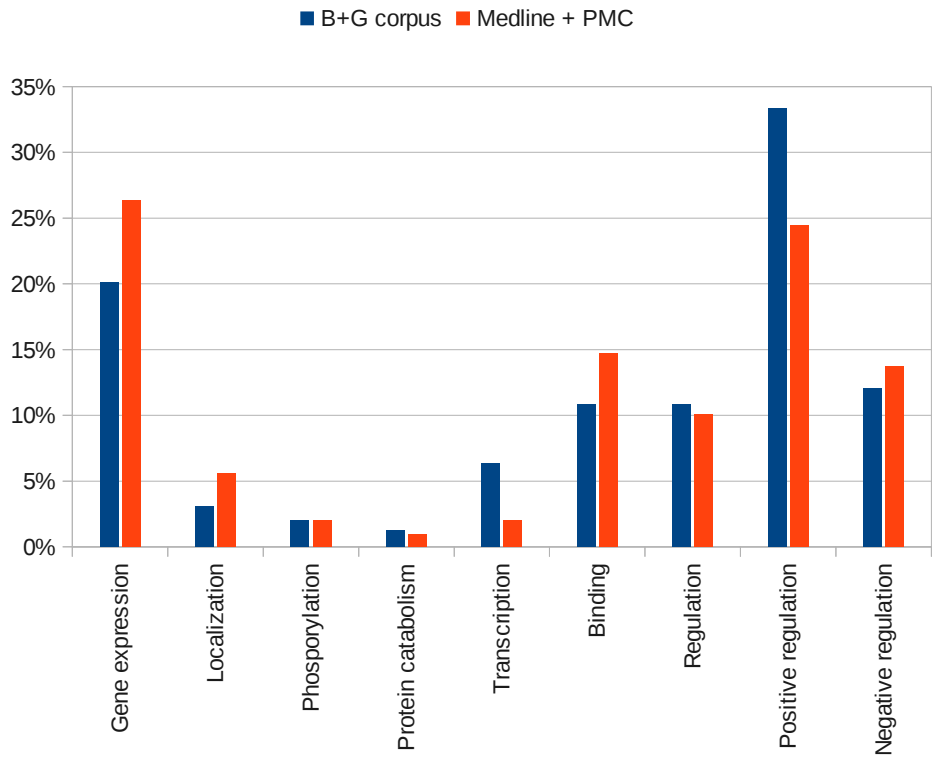


Figure 6.1: Type-specific comparison between the B+G corpus and the extracted events

6.3.4 Event inference

Of the 80 million gene/protein mentions in the MEDLINE and PMC union sets, 11.3 million (14%) were part of enumerated groups as detected by our patterns (see Table 5.1), i.e. joined with a conjunctive structure, and presumably contributed towards inferring events associated with each of the entities in the enumerated group.

Of the 36.1 million events in the MEDLINE and PMC union sets, 1.05 million (2.9%) were created through the event inference methods (see Section 5.1.7). While the percentage of events inferred is low, the absolute number of events is still large enough to show the utility of the method.

Post-hoc manual inspection of 100 randomly selected inferred events showed a precision level of 44%. Most false positive results were due to the original entities being wrong in the first place, or an incorrect event detection, rather than an error in the enumeration detection. Example 6.2 shows one such FP instance (incorrect entity *"factor"*) as well as a TP instance of event inference based on enumeration. The underlined entities have been recognised as belonging to the same enumerated group.

Example 6.2

TP: *"Many cartilage matrix proteins or domains such as collagen types II, IX, and XI, GP39, AG1, VG1, and LP are potential antigens that might induce polyarthritis in susceptible animals."*

(From PMID 12951872)

FP: *"Somatostatin was first identified as a hypothalamic factor which inhibits the release of growth hormone from the anterior pituitary (somatotropin release inhibitory factor, SRIF)."*

(From PMID 11430867)

A total of of 57.1 million anatomical mentions were found in the union set of MEDLINE and PMC. Out of this number, 4.0 million (7.0%) were enumerated. Example 6.3 shows an instance of an inferred event (gene expression of Neurturin) that was initially associated with *"retina"*, and through the event inference method also reported in *"photoreceptor"*.

Example 6.3

TP: *"Neurturin mRNA expression was modulated through normal postnatal retinal development and was localized primarily to the inner retina and photoreceptor outer segments.."*

(From PMID 10067959)

6.3.5 Confidence evaluation

To evaluate how well our confidence scoring (see Section 5.3) corresponds to the actual quality of extraction, we measure the confidence of extraction on the BioNLP'09 data, and compare it to the precision of extraction against gold annotations.

Since we calculated the confidence scores based on the precision of different components in the first place, we expect to see such a correspondence anyway. However, since the formula to calculate confidence from different precision measures was simple and heuristic, it is reasonable to evaluate how well it reflects the quality of the extracted data.

We ran the event extraction pipeline on the 950 abstracts of the collective BioNLP'09 data sets (training + development), and calculated the confidence for every extracted event. The graph in Figure 6.2 shows the distribution of different confidence scores.

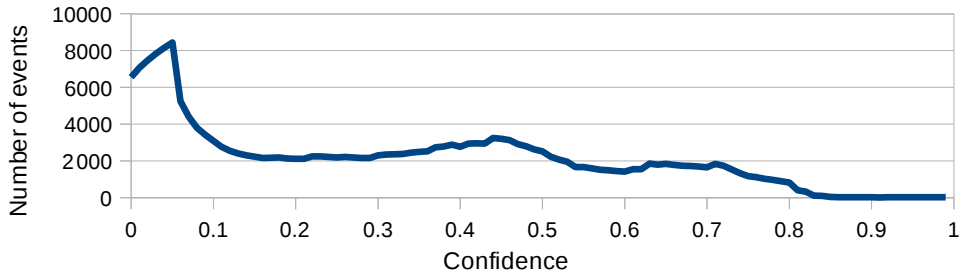


Figure 6.2: The number of events against the confidence scores

This is reported on the training and development gold annotated BioNLP'09 data sets.

As we can see in this graph, there are very few events (although not zero) having a confidence of above 0.8. We will subsequently see that this sparseness of data within the high-end confidence area causes some irregularities in that region.

We expect a steady increase in precision when we look at the precision of the events in intervals with increasing confidence. We calculate the

precision of the extracted event amongst events that fall in the same confidence neighbourhood. The interval size in which we calculate the precision is 0.1 confidence points, and the intervals are 0.01 confidence points apart.

The graph of Figure 6.3 shows the precision of every interval. As we noted earlier, the data with confidences above 0.8 is very sparse, and therefore does not strictly follow the increasing pattern of the precision. But we are still observing the correlation between the confidence levels and the precision, suggesting that we can assume the extracted events of higher confidence values to have better quality.

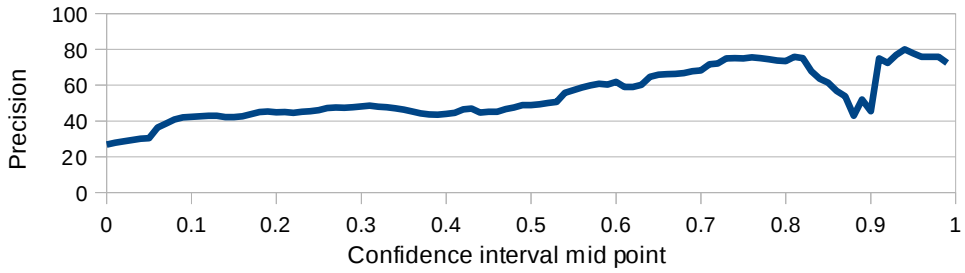


Figure 6.3: Precision against confidence scores

The precision of the events of confidence falling within a window of 0.1 precision points. The intervals are overlapping, with a midpoint every 0.01 interval point. This is reported on the training and development gold annotated BioNLP'09 data sets.

We also calculate the cumulative precision, recall, and F-scores as we start by only considering the extracted events of the highest confidences, and gradually adding those with lower qualities. The cumulative graphs are shown in Figure 6.4. As expected, the precision decreases as lower quality data is added, whilst recall increases, as more FN results are detected by adding more data.

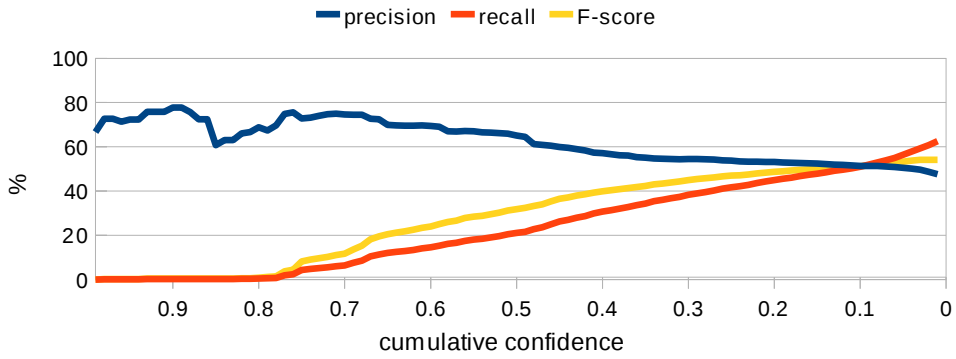


Figure 6.4: Quality of extracted data against cumulative confidence

Precision, recall, and F-scores of extracted data as we cumulatively add events of decreasing confidence.

It is interesting to note that the F-score is constantly increasing, even beyond the P/R balance point. This indicates that that the increase in recall is large enough to compromise the decrease in precision, and therefore if a high F-score is desirable, it is beneficial to include all the extracted results in the system output, even those with lower confidences.

6.3.6 Discussion

As a part of this research, we have presented an integrated text mining framework, BioContext, and the data produced by it after applying to 10.9 million abstracts in MEDLINE and 235,000 full-text articles in the open-access subset of PMC. The data contains 36.1 million event mentions, which represnt 11.4 million distinct events discussing biomedical processes involving genes and proteins. The event participants are linked to the Entrez Gene database whenever such a normalisation was possible. The data contains contextual information about the events including the associated anatomical locations and whether they are reported as negated or speculative.

In addition to the gene/protein entities and the events, the process of extracting events from MEDLINE and PMC also produced large volumes of other intermediary data that should prove useful to the biomedical text-mining

community. This data includes 70.9 million LINNAEUS species entity mentions, 57.1 million anatomical entity mentions, and 133 million parsed sentences from each of the Gdep, Enju and McClosky-Charniak parsers.

Compared to the previously released dataset of 19.2 million events extracted from MEDLINE by TEES (Björne et al. 2010); (Van Landeghem et al. 2011), the data set described here provides additional data in a number of ways, including the addition of full-text PMC corpus, negation and speculation detection, anatomical association, and normalisation of genes and proteins to species-specific identifiers.

We observed that locally running and evaluating a publicly available tool for event extraction (TEES) results in significantly lower results than evaluating the data resulting from running the same tool provided by the developers, as we were unable to reproduce the levels of precision and recall that was originally reported. This could be due to the exceptional termination of the tool in around 20% of the documents when run locally, caused by any configuration disparity between our system and the developers'.

Evaluations performed using the B+G corpus are limited by the fact that it was derived from the set of MEDLINE abstracts containing the MeSH terms *"humans"*, *"blood cells"*, and *"transcription factors"* by the BioNLP'09 corpus curators. Because of this, evaluation results may not be completely representative for MEDLINE as a whole. Despite this discrepancy, the distribution of event types in the two corpora are similar. The inferred and anatomically associated events used for evaluation of those stages were selected completely randomly, and while the sample size was limited, they should provide a representative sample.

Looking at the evaluation results as the data moves through the different stages of the pipeline, the impact of the multi-tiered nature of the system becomes evident. Many FPs and FNs that occur in the NER stage are propagated to the event extraction stage, and additional FPs and FNs introduced there are in turn propagated to the context association stage. In

other words, errors (in particular those occurring early in the pipeline) can have a large impact on the final results.

Some text-mining systems are evaluated as part of challenges that eliminate these issues by providing gold-standard data for the earlier stages (typically NER). This allows researchers to focus on a particular task (e.g. event extraction) rather than having to divide their attention between both NER and event extraction. However, it also means that any evaluation result coming out of these challenges needs to be adjusted for more realistic constraints when used for information extraction on a large scale where gold data is not available. We note the drop in precision and recall when applying these tools in a realistic environment.

A common theme in the evaluation of text mining tools is the balance between precision and recall. Applications prioritise and value precision and recall differently. Looking specifically at the evaluation results for gene/protein NER and event extraction, the utility of merging data from multiple similar tools becomes evident: by applying multiple different tools and creating datasets from both the intersection and the union of the extracted data, we can shift this balance between precision and recall in different directions, depending on how the data is used.

In addition, the use of multiple tools for the more challenging aspects (gene/protein NER and event extraction), allows users to handle data differently depending on whether it was extracted by e.g. both event extractors or whether only a single tool found it. The differences between the tools is evident from the difference between the union and intersection data sets.

After performing the large-scale data extraction experiments, it is clear that text-mining on this scale comes with a range of challenges, beyond the technically relatively simple matter of having access to powerful enough computational systems. Here, we mention a few of these challenges, and our approach to addressing them.

Most text-mining software operates on plain-text files. Because of this, it

can be tempting to store documents as text files, but it quickly became clear that if this is done, the file system becomes too much of a bottleneck in the computational process due to the huge number of files.

We stored the data in a databases running on a powerful server instead of the file system which mitigated the problem, but was not perfect as it still remained the bottleneck for some tools. Distributing the documents on multiple database servers each having local storage should provide further mitigation for the problem.

While the documents in MEDLINE and PMC are generally well-structured, there are always exceptions. Although these outliers are very rare in relation to the total number of documents, the very large number of documents in MEDLINE and PMC still means that odd outliers become significant problems. Examples we have found include documents over 300 pages long (causing some tools to crash when running out of memory, and others never to terminate due to inefficient algorithms), documents containing programming source code (causing every single grammatical parser to crash), and seemingly "innocent" documents that for some reason give rise to hundreds of thousands false positive events, which in turn crashes downstream tools. Document issues that are more common include PMC documents with embedded TeX code or non-ASCII characters as the parsers typically cannot handle either.

We have implemented robust general error detection and recovery methods within TextPipe to address problems with unusual processing time, frequent crashes and other external problems, such as network connection time-outs or machine failures.

We note that processing for the number of tools and documents described in this research is computationally very heavy. For such a large-scale task, processing time requirements depend on a range of factors such as the speed of the computational hardware available, potential database or network bottlenecks, etc., making such estimates difficult to make. We estimate that a matter of months would be a fairly accurate approximation, using a cluster of

100 processing cores. This assumes that everything works flawlessly and no re-computation is necessary, which may not be the case in practice.

It is unfortunate that only roughly 2% of MEDLINE entries have full-text articles that are available for text-mining. If the open-access subset of PMC is a representative sample of all full-text articles, we would expect that about 400 million further events are mentioned in full-text articles, but unavailable for automatic extraction due to copyright restrictions.

This work provides a foundation for future work: Protein complexes are currently not linked to any protein complex knowledge bases. Additionally filtering of results based on the journal or document subject area could improve the performance.

6.4 Context association

6.4.1 Anatomical association evaluation

Of the 36.1 million events in the MEDLINE and PMC union sets, 13.5 million events (37.5%) could be associated with an anatomical entity.

Post-hoc manual inspection of 100 randomly selected events associated with anatomical entities showed a precision of 34%. Here, we consider an event a true positive only if all the components were extracted correctly. Therefore this precision refers to the cumulative precision of all the automated components, and not only to the precision of the anatomical entity association phase.

While not evaluated, this value is expected to be higher if the events are constrained to the intersection set, similar to the precision levels in Tables 6.11 and 6.12 that were higher for the intersection set than for the union set.

Although not theoretically considered a comprehensive way to evaluate the performance of a system, post-hoc manual examination of predictions have been used by researchers in the absence of gold-standard annotations. In this method, a (usually small) number of predictions are randomly selected and analysed to calculate the percentage of the false positive results. With this

method only the precision and not recall can be approximated, as it does not take into account any false negative instances. In addition, the error analysis in this method can only include the false positive and not the false negative instances.

Moreover, post-hoc examination of the predictions introduces bias towards the automatically extracted information by considering instances that may look reasonable and are considered as true positive whereas—had they been annotated independently—they would not have been annotated as positive instances. This effect will be stronger if several properties of the gold and the predicted instances have to match in order to be considered true positive. In the case of the events, these properties include trigger word boundaries, event type, theme and cause mentions, and the association between them. Amongst the related previous work, (Björne et al. 2010) have used this method to evaluate the precision of their system by manual examination of the predictions.

6.4.2 Negation and speculation extraction as part of context extraction

Evaluation of event extraction results after performing negation and speculation detection by Negmole can be seen in Table 6.15. Here, events are required to have both their negation and speculation status correctly identified to be classified as a true positive.

	P	**R**	**F1**
Intersection	62.6%	34.6%	44.6%
Union	38.8%	58.3%	46.6%

Table 6.15: Evaluation of event extraction after processing by Negmole

Evaluated on 2,607 instances, on the B+G corpus.

Relatively small differences in data quality are observed before and after applying Negmole. This is expected, since only a small subset of events are affected by negation and/or speculation. Table 6.16 shows the numbers and

percentages of negated and speculated events for each event type. Interestingly, regulation events show specially high ratios of negation and speculation.

Of the 36.1 million events in MEDLINE and PMC, 1.49 million (4.1%) are negated, and 1.25 million (3.5%) are speculative. The negation and speculation ratios are slightly lower than those of the combined BioNLP'09 training and development sets (6.8% and 5.3%, respectively).

Event type	Number of negated events	% of negated events	Number of speculated events	% of speculated events
Gene expression	335,774	3.47%	297,992	3.09%
Localization	31,175	1.51%	41,244	2.01%
Phosphorylation	11,406	1.52%	8,606	1.15%
Protein catabolism	5,898	1.69%	5,488	1.58%
Transcription	21,429	2.91%	25,222	3.44%
Binding	172,068	3.14%	132,316	2.45%
Regulation	478,383	12.86%	352,139	9.55%
Positive regulation	351,368	3.86%	298,939	3.34%
Negative regulation	80,001	1.59%	91,187	1.82%
Total	1,487,502	4.06%	1,253,133	3.43%

Table 6.16: The number and percentage of negated and speculated events in MEDLINE and PMC

It has been observed previously (Cohen et al. 2010) that the incidence of negation (measured by the distribution of the words "*no*", "*not*", and "*neither*") is significantly different between the full-text articles and abstracts. They reported a higher incidence of these words (5.3 per thousands tokens of text) in article bodies, compared to their incidence in abstracts (3.8 per thousands tokens of text).

Table 6.17 shows the distribution of negated and speculated events in MEDLINE and PMC. There are very few events that are both negated and speculated. Therefore, these events hardly affect any aggregate analysis presented in this chapter.

Polarity & certainty of events	MEDLINE	% MEDLINE	PMC	%PMC	Total	% Total
Affirmative & certain	26,406,361	92.16	7418250	93.94	33,824,611	92.54
Negative & certain	1,203,376	4.20	268838	3.40	1,472,214	4.03
Affirmative & speculated	1,032,393	3.60	205452	2.60	1,237,845	3.39
Negative & speculated	10,946	0.04	4342	0.05	15,288	0
Total	**28,653,076**	**100**	**7896882**	**100**	**36,549,958**	**100**

Table 6.17: Distribution of negated and speculated events on MEDLINE and PMC

Figure 6.5 shows the frequency distribution of negated and speculated events for all event types. Due to the small numbers of events that are both negated and speculated, we have grouped them together with the speculated category. Others have previously considered the three-class categorisation of events into ‘affirmed and certain’, ‘negated and certain’, and ‘speculated’ categories (Elkin et al. 2005).

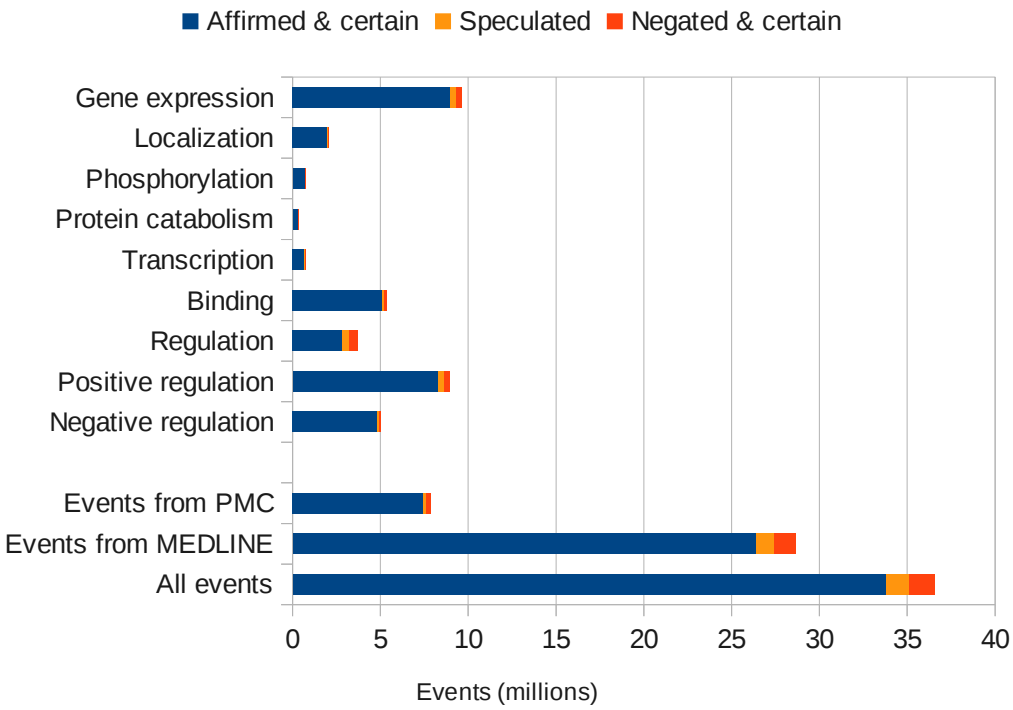


Figure 6.5: The frequency distribution of negated and speculated events on MEDLINE + PMC

The speculated events contain both polarities. These numbers have been grouped together because of the negligible size of the set of events that are both speculated and negated.

Our analysis shows that 4.2% of the events reported in MEDLINE are negated, as opposed to only 3.4% of the events in the the PMC corpus. The difference is statistically significant ($p<0.0001$). This finding seems to contradict the previous findings that there is more negation in the body of journal articles compared to the abstract. However, Cohen et al. only analysed the occurrence of negation cues, which we have seen that does not necessarily indicate the reporting of negative results. This difference could be due to the article body text using less direct and more elaborate prose style, including more use of figurative negation, but not necessarily reporting more negative results. On the other hand, of the total negative results reported in the literature, a higher proportion of them are likely to be highlighted in the abstract of the

articles.

The proportion of speculated events in MEDLINE (3.6%) is significantly higher ($p<0.0001$) than that of PMC (2.6%). This shows that authors speculate about their own or others' findings in the abstract more often than they do in the body of the articles, possibly because the article body presents speculated findings in the context of established facts. These results confirm previous findings that the composition and characteristics of full text scientific literature differs from that of the abstracts, and as text mining on scientific literature moves from abstracts towards more full text approaches, these discrepancies will cause challenges and opportunities for further findings.

It is worth noting that with such large numbers of instances, almost any trend that is observed would have statistical significance. On the one hand, this is meaningful, since the observation is done on the entire available data, as opposed to a sample. On the other hand, care should be taken when analysing such statistically significant trends, and the effect of the large data size should not be overlooked.

Figure 6.6 shows the percentages of negated and speculated events, normalised for the size of each event type. Although the proportions differ across all event types, regulation events (but not positive regulation or negative regulation events) are dramatically more frequently detected as negated or speculative. This could indicated that the authors tend to refer to an event simply as negated when they speculate or report the lack of that event, whereas if the event certainly exists, it would be explicitly reported as upregulation or downregulation.

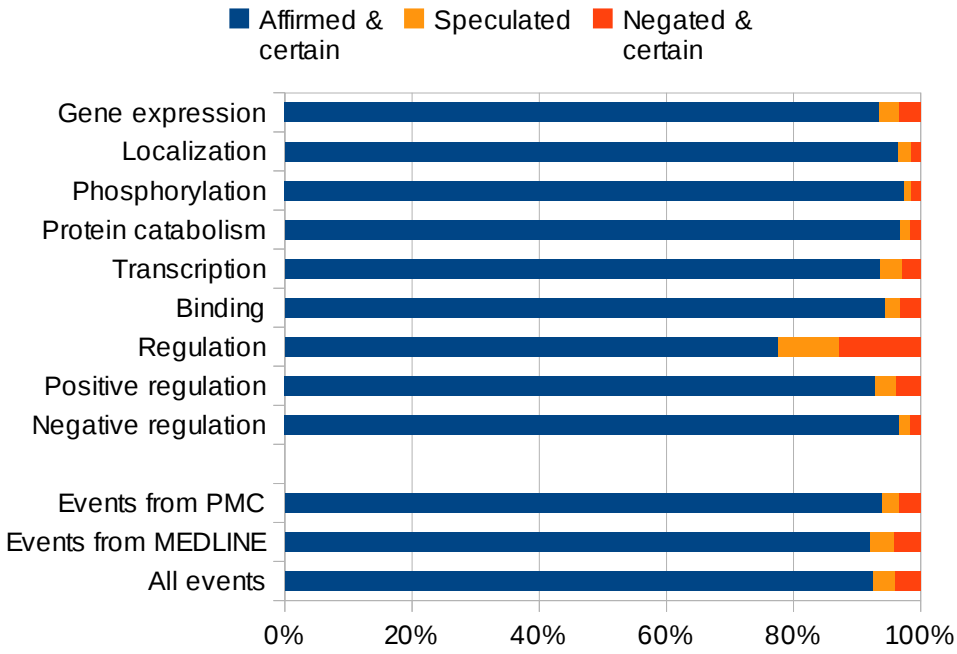


Figure 6.6: Normalised distribution of negated and speculated events for each type on MEDLINE and PMC

Figure 6.7 shows the distribution of the most common negation cues that affected some extracted event in the data set. Figure 6.8 shows the same statistics for the most common speculation cues. We observe that the trends are quite dissimilar, with negation cues "*not*" and "*no*" dominating the set of cues, whereas there are more variation amongst the speculation cues, showing a more gradual decrease in the frequency of the most frequent speculation cues.

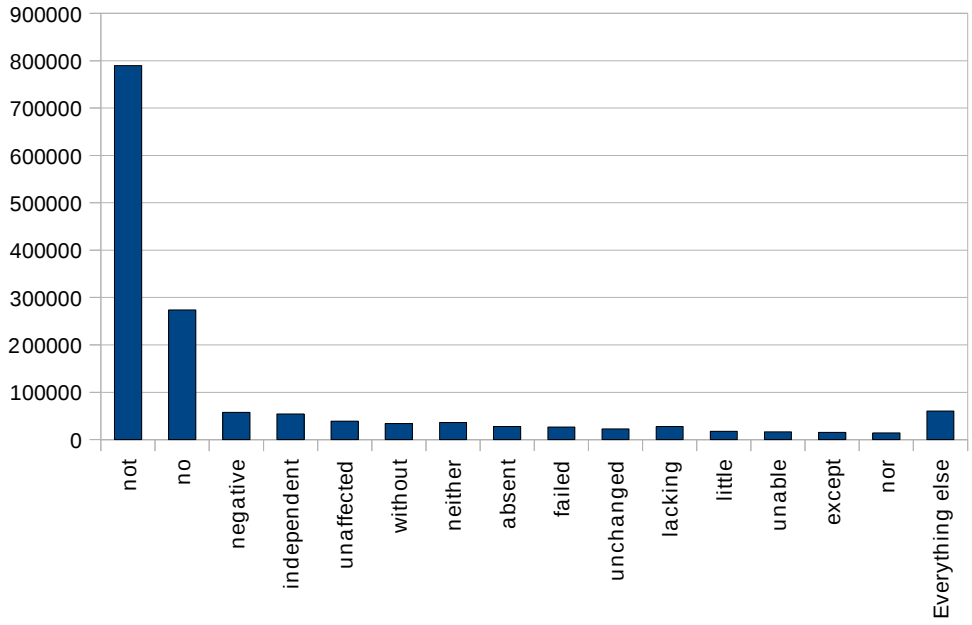


Figure 6.7: The distribution of the most common negation cues

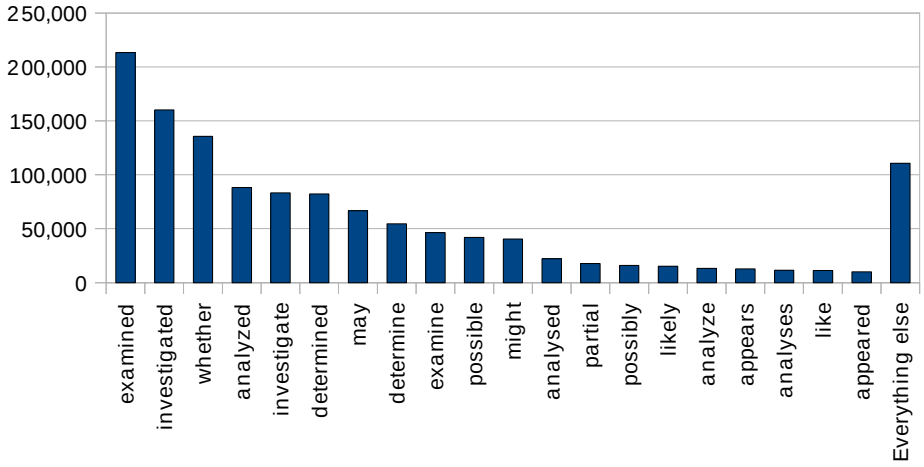
This distribution is reported on the entire extracted data from MEDLINE and PMC. Note that the cues with the same stem are grouped together, and a representative member of every stem class is shown.

These trends corresponds to those of the BioNLP'09 data, displayed in Figures 3.16 and 3.17. However, the ranks at which each cue appears differ from the BioNLP'09 data. This is due to the fact that the distributions in the BioNLP'09 data only refer to the incidence of the cues, regardless of whether these cues have affected any event. As discussed in Section 3.4.1, this shows that words with ambiguous function such as "*inhibit*" have not affected the performance of negation detection. This word is the second most frequent cue in Figure 3.16, but much further down the ranked list (not shown) in Figure 6.7.

The cues in Figures 6.7 have been stemmed. However, Figure 6.8 shows speculation cues before and after stemming. Since most high-ranked speculation cues are words with possible verb, noun, and other forms, the effects of stemming on the cues can be noticeable. However, this does not

change the order of the cues in the frequency-based ranking.

(a)



(b)

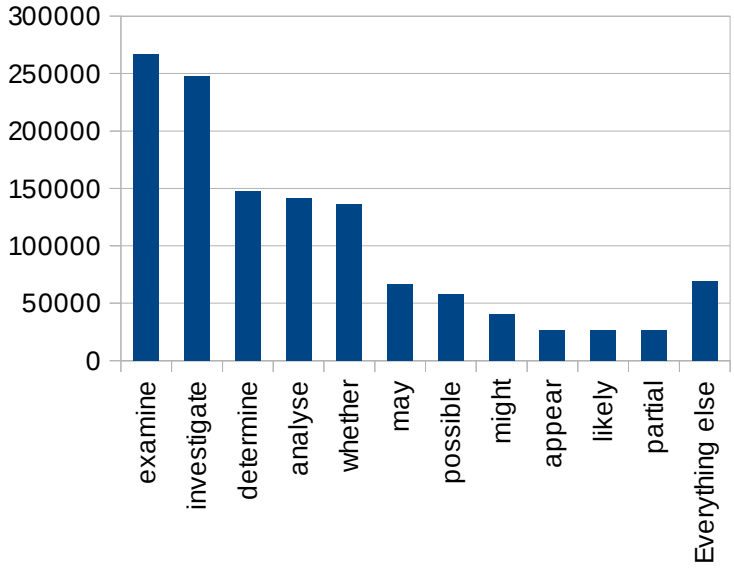


Figure 6.8: The distribution of the most common speculation cues

This distribution is reported on the entire extracted data from MEDLINE and PMC. In (a) the cues are not stemmed or otherwise normalised. In (b) the cues with the same stem are grouped together, and a representative member of every stem class is shown.

6.5 Temporal analysis

With information extracted from the entire available life sciences literature, and specially with storing the data in a denormalised format, we can perform temporal analysis on the reported claims. Figure 6.9 shows the increase in the number of the reported events as well as the number of negated and speculated events since the beginning of the recorded literature on a logarithmic scale.

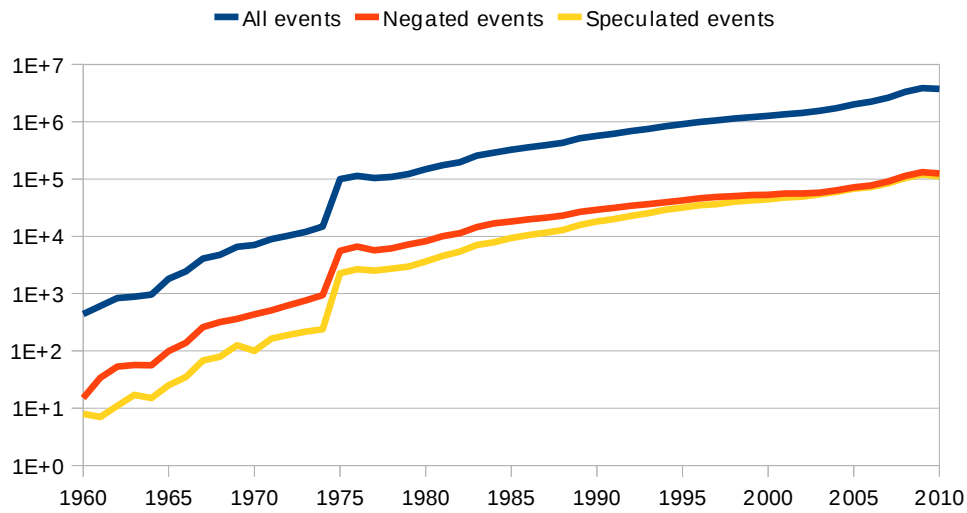


Figure 6.9: Event numbers in the literature over time

The total number of events and the total number of negated and speculated events extracted from the MEDLINE and PMC corpora over time, on the logarithmic scale.

It is interesting to compare Figure 6.9 with Figure 2.1 which shows the total growth of the literature. Notice that the sudden increase in the number of reported events around year 1975 corresponds to the rise in the number of additions to MEDLINE around the same time (see Figure 2.1).

To test the hypothesis that papers are reporting more molecular events over time, we calculate the ratio of the number of events reported per publication over time (see Figure 6.10). As can be inferred from this figure, with the growth in the volume of scientific publications, the number of molecular events reported and discussed in the literature increases even more

dramatically, suggesting that not only there are increasingly more papers published in life sciences domain, but also these papers have become much richer in content over time with regard to reporting molecular events. This can be attributed to the growth in molecular biology research in the last few decades.

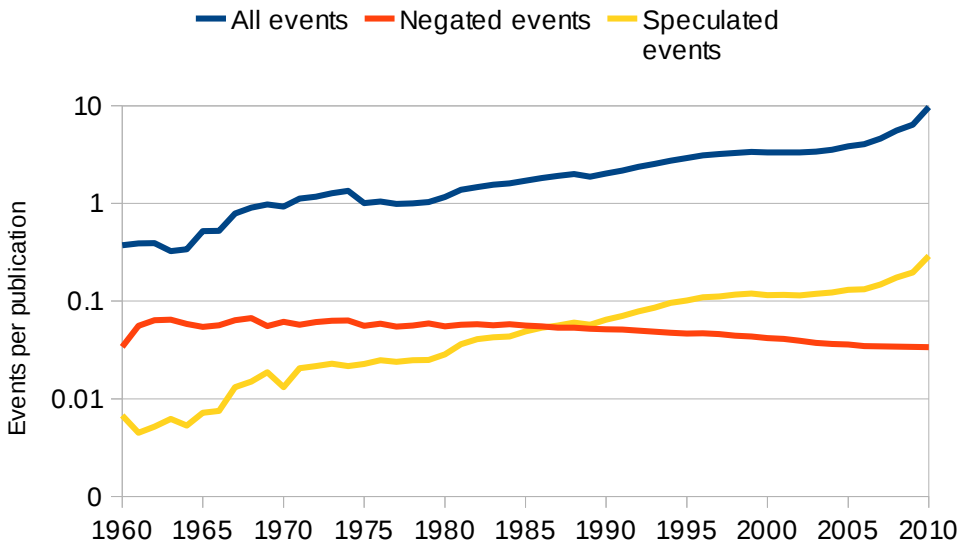


Figure 6.10: The number of events reported per publication over time

Note that the ratio is displayed on a logarithmic scale.

It is interesting to see in Figure 6.10 how the reporting of negated and speculated events has changed over time. An analysis, summarised in Figure 6.11, shows that while the ratio of the reporting of negated events are generally higher than that of the speculated events, this ratio has been decreasing over time. This suggests that scientists report their findings more speculatively than they used to, and use less assertive tone, whether strongly negative or affirmative. It also shows that scientists publish fewer negative events than they used to.

The fluctuations prior to 1975 in Figures 6.10 and 6.11 are probably due to small data sizes (see Figure 6.9).

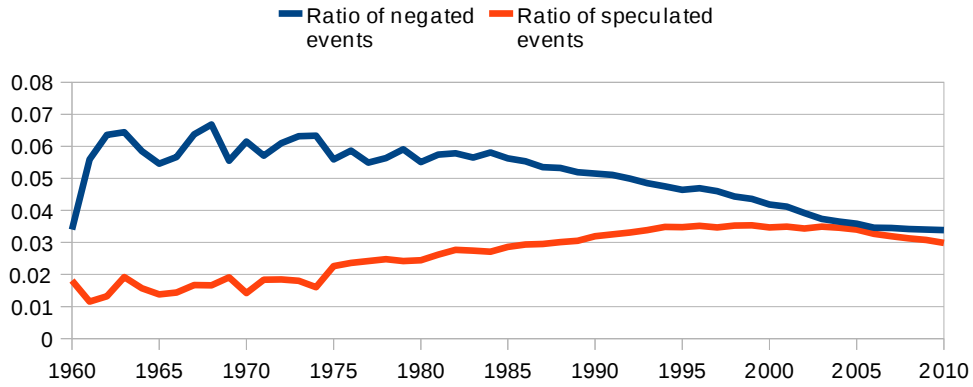


Figure 6.11: Ratio of negated and speculated events over time

The proportion of events that are negated or speculated, compared to the total number of events mined from the literature over time.

6.6 Mining conflicting statements

6.6.1 Results

Using the methods described in Section 5.4, we found 72,314 potentially conflicting pairs in the set of events extracted from the literature that had sufficient associated data which qualified them for strict conflict extraction. Table 6.18 summarises this data. It shows the number of event mentions that were sufficiently rich in context to be included in the analysis, number of distinct events included, and the number of conflicting pairs detected in the data.

The majority of conflicting pairs are gene expression events (78%) followed by localization (8%) and transcription (4%).

Type	Event mentions	Distinct events	Conflicting pairs
Gene expression	3,450,494	1,089,937	56,367
Localization	773,844	285,420	5,584
Phosphorylation	172,482	74,558	1,238
Protein catabolism	102,520	48,168	720
Transcription	260,385	149,066	3,199
Binding	420,758	288,030	2,239
Regulation	148,737	129,150	817
Positive regulation	326,800	268,792	2,047
Negative regulation	82,961	72,860	103
Total	**5,738,981**	**2,405,981**	**72,314**

Table 6.18: Summary of the events extracted in the conflict analysis.

Only a subset of the entire extracted events were included in the conflict analysis. The numbers of these events as well as the conflicting pairs extracted are shown.

Each extracted pair has a score associated with it which is a measure of how common and how confident the supporting mentions of that pair are (see section 5.3).

While these pairs should have their implicit potential in exploring biological claims, we also looked at how accurate they are from a textual perspective.

To evaluate the extracted conflicts, we manually examined a selection of conflicting pairs from the highest ranks. We selected the 10 top rank gene expression pairs, and 5 top rank pairs from each of the other 8 types for manual investigation. The numbers were chosen due to the relatively larger size of the gene expression type. Overall, a total of 50 event pairs were selected.

For each of these pairs, we selected a maximum of five supporting sentences containing affirmative supporting mention, and five supporting sentences containing a negative supporting mention. In total, a maximum of 500 supporting sentences would have been selected. However, since some of the events had fewer than 5 supporting sentences, a total of 434 sentences

where included in the analysis. The full list of the pairs and sentences that were used for evaluation can be found in Appendix C.

We manually inspected every one of these 434 sentences to initially determine whether the event extraction and contextualisation has been performed correctly, and secondly to determine whether any two of the supporting sentences are actually reporting conflicting events.

A summary of the results of this evaluation can be seen in Table 6.19. It classifies the number of sentences with correct extraction, as well as the source of error in the others.

On several occasions, the actual cause of these errors might have lied within other text mining stages, from the sentence splitter to parsers, but we have not included those errors in our analysis, as evaluating them would have been difficult. We only concentrate on the first point at which our event extraction pipeline is affected.

	Number of events	**Percentage**
Correct information extraction	206	47%
Gene name recognition or normalisation error	33	8%
Event extraction error	61	14%
Anatomical entity recognition or anatomical association error	94	22%
Negation detection error	39	9%
Speculation detection error	1	<0%
Total number of events	**434**	**100%**

Table 6.19:The summary of the conflicting pairs evaluation

Table 6.19 shows that in almost half (47%) of the cases the information extraction was performed without error. Most of the errors (41% of the errors) were due to the anatomical location association. In the sample studied, this was mostly due to the fact that the anatomical association method did not require the anatomical entity to differ from the other components of the event (trigger

and participants). Due to the high ambiguity between trigger terms, gene and protein entity names, and anatomical location names, many of them could overlap in a sentence. Since the anatomical location association method is effectively a distance-based method, it commonly associates one of the components (trigger or participant) as the anatomical location, resulting in an error. Event extraction errors happen in only 14% of the pairs, but constitute 27% of all the errors in the sentences.

Amongst the 50 event pairs examined, 32 showed some degree of conflict between the stated claims, 8 were definite errors, and 10 were undecidable by the non-biologist annotator. Projecting these results over the entire set of extracted conflicting pairs, more than 46,000 of the event pairs would show some degree of contrast.

Of course, not all the 32 positive cases were real contradictions. Most of them were conflicts due to underspecified context. Several contrasting statements were in the presence/absence of drugs or auxiliary molecules, some were due to some procedures or treatments, and others were in different populations, affecting different types of the same entity, or happening in different types of the same cell line. Alas, these contextual information has not been captured in our current setting. Nevertheless, even if they had been extracted, these sentences would still be indicating some form of conflict. A handful, however, hinted at a true contradiction. Example 6.4 is one such pair.

Example 6.4. Regulation of leptin by insulin in plasma

Affirmative supporting sentences

1. 9568685: Whether insulin acutely *regulates* plasma leptin in humans is controversial.
2. 9398728: In animal models, insulin and agents that increase intracellular cAMP have been shown to similarly *affect* plasma leptin in vivo.

Negative supporting sentences

1. 8954052: These results suggest that insulin does **not** acutely regulate plasma leptin concentrations in humans.
2. 11832440: Insulin and pentagastrin did **not** modify plasma leptin, whatever HSV status.
3. 10856891: Adrenaline, insulin and glucagon do **not** have acute effects on plasma leptin levels in sheep: development and characterisation of an ovine leptin ELISA.

All of these sentences hint at the fact that the event in question is somehow controversial and that there seems to be no consensus about it. The affirmative supporting sentence number 2 is in direct contradiction with the negative supporting sentence number 3, as one states that in animal models insulin affects plasma leptin, and the other states that it does not have such an effect in sheep.

Obviously, the procedure that is required to infer that sheep is indeed an animal would rely on some background knowledge (e.g. an ontology) and was not implemented in our method, but this example demonstrates that even without knowledge and inference integration, the results of the conflict detection can be beneficial and a good starting point for the researchers.

In Example 6.5, event extraction and contextualisation in all but the last one of the affirmative supporting sentences is correct. In the positive sentence number 5, a negated event is missed, as the authors express that certain conditions were not sufficient for the desired regulation to happen.

Similarly, the events and their context in all but the last one of the negative instances also seem to be correctly identified. In the negative sentence number 5, negation is incorrectly assigned to the regulatory event in question.

Example 6.5. Positive regulation of IgE caused by IL4 in B cells

Affirmative supporting sentences

1. 10887336: IL-4 is *important* for B-cell production of IgE, and the

human IL-4 receptor alpha chain (hIL-4Ralpha) is crucial for the binding and signal transduction of IL-4, so hIL-4Ralpha may be a candidate gene related to atopy.

2. 7722171: In contrast, terminally differentiated, IgE-producing B cells no longer express functional IL-4R because DAB389IL-4 only modestly inhibited ongoing IgE synthesis by B cells from patients with hyper-IgE states and only minimally affected IL-4-*induced* IgE synthesis in normal B cells when the toxin was added at day 7.
3. 2172384: We demonstrate here that EBV and IL-4 *induced* the synthesis of IgE by surface IgE-negative B cell precursors isolated by cell sorting.
4. 2967330: Like IL-4-containing SUP, rIL-4 also showed the ability to *induce* IgE production in B cells from both atopic and nonatopic donors.
5. 2789139: However, a combination of IL4, IL5 and IL6 (with or without IL1) at optimal concentrations could not induce IgE synthesis by purified normal B cells, indicating that cytokine-mediated signals, although essential, are not sufficient for the IL4-dependent *induction* of IgE synthesis.

Negative supporting sentences

1. 2172384: IL-4 **failed** to *induce* IgE synthesis in established EBV B cell lines and failed to induce 2.0-kb mature C epsilon transcripts but induced 1.8-kb germ-line C epsilon transcripts.
2. 2789139: Recombinant IL4 could *induce* IgE synthesis by peripheral blood mononuclear cells and autologous T/B cell mixtures, but **not** by highly purified B cells.
3. 1383379: In contrast to these observations with MNC, IL-4 **failed** to *induce* IgE and IgG4 production by purified B cells.
4. 1382870: IgE production was **not** *induced* by IL-4 in purified B cells.
5. 1904400: Similarly, DSCG did not enhance IgG2, IgG3 or IgG4

production from sIgG2-, sIgG3- or sIgG4- B cells, respectively, Interleukin-4 (IL-4) or interleukin-6 (IL-6) also *enhanced* Ig production **except** IgG4 from large activated B cells.

Overall, the first set of sentences in Example 6.5 seem to suggest that the regulation in question happens under certain conditions, whereas the second set seem to discuss some cases where it does not happen. Although some of the sentences offer more context such as an auxiliary molecule, the others do not explicitly mention such context and seem to be contrasting on the sentence level. Further document-level analysis is required in future work to determine whether these events are truly contradictory.

An interesting example (see Example 6.6) was when the event extractors made an error in virtually every aspect of the event extraction and contextualisation, but as the errors were identical in every supporting sentence, the resulting sentences still presented contrasting information, however in a completely different realm than molecular events:

Example 6.6. Transcription of Angiotensin-converting enzyme (ACE) in heart

Affirmative supporting sentences

1. 8461246: Angiotensin-converting enzyme (ACE) inhibitors are now widely *prescribed* for the treatment of hypertension and heart failure.
2. 9562936: Patterns of angiotensin-converting enzyme inhibitor *prescriptions*, educational interventions, and outcomes among hospitalized patients with heart failure.
3. 9562936: BACKGROUND: Among hospitalized patients with heart failure, we describe characteristics associated with *prescription* of angiotensin-converting enzyme (ACE) inhibitors in the doses recommended by clinical practice guidelines.

Negative supporting sentences

1. 10908091: Therefore, we recommend that physicians continue to *prescribe* ACE inhibitors for patients with heart **failure** based on the target doses used in the placebo-controlled trials and not on the "high" dose target used in ATLAS.
2. 11831455: Captopril, enalapril, and lisinopril are angiotensin-converting enzyme (ACE) inhibitors widely *prescribed* for hypertension and heart **failure**.
3. 11052861: Although most primary care physicians stated they *prescribe* ACE inhibitors in heart **failure**, this was for only 47-62% of patients, and at doses below those identified as effective in trials.
4. 9491949: BACKGROUND: Angiotensin-converting enzyme (ACE) inhibitors were *underprescribed* for patients with congestive heart **failure** (CHF) treated in the community setting in the early 1990s despite convincing evidence of benefit.

In Example 6.6, there have been a number of errors in the collective information extraction pipeline. *"Angiotensin-converting enzyme (ACE)"* is the name of a blood enzyme, which has been correctly recognised as a gene or protein name, despite the ambiguity of the acronym ACE. The word *"prescription"* referring to the medical prescription of a drug has been confused with the molecular event transcription, as it is often used as a term in expressing this type of event (again, word sense ambiguity). The disease name *"heart failure"* has caused both Negmole and anatomical association to report erroneous results, as *"failure"* has been marked as negation cue and *"heart"* has been reported as the location in which the event has occurred. Using domain-specific negation cues or semantic tokenisation for disease names could have addressed this error.

However, these errors seem to be generally consistent amongst all the supporting sentences, perhaps with the exception of *"failure"* that has not been

consistently recognised as an indicator of a negated event. Therefore, the sentences that have been grouped together still refer to the same concept, although not a molecular event, as expected. As a result, these sentences reveal a disputed subject in medicine (i.e. whether ACE inhibitor drugs should be prescribed for heart failure, and whether they are prescribed in reality).

The results of the strict contrast extraction evaluation show that in many cases, incomplete or incorrect text mining causes the resulting pair to only contain a contrast in the relaxed sense. Based on this observation, we decided not to evaluate the method on the relaxed conflicting pairs. However, the relaxed pairs are available to download for browsing on the web interface, and can be evaluated independently.

6.6.2 Discussion

Here we present a model to identify conflicting statements across the literature. Although a relatively new problem, the problem of finding contrasts, contradictions, and conflicts in the literature has been explored by many researchers in this area. Our review of the previous research in this area revealed that there has not been a consensus over the definition of the main terms and concepts related to this area. Some (e.g. BioContrasts) have explored contrasting entities, whereas others (e.g. Sanchez et al.) have considered implicit and explicit contradictions between statements.

Our conflict mining methodology addresses many limitations that the previous conflict detection systems were subject to. We have investigated conflicts at the event level. This level of investigation is finer than some of the previous sentence-level or scope-level approaches, focusing on the exact claim that is being affected. At the same time, it is not too fine-grained not to contain the smallest piece of self-contained information, as some of the previous methods that only concerned contrasting entities have done by considering contrasts on the entity level.

Although this research does not incorporate any ontological relation

among the biological events, it uses a widely accepted and commonly used set of relations as the basis of molecular events. Some previous research has included ontological information about the events they study. However, they lack enough biological justification for the ontology they introduce. Moreover, they do not use this ontological information in a computationally useful way.

By addressing the problem at the event level, and finding conflicting events across the literature, we have eliminated the need to differentiate between events that are described in the same document or even in the same sentence and those appearing in different documents.

By evaluating our method on the large-scale corpus of the entire available biomedical literature, we have demonstrated that the output contains potentially useful conflicts and potential contradictions, even despite the text mining errors occurring in the previous stages. We observed that often the errors are duplicated in both affirmative and negative cases, still leading to potentially valid examples of contrasting output.

In the conflict analysis, we focused on the events with “complete” context, discarding any event that did not have every contextual attribute extracted. Also, we treated all the regulatory events as if they are first level events, reducing higher order events to first level by assuming their indirect nested participants are direct participants. These simplifications, although presumably led to more accurate conflict results, discarded much higher quality data (as text mining accuracy decreases by the addition of contextual information.)

In many cases, this incomplete context was not even mentioned in the sentence that was being examined, and could only be inferred from the entire document. An approach that heuristically derives context from the document and assigns it to the events that miss sentence-level context could help enrich the events and include them in the conflict analysis.

Studying the sample of detected conflicts showed that, although thousands of conflicting pairs were mined in certain event types (such as *Gene*

expression), others (such as regulation events), despite having high overall frequency, showed relatively low numbers of conflicting pairs. This was due to the larger attribute set of these events which had to exist and match in order to be considered as a pair. This resulted in lower numbers of supporting sentences, to the point that even the top rank pairs did not always have a maximum of five sentences of each polarity for manual examination.

This shows that simply including regulatory events of higher orders will not necessarily lead to the discovery of more contrasts, since the more attributes need to be matched, the fewer supporting mentions there will be. Therefore, for higher order regulation events, more sophisticated contrast extraction rules are required in order to mitigate the sparseness of highly specified data.

Analysis of the results shows that in many cases, compatible instances were reported as conflicting due to incomplete context extraction, either because the context was not explicitly mentioned in the sentence, or because it was not included in the model. To explore this further, we categorise the missing context into UMLS categories. As a small case study, we consider the gene expression events of *IgM* in the anatomical location *blood.* As a prerequisite for conflict extraction, these events have to be asserted (i.e. not speculated). Table 6.20 shows affirmative and negative cases of this event. The missing context is highlighted in each case, and the UMLS category of the missing context is displayed in the left-hand column. In conclusion, more comprehensive context modelling is important and necessary for more accurate conflict detection.

Type of context (UMLS)	Example	negation
Population / Species	**Dogs** treated with UV-irradiated blood did **not** produce anti-donor PBL antibody, or IgG, IgM and C3 determined by the indirect Coombs test.	negative
	Gel diffusion analysis of sera from 73,569 **healthy volunteer** blood donors revealed apparent lack of IgA in 113 (1:650) samples, all with normal *levels* of IgG and IgM.	affirmative
	Dichotomy between immunoglobulin *synthesis* by cells in gut and blood of **patients with hypogammaglobulinaemia**.	affirmative
[absence of] Disease	Pokeweed mitogen (PWM)-induced immunoglobulin (Ig) *synthesis* by peripheral blood mononuclear cells (PBL) from 33 **patients with systemic lupus erythematosus (SLE)** was compared to that synthesized by PBL from 22 normal individuals.	affirmative
Temporal	Spontaneous immunoglobulin production by peripheral blood lymphocytes was increased **during the acute phase of illness**.	affirmative
Phenomena / Environmental	We have quantified the levels of IgG, IgA and IgM in Nigerian cord blood samples **during the dry and the wet seasons**.	affirmative
Human-caused phenomenon / Environment	The *levels* of IgG, IgA and IgM were measured in the blood serum of uranium miners in a minin-district, where after a geological disturbance **exposure to a high level of ionizing radiation** took place.	affirmative
Population / Age group	In vitro immunoglobulin *production* in mitogen-stimulated cultures of peripheral blood and bone marrow cells from **young and old adults**, and cases of benign monoclonal gammopathy. [...] Comparing the **age-related** increase in Ig synthesis in peripheral blood lymphocytes, PWM-induced Ig synthesis of bone marrow cells was only slightly increased.	affirmative
	The distribution of antibodies to HHV-6 compared with other herpesviruses in **young children**. [...] HHV-6-IgM was **not** *detected* in 235 cord blood samples.	negative

Quantitative attribute of protein	**Low molecular weight** IgM is the monomeric subunit of **pentameric** IgM and is **not** generally *found* in the blood of healthy individuals.	negative
Population (case report)	**A second blood donor (Donor B)** had a low *level* of B19 DNA but was IgM **negative**.	negative
Biological process	With all three proteases, **no** *changes* in the relative rates of MAO-A and MAO-B inactivation were observed **after disruption of the mitochondria**.	negative
Clinical drug: causes other than gene/ protein	**Pyrazidol** did **not** *affect* significantly the type B MAO in bovune liver and kidney tissues, although the latter enzyme catalyzed deamination of serotonin.	negative

Table 6.20: Examples of missing context

Examples of supporting sentences from which a gene expression event of "IgM" in the anatomical location "blood" has been extracted. Although eight affirmative and five negative supporting sentences are listed, the conflicts are not necessarily contradictions. The contexts that could have been extracted to enrich the events and emphasize the differences are highlighted. The first column shows the type of context in terms of UMLS categories.

We make the important observation that events are not static concepts, and different degrees and qualities of events are reported at different times, in different populations, and in different experimental settings. We conclude that the addition of further context could improve the accuracy and usefulness of the extracted contrasts.

To capture these nuances in the reporting of molecular events, we identify the following additional contextual information regarding the experimental settings that can be added to the extracted events.

- Temporal information
- Population type (patients, species, etc.)
- Interventions (therapeutic e.g. drugs or surgery, or diagnostic e.g. tests)
- Exposure (the main variable that is being examined or researched,

including a disease

- Co-variates (e.g. gender, environmental factors, existing treatments, age, etc.)
- Environment (experimental environment e.g. temperature, etc.)
- Other exposures

We also note that the event representational model is a simplification, and cannot represent every molecular biology phenomenon that is being reported. A more complex representation is needed if a systems biology perspective is taken (e.g. formal models, pathways, etc.)

It remains future work to combine the sum of all knowledge, not only from elsewhere in the document, but from other documents and resources to add richer context to the extracted information. Currently, we try to extract as much context as possible from the sentence to enrich the extracted information without compromising the quality of the data or producing overly specified, sparse data, and to compensate for the information which is lost due to this local perspective.

An immediate next step can be to extract context from a level beyond the sentence containing the statement, and determine the type of context that can be minimally added to the events in order to improve the likelihoods of correct conflict detection.

Another issue to have in mind when comparing automatically extracted strict and relaxed events is the likelihood of correct extraction. As more contextual information is extracted, the precision inevitably drops, due to the extra layers of sub-optimal automatic processing that is required.

The identification of conflicting findings should prove useful for biologists. For any particular interaction or process a biologist is interested in, the results of this research would enable her to quickly identify not only the papers that discuss that particular process, but also what context it has appeared in, and whether the findings support, conflict with, or merely speculate the

process. It allows her to actively discover whether the process has been reported to be observed or not. Subsequently, this would reduce the effects of confirmation bias in the scientific literature by highlighting negated or conflicting findings.

The language of scientific literature is typically formal and neutral, and authors rarely express their personal opinions explicitly. Therefore, methods used for sentiment analysis in other domains do not directly apply to the domain of scientific literature. By mining conflicting statements, more controversial areas of science and areas in which scientists have conflicting views are highlighted and brought to the attention of the research community. This can be investigated with regard to publication year, journal, species, etc.

We briefly explored a few temporal trends in the reporting of biomedical events. Using the data produced as part of this research, more sophisticated trends can be easily studied. For example, a biologist interested in a particular event can ascertain when the finding was first reported, and how the speculative and negated reports of the finding are spread throughout history compared to the affirmative findings.

Investigation of the reports of specific events and processes in aggregate would enable the readership to judge how reliable a particular finding is, how often it has been reported or refuted, in what experimental contexts it has been studied, and how much agreement there is amongst the scientific community regarding that event. Finally, it could point towards new areas for future research.

The results of this research can be integrated with scientific search engines, and in particular PubMed, to annotate the articles with useful metadata and enable the users to find relevant or conflicting information from across the literature. Finding this information through mere keyword search, if at all possible, would require a lot of time, effort, skill, and thought.

6.7 Summary

In this large-scale analysis, we extracted 36 million events and enriched them with contextual information. This was achieved by combining several entity identifiers, parsers, event recognisers and context extractors in an event extraction pipeline. By mining the resulting data, we identified interesting patterns regarding the reporting of negated and speculated findings in the scientific literature. Specifically, by mining the normalised form of this data, we identified 72 thousands conflicting pairs, potentially 2/3 of which give some notion of conflict as indicated by manual analysis of 50 event pairs.

With more accurate event extraction and more comprehensive context association (including experimental settings such as population, drug, etc.) more accurate conflict extraction can be achieved, leading to the discovery of contradictory findings in the literature.

Chapter 7
Conclusion

Biomedical literature is growing at the rate of 2,000 new documents every day. Automatic systems are needed to access the knowledge that is contained in this vast body of textual information. Many articles in the biomedical domain make claims that have been found experimentally, and/or refer to claims that have been reported previously. Specifically, molecular events have been of increasing importance to scientists since the discovery of genes and their role in biological processes.

With such large numbers of claims reported in the literature, conflicts or inconsistencies are highly probable. Finding these conflicts are of value to researchers as they can be a source for hypothesis and future investigations, as well as a method to facilitate data and knowledge consolidation.

7.1 Summary of contributions

In this thesis we sought to address the problem of mining conflicting statements from the literature at the level of molecular events. As a preliminary step for finding conflicts, it was necessary to extract biomedical events enriched with context that can be used to indicate inconsistent claims.

We have shown that automatic extraction of contextualised event information from textual data can facilitate identification of conflicting statements. Specifically, we have shown that negation is a key contextual clue that contributes towards mining conflicts. This was the main research question, i.e. whether it is possible to use negations and speculations as clues to detecting conflicting and contrasting information in the biomedical literature.

The contributions of this thesis are listed below. They show that every objective set at the beginning of this thesis has been met.

1. Proposing a general event representational model suitable for the purposes of this study, and in particular for conflict detection at the molecular event level, which is also expandable to other similar tasks.
2. Design, implementation, and evaluation of a method for event extraction (Evemole) using hybrid rule-based and machine learning methods with an overall F-score of 35% (58% for the non-regulatory events). The method uses dependency trees, and was expanded to extract various event context attributes such as anatomical locations and species.
3. Design, implementation, and evaluation of a negation extraction method (Negmole) with an overall F-score of 63% and specificity of 99%. Negmole uses the command relation as a feature, which was previously suggested by linguistics as having a link with negations. It considers every event component (trigger or participants) that is affected by negation as an indication of a negated event. Lexical, syntactic, and semantic features were used and the effects of each group of features were analysed.
4. Adaptation of the negation extraction system with minor modifications to detect information regarding statements and findings that are reported speculatively.
5. Design and producing a text mining framework, TextPipe, upon which an information extraction pipeline was built to combine several text processing modules in order to extract and contextualise molecular events from the literature.
6. The pipeline was applied to the entire accessible biomedical literature and more than 36 million events reported in the literature were extracted, 38% of which are associated with a species-specific anatomical location. The participants of these events are a subset of over 80 million extracted gene and protein entity mentions, 87% of which were normalised and linked to their species-specific database

identifiers.

7. Proposing a method that benefited from the event representational model to analyse the automatically extracted event data in order to detect potentially conflicting statements in the literature. Over 72,000 potentially conflicting pairs of events were detected, each pair having an average of five supporting event mentions per each event in the pair. This information could be used as a means for biomedical scientists to explore disputed areas of research, and find relevant contrasting or conflicting findings in the literature.
8. Providing the data regarding biomedical events, negations and speculations, and conflicting pairs from the entire MEDLINE and the open access part of PMC at www.biocontext.org both for download and for browsing through a web search interface.
9. In addition to the molecular event data, which is intended primarily for biologists and bioinformaticians, we also provide the entire intermediary data including the syntactic parse trees, genes, proteins, species, and anatomical location named entities, as well as the integration framework, TextPipe, which can be used either for new projects or to reproduce or modify the system described in this thesis.
10. Constructing gold standard corpora and evaluation data, including a manually annotated corpora derived from the GENIA corpus and the BioNLP'09 corpus, which includes protein complexes as well as genes and gene products. We manually examined the anatomical association and event inference methods on two sets of 100 sentences each. We also analysed a set of 50 extracted conflicting event pairs and manually evaluated 10 supporting sentence per each pair for the quality of text mining as well as conflicting facts.

The contributions numbered 5, 6, and 8–10 are outcomes of a larger joint project with Martin Gerner (Faculty of Life Sciences, University of

Manchester). The design and implementation of TextPipe was led by Martin, and he contributed to a greater proportion to the framework. The design and implementation of BioContext and the database was performed collaboratively, and the web interface was an expansion of one of Martin's earlier projects, GETM.

We manually examined the anatomical association and event inference methods on two sets of 100 sentences each. To evaluate the conflict detection method, a selection of the events with the highest extraction accuracy were manually examined. As a result, we found that 64% showed some degree of conflict. Moreover, we found that many of the conflicts were due to underspecification and incomplete data, rather than indicating a true contradiction.

The process of generating and integrating this large volume of data proved challenging, as the results of individual tools are not always easily reproducible on a large scale. Still, the integrated results proved to be useful when deciding the balance between precision and recall, depending on how the data will be used.

7.2 Future work and open questions

Although each component in the pipeline performed with state-of-the-art accuracy, aggregating the results meant that the quality of the extracted data was still far from perfect. Even relatively simple modules such as sentence splitting often produced errors early in the pipeline which was propagated through all the later stages, resulting in errors in the final results.

With more training data and further feature engineering, the performance of negation and speculation detection could be improved. Examples of feature that could have been helpful include the occurrence of certain adverbs and conjunctions that might be predictive of negations and speculations, such as *"however"*, *"nonetheless"*, etc.

It remains a question how expandable the method used in Negmole is if applied to other similar tasks. One such task could be the detection of manner in events, with cue words such as *"slightly"*, *"rapidly"*, *"strongly"*, *"partially"* or *"under"* in terms such as *"under-expressed"*. The generalisation of the method to relations between entities in domains other than the biomedical domain can also be investigated.

Our proposed event representation model regards events as atomic concepts. A more sophisticated representation is needed to allow more in-depth analysis, including the study of pathways, or other studies from a systems biology or ontological perspective.

Adding further context to the events, possibly from beyond the sentence boundaries would improve the quality of conflict detection and would help us move from merely detecting conflicts amongst statements towards detecting more reliable contradictions amongst claims of facts.

In this research we treated the nine event types as independent concepts. However, it is possible to imagine that they are semantically related. For example, positive regulation and negative regulation can be regarded as types of regulation, and can be considered opposites. In this sense, the two claims "A upregulates B" and "A downregulates B" could be considered conflicting by biologists, and therefore may be of interest. We have not included such semantic conflicts in our analysis of conflicting events. However, accessing this information is straightforward through the web interface as well as in the data downloads. It remains future work to systematically integrate background knowledge and semantic relations between the concepts into the conflict analysis.

Not many of the errors were due to the "double negation" phenomenon, possibly because it is not very common in scientific discourse. However, any complete negation detection method should systematically address this as well as other complicated guises of negations.

We used speculations as a way to filter out claims that are not asserted

from the analysis of conflicting statements. By capturing more nuances in speculative language, we could improve the conflict detection model to accommodate for wider shades of tone and certainty. This could prove useful for biomedical literature exploration and knowledge consolidation.

It would be useful to detect and integrate explicit conflicts appearing in the same sentence with our conflict detection method. This could include contrasting entities similar to the BioContrasts database (Kim et al. 2006) or "explicit contradictions" as studied by Sanchez (2007).

We briefly analysed temporal patterns in the reporting of molecular events. However, temporal text mining would be beneficial in finding how claims change over time. Such temporal perspective would particularly shed light on the nature of some of the conflicts that have been detected.

We used molecular events as a case study for our methods. Expanding the relation extraction, negation and speculation detection, and finally conflict detection methods to relations in the literature other than molecular events would be a future step for this research.

7.3 Conclusions

As a result of this research we have contributed towards the understanding of the phenomenon of contradictions and contrasts within the biomedical literature by defining and focusing on the conflicting events. We proposed, implemented, and evaluated a novel way for automatic mining of conflicting statements from large-scale corpora of biomedical literature.

We found that conflicts do exist among the claims made in the biomedical literature, and that text mining could provide useful support for the scientists to explore the biological knowledge by focusing on potentially conflicting statements.

Detection of negations proved to be an essential step in finding conflicts. Although not every pair of conflicting claims necessarily contain a negated

claim (e.g. semantic conflicts), the simplest and potentially most straightforward way to express conflicting claims is through negations.

Finding conflicting statements in the literature does not mean that biology as a science is self-contradictory. Rather, it emphasises the areas in which experts have expressed conflicting opinions or where experimental evidence has been found to support conflicting hypotheses. This provides support for researchers by highlighting a very specific area of literature, assisting further investigations, data exploration, and knowledge consolidation.

Appendix A

Definitions of biological event types

Definitions of biological events, partly from their respective Wikipedia articles.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA), transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA.

Transcription is the process of creating a complementary RNA copy of a sequence of DNA.

Localization refers to the location where a protein resides in a cell. The study of proteins in vivo is often concerned with the synthesis and localization of the protein within the cell. Although many intracellular proteins are synthesized in the cytoplasm and membrane-bound or secreted proteins in the endoplasmic reticulum, the specifics of how proteins are targeted to specific organelles or cellular structures is often unclear.

Phosphorylation is the addition of a phosphate (PO_4^{3-}) group to a protein or other organic molecule. Phosphorylation activates or deactivates many protein enzymes.

Protein catabolism is the breakdown of proteins into amino acids and simple derivative compounds, for transport into the cell through the plasma membrane and ultimately for the polymerisation into new proteins via the use of ribonucleic acids (RNA) and ribosomes. Protein catabolism, which is the breakdown of macromolecules, is essentially a digestion process.

Binding occurs when two or more proteins form a chemical bond together, often to carry out their biological function. Many of the most important molecular processes in the cell such as DNA replication are carried out by large molecular machines that are built from a large number of protein

components organised by their protein–protein interactions.

Negative regulation or downregulation is the process by which a cell decreases the quantity of a cellular component, such as RNA or protein, in response to an external variable. An increase of a cellular component is called **positive regulation** or upregulation. The external variable responsible for this change is known as ‘cause’. **Regulation** is a more general term referring to a change not necessarily in any of the above directions.

Appendix B

List of known trigger terms

This appendix contains the list of triggers used in the event extraction task.

Triggers to positively discriminate

Gene expression:
Transfection
coexpressed
low levels
Overproduction
Cotransfection
co-transfected
overexpression
allele-specific expression
allele specific expression
biallelic expression
positive staining
transduced
Biosynthesis
stably transfected
Paternal expression
maternal expression

Localization:
secretion
translocation
mobilization
retention
release
localized

Transcription:
transcription
inducibility
transcriptional activity
gene transcription
mRNA expression
induction
abundance of mRNAs

Binding:
recruit
form complexes
association
ligitation
form heterodimers
exist in separate hetero-complexes

Phosphorylation:
underphosphorylated form

Positive regulation:
transcriptional induction

Triggers to negatively discriminate

Transcription:
gene expression
lack
have
expression
level
expressed
absence
transcriptional induction

List of trigger stems and their distributions amongst event types

This data contains word stems and the frequency of their incidence as the trigger for each event class. The first few stems are displayed fully with the frequencies, and the rest are just listed without the frequencies for space constraints. The full list can be found in supplementary materials available at www.cs.man.ac.uk/~sarafraf/thesis-supplementary.html.

	Gene expression	Localization	Transcription	Binding	Phosphorylation	Positive regulation	Regulation	Protein catabolism	Negative regulation
+	1	0	0	0	0	0	0	0	0
-	0	0	0	1	0	0	0	0	0
:	0	0	0	1	0	0	0	0	0
aberr	0	0	0	0	0	0	1	0	0
aberr in the regul	0	0	0	0	0	0	1	0	0
abnorm	0	0	0	0	0	0	0	0	1
abnorm low level	0	0	0	0	0	0	0	0	1
abolish	0	0	0	0	0	0	0	0	18
abrog	0	0	0	0	0	0	0	0	7
absenc	1	0	1	1	0	1	1	0	5
absent	1	0	1	0	0	0	0	0	1
absent or detect at a veri low level	0	0	0	0	0	0	0	0	1
abund	0	2	2	0	0	0	0	0	0
acceler	0	0	0	0	0	2	0	0	0
accompani	0	0	0	0	0	1	1	0	0
accompani by upregul	0	0	0	0	0	1	0	0	0
accomplish	0	0	0	0	0	1	0	0	0
account	0	0	0	0	0	0	1	0	0
account for	0	0	0	0	0	1	0	0	0
accumul	0	4	0	0	0	19	0	0	0
act	0	0	0	0	0	3	3	0	0
act as a cofactor by sustain	0	0	0	0	0	1	0	0	0
act as enhanc	0	0	0	0	0	1	0	0	0
act upon to mediat	0	0	0	0	0	1	0	0	0
activ	1	0	0	0	0	234	1	0	0
activ cooper	0	0	0	0	0	1	0	0	0
activ pathway	0	0	0	0	0	1	0	0	0
advers affect	0	0	0	0	0	0	0	0	1
affect	0	0	0	0	0	1	38	0	1
affect the half-liv	0	0	0	0	0	0	1	0	0
affin	0	0	0	3	0	0	0	0	0
after	0	0	0	0	0	9	0	0	0

after incub with
aggreg
allow
alter
amplifi
analyz
antagonist
antisens
appear
as a consequ of
as a tempor consequ of
as stimulus
as the minimum sequenc
assembl of a protein complex
associ
associ to form a complex
associ with superinduct

at
at the level of transcript
at the mrna level
at the transcript level
attenu
attenu function
attribut
augment
autoinduc
autoregul
autoregulatori control
be
be a key molecular mechan
be induc
be the predomin subunit
be true
be undetect
becaus of
becom capabl
becom transcript activ
behav as an authent enhanc
bind
bind activ
bind affin
bind complex
bind genotyp
bind interact
bind mutant
bind partner
bind protein
bind site
bind specif
bind studi
bind subunit
biosynthesi
block
blunt
bnormal low level
breakdown
bright focus
by
by mean of
by stimul with
by the altern use of
can not
capabl of control
capabl of form function heterodim
caus
caus an increas
central role
chang
cis-activ
cleav
cleavag
co-activ
co-express
co-loc
co-transfect
coactiv
coengag
coexpress
coimmunoprecipit
colig
combin
comigr
compar level
compet
competit
complet degrad
complex
complex bind
complex form
complex format
compos
concentr
confer
confer direct transcript control
confer strong transcript activ
confin
conjug
consequ
constant
contain
contain a bind site
contain function promot activ
content of ap-1
continu to
contribut
control
control at transcript and post-transcript level
cooper
cooper effect
cooper in augment
cooper to mediat
coregul
costimul
cotransfect
coupl
critic
critic role
cross-link
cross-react
crosslink
crucial
culmin
cytokin product
de
declin
declin in the level
decreas
decreas number
defect
defici
defin as a respons element
degrad
degrad loss
delay
delet
demonstr
depend
deplet
depress amount
depriv
deregul
derepress
deriv
desensit
despit
destabil
detect
determin
development regul
dimer
dimeriz
diminish
direct
dispens
display dispar effect
display low level
disrupt
distinct from that regul
distribut
domin negat regul
domin negat regulatori effect
domin role
down regul
down-regul

downmodul
downregul
downstream effector
drive
due
due to
dure
dysregul
ec50 valu
effect
effect on the half-lif
efficaci
elev
elev level
elicit
elimin
enabl
enforc
engag
enhanc
enhanc and prolong
equival
essenti
essenti and suffici
essenti in the control
essenti role
establish
evolv independ
excess
exclus
exert
exert a posit effect
exert a stimulatori effect
exhibit
explan
export
expos to
express
express and transcript
express at the transcript level
express level
express mrna
facilit
fail
fail to interact
failur
find
follow
for
forc
form
form a complex
form specif dna-protein complex
form the function core
format
from
function
function activ
function in promot
function role
gene activ
gene express
gene transcript
gene transfer
general role
generat
generat by
have a promin increas
have a silenc effect
have littl if ani effect
have no effect
have similar effect
have the high bind affin
heterodim
heterodimer
heterodimer bind complex
heterodimeriz
heteromer complex format
high
high affin
high express
high level
high stabil
high-level
higher-affin site
hindranc
homodim
homodimer
hybrid signal
hyperphosphoryl
hyporespons
identifi
immobil
immun modul effect
immunoreact
impair
imped
implic
import
import factor
import for regul
import or essenti
in
in concert to regul
in favor of
in respons
in respons to
in the amount
in the case of
in the presenc of
in transcript activ
inactiv
includ
increas
increas level
increas number
increas stabil
increas the proport
independ
indispens for the activ
induc
induc a down-regul
induc an enhanc
induc complex
induc hyper
induc the format
induct
induct be obtain
ineffect
influenc
influenc the level
inhibit
inhibit effect
inhibitor
inhibitori
inhibitori capac
inhibitori effect
inhibitori role
initi
insuffici
intact
intens
interact
interact receptor-ligand pair
interact to exert differ effect
interfer
intermediari link
into
introduc
invers
involv

involv in regul
involv in the regul
it specif receptor
joint requir
key enzym
key role
lack
larg amount
lead
lead to an enhanc
lead to synergist enhanc
lead to the prevent
lead to their acceler
less
less import
level
level peak
level stay low
liber
ligand
ligat
limit
link
linkag
local
localiz
lose
loss
low
low amount
low level
lower
lower-affin site
maintain
mainten
make
mask
maxim
maxim express
measur
mechan
mediat
mediat a reduct
migrat
mimic the effect of the lectin
mobil
modif
modifi
modul
more
mrna
mrna accumul
mrna express
mrna level
mrna synthesi
mrna transcript
multim
mutual exclus
necessari
necessari and suffici
necessari but not suffici to mediat
need to revert
negat
negat autoregul
negat effect
negat regul
negat regulatori
negat regulatori role
negat transcript effect
negat transcript regul
neutral
non-express
nonexpress
nonproduc
nonrespons
nonsecret
normal level
not
not affect
not lead to detect activ
not requir
not transcrib
number
observ
occup
occupi
occur
oligomer
oper
oppos
opposit
optim induct
over
over-express
overcom
overexpress
pair
particip
particip in the regul
pathway synerg
pattern
perpetu
persist
pg490
phosphoform
phosphoryl
phosphoryl form
phosphorylation-defect form
physic associ
physic interact
play a critic role
play a key role in defin
play a major role
play a role
poor
posit
posit autoregul
posit control
posit induct
posit regul
posit regulatori
posit regulatori role
posit role
post-transcript
posttranscript effect
posttranscript regul
potent
potenti
potenti role
preced
preform
presenc
present
preserv
prevent
primari sourc
process
produc
product
prolong the stabil
promin
promot
promot activ
promot function
protect effect
protein level
protein secret
proteolysi
proteolyt
proteolyt degrad
provid
provid costimulatori signal
provid high level

reach a maximum
reach a peak
react
reactiv
read through
receptor
recogn
recognit
recov
recruit
reduc
reduc level
reduc the level
reduct
regain
regardless of
regul
regulatori
releas
relief
remain constant
remain elev
remov
replac
replenish
repress
repress effect
repressor
requir
requir to induc
reservoir
resist
respond
respons
respons element
respons for enhanc
restor
restrict
result
result in a reduct
result in abnorm
result in an increas
result in increas
result in peak level
result in up-regul
result in veri limit
resynthes
resynthesi
retarget
return
return to baselin level
revers
rise
rna
rna transcript
role
screen
secret
select
sensit
serv as mediat
serv to target
show
show an earli peak and more activ
shuttl
signal
signal role
signific role
similar effect
simultan engag
slow migrat form
sourc
specif
specifi
spontan express
squelch
stabil
stabiliz
stimul
stimul the activ
stimulus
strong
subject
subsequ to
suffici
suffici for bind
suffici for the up-regul
suffici to respond
super
super-induc
superinduc
superinduct
support
suppress
suscept
sustain
switch
synerg
synergist action
synergist activ
synergist effect
synergist induc
synergist induct
synergist regul
synergist transactiv
synthes
synthesi
target
the cdna hybrid
through
to diminish
to induc
to inhibit
tran activ
tranfect
trans-activ
transactiv
transactiv pathway
transcrib
transcript
transcript activ
transcript blockad
transcript complex
transcript control
transcript induct
transcript inhibit
transcript inhibitor
transcript initi
transcript level
transcript mediat
transcript rate
transcript regul
transcript repress
transcript repressor
transcript stimul
transcript up-regul
transcriptionally-act
transfect
transfer
transfer of tyrosin phosphoryl group
transloc
treat
trigger
trigger signal cascad
trigger the activ
tripl autoregulatori loop
turn
ubiquitin-proteasom pathway
unabl to induc
unaffect
unalt
unchang
under

under the control
under transcript control
undergo
underli
underli the abil
underphosphoryl
undetect
up-regul
upon
upregul
upstream
use
util
vari
via
when
with
without
without stimul
yield constitut repressor that escap

Appendix C

Sentences selected for conflict evaluation

1 Gene expression

Affirmative:

- 3138231: We have described a human tumor T cell line, IARC 301, which constitutively *expresses* high affinity interleukin 2 (IL2) receptors, and showed that after binding to its receptors, IL2 is endocytosed and degraded.
- 1975810: To produce a molecule that will kill activated T cells as well as lymphomas and leukemias *expressing* interleukin 2 (IL2) receptors, we have created a recombinant chimeric protein in which IL2 is attached in peptide linkage to a truncated mutant form of Pseudomonas exotoxin (PE) (Lorberboum-Galski, H., FitzGerald, D.J.P., Chandhary, V.K., Adhya, S., and Pastan, I. (1988) Proc.
- 1975810: Our results indicate that IL2-PE664Glu should be evaluated as an immunosuppressive agent for the treatment of human immune disorders in which activated T cells *expressing* the IL2 receptor are prominent.
- 6255556: TCGF *production* for cloning and growth of functional human T lymphocytes.
- 6255556: In an effort to increase the potency of T cell growth factor (TCGF), several variables were examined for their effects on the *production* of TCGF.

Negative:

- 3491139: The sarcoid lung T cells, however, did **not** *express* the IL 2 gene constitutively; when placed in culture with no stimulation and evaluated after 24 hr, they demonstrated down regulation of the amounts of IL 2 mRNA transcripts, despite the fact that they were capable of re-expressing the IL 2 gene and releasing more IL 2 in response to added activation signals.
- 2990687: Resting T-cells do **not** *express* IL-2 receptors, but receptors are rapidly expressed on T-cells following interaction of antigens, mitogens, or monoclonal antibodies with the antigen-specific T-cell receptor complex.
- 2990687: Normal resting T-cells and most leukemic T-cell populations do **not** *express* IL-2 receptors; however, the leukemic cells of the 11 patients examined who had human T-cell lymphotropic virus-associated adult T-cell leukemia expressed the Tac antigen.
- 3918105: In the resting state, the T3-positive, human T cell line Jurkat does **not** *synthesize* detectable amounts of either interleukin 2 (IL 2) or gamma-interferon (IFN-gamma).
- 3918105: In the resting state, the T3-positive, human T cell line Jurkat does **not** *synthesize* detectable amounts of either interleukin 2 (IL 2) or gamma-interferon (IFN-gamma).

2 Gene expression

Affirmative:

- 2311095: *Expression* of desmin gene in skeletal and smooth muscle by in situ hybridization using a human desmin gene probe.
- 2311095: We have used a probe encoding for the human desmin gene to study the *expression* of the desmin gene in skeletal and smooth muscle by in situ hybridization.
- 3300387: The intermediate filament typing of skeletal and smooth muscle tumors has shown that these neoplasms are characterized by the combined *expression* of desmin and vimentin intermediate filaments.
- 1697612: From these results, we suggest that the large tumor cell of IDF is a myofibroblast and may originate from or differentiate toward vascular smooth muscle cells, because only this type of smooth muscle can *coexpress* desmin, vimentin and cytokeratin.
- 2024709: Using light and electron microscopic immunolocalization techniques, a series of 207 normal and pathologic human liver specimens were evaluated for the *expression* of alpha smooth muscle (SM) actin and desmin in this and other nonparenchymal cell types.

  Negative:
- PMC1878495: The neoplastic cells were positive for vimentin, alpha smooth muscle actin, osteonectin, CD99, and S100 in the chondroblastic portion, but negative for cytokeratin, epithelial membrane antigen, desmin, myogenin, CD34, and c-kit.
- PMC2975002: Mesenchymal markers including S100 protein, a-smooth muscle actin, CD34, myoglobin and desmin were absent in the present tumor, indicating that it was not a sarcoma but a carcinoma and that these antigens did not emerge during the spindle cell transformation of the adenocarcinoma of the present case.
- PMC2895884: Markers for epithelial membrane antigen (EMA), CD34, desmin, smooth muscle actin and keratin cocktail were negative.
- 1317998: The tumor cells were immunoreactive for alpha-smooth muscle actin, vimentin, laminin, and type IV collagen, but did **not** *express* desmin.
- 8291651: Others were negative: S-100, MAC387 (L1 antigen), LeuM1 (CD15), desmin, smooth muscle-specific actin, and QBEND10 (CD34).

3 Gene expression

Affirmative:

- 301152: The necessity for T cell help for human tonsil B cell responses to pokeweed mitogens: induction of DNA synthesis, immunoglobulin, and specific antibody production with a T cell helper factor *produced* with pokeweed mitogen.
- 68013: One line (JM) *expressed* T cell characteristics and complement receptors.
- 381769: T-helper cells *produce* a T-cell replacing Factor (TRF) upon mitogenic or antigenic stimulation.
- 89129: The THI and the deficient *production* of T cell--helper factor resolved after the age of 20 to 24 mo.
- 315319: Requirements for the mitogen-dependent *production* of T cell growth factors.

  Negative:
- 6461917: Hydrocortisone abrogates proliferation of T cells in autologous mixed lymphocyte reaction by rendering the interleukin-2 Producer T cells

unresponsive to interleukin-1 and **unable** to *synthesize* the T-cell growth factor.

- 6220110: AMLR killer activity was virtually eliminated by treatment with C' and 9.6 or 4F2, but the cytotoxic cells did **not** *express* NK-specific antigens, OKM1 and Leu-7, nor cytolytic T lymphocyte-specific antigens, 9.3 and OKT8.
- 6607843: We found that T-cell-associated antigens were **not** *expressed* on Tdt+ bone marrow cells and that T cells in bone marrow have a phenotype similar if not identical to peripheral-blood T cells.
- 6232854: Lymphomas from 28 patients (31%) did **not** *express* immunoglobulin or T-cell antigens but commonly expressed the B-lineage antigen B1; and the remaining 9 cases generally expressed Ia antigens, common ALL antigens, or both.
- 3081578: A T cell surface membrane-associated glycoprotein, Tp40 (40,000 mol wt), also designated as CD-7, was **not** *expressed* by the T cells of a patient with severe combined immunodeficiency.

4 Gene expression

Affirmative:

- 8047166: Cells expressing the expanded V beta s predominantly *expressed* the CD8 T-cell differentiation antigen and mediated HIV-specific cytotoxicity.
- 12660941: In this in vitro study, in a human autologous CD8(+) T cell/dendritic cell (DC) coculture system, thalidomide and a potent thalidomide analogue were shown to enhance virus-specific CD8(+) T cell cytokine *production* and cytotoxic activity.
- 15371918: CD4+CD8+ human small intestinal T cells are decreased in coeliac patients, with CD8 *expression* downregulated on intra-epithelial T cells in the active disease.
- 15371918: Cell yield and viability were assessed and flow cytometric analysis was used to examine CD4CD8 T cells and to quantify CD8 *expression*.
- 15371918: Levels of CD8 *expression* by CD4CD8 T cells in the epithelial layer were decreased significantly in patients with active coeliac disease.

Negative:

- 1335323: Both the CD4 and CD8 T cell subsets, and a hitherto undefined T lineage **lacking** CD4/CD8 *expression* have been involved.
- 2978114: Whereas the majority of T cells use alpha and beta chains to form their T-cell receptor, a small minority of T cells, which do **not** *express* the CD4 or CD8 surface markers, use other chains termed gamma and delta to form their receptor.
- 2467704: The T-cell surface antigens CD5 and/or CD2 and focal acid phosphatase were additional markers of this subgroup traditionally called pre-T ALL, whereas thymocyte antigen CD1 as well as CD4 and CD8 antigens were **not** *expressed*.
- 8750571: CD4 and CD8, gamma/delta TCR bearing T cells and CD45R0 on CD4+ T cells as a marker for memory cells, on TL no differences could be *detected* between patients with or **without** anti-TPO.
- 8977296: Here, we analyzed both early (intracellular Ca2+ mobilization), and late (interleukin-2 production) signal transduction events induced by a cognate peptide or a corresponding altered peptide ligand using T cell

hybridomas *expressing* or **not** the CD8 alpha and beta chains.

5 Gene expression

Affirmative:

- 2552810: Cells *expressing* the CD4 and CD8 antigens were both increased in number, with the former accounting for approximately two-thirds of the T lymphocytes.
- 19943952: RESULTS: Here we report that, in a virus-free mouse model, conditional ablation of activated CD4(+) T cells, the targets of immunodeficiency viruses, accelerates their turnover and *produces* CD4(+) T cell immune deficiency.
- 17892128: Flow cytometry was used to detect the *expression* of CD4 and CD25 molecules of the T cells which came from the tumor-bearing mice.
- 10929051: In addition to the recognized phenotypic distinctions of resident vaginal T lymphocytes, we recently provided evidence by fluorescence-activated cell sorter (FACS) that murine vaginal CD4+ T lymphocytes, are differentially recognized by two epitope-distinct anti-CD4 antibodies, suggesting that the CD4 protein on vaginal CD4+ cells is atypically *expressed*.
- 10553102: We have recently reported that all "conventional Ag" reactive CD4+ T cells are contained within the subpopulation *expressing* high levels of the CD4 molecule, termed CD4high.

Negative:

- 7876540: The IEL compartment from SCID mice injected with highly purified CD4+/CD45RBhigh T cells or CD4+ T cells contained significant numbers of T cells that *expressed* both CD4 and CD8 alpha, but **not** CD8 beta.
- 9317137: Here we show that a CD4 minigene comprising a combination of these elements is specifically *expressed* in mature CD4+ T cells of transgenic mice, but **not** in CD4+CD8+ double positive thymocytes.
- 8350060: A small subset of T cells of mature phenotype *express* the alpha/beta T cell receptor, but **not** CD4 and CD8 coreceptors (alpha/beta double-negative [DN] cells).
- PMC2585832: The earliest T cells *express* **neither** CD4 nor CD8 and are known as double-negative (DN) cells.
- PMC1636060: The gating on CD3+ T cells removes the monocytes (*expressing* CD4 but **not** CD3 on their cell surface) from the gate.

6 Gene expression

Affirmative:

- 3878829: Such PtLN cells exhibited augmented proliferative responses to T cell mitogens and exogenous interleukin 2 (IL 2) and showed a great ability to *produce* IL2, which suggests an increase in mature T cells in the PtLN.
- 1717798: Lymphocyte proliferation and IL2 *production* in response to a T cell mitogen are greatly diminished during the whole life of the animals, on the contrary B cell proliferation in the presence of lipopolysaccharide is not modified.
- 2439327: *Expression* of the gene for the T-cell growth hormone, interleukin 2 (IL2), is subject to at least two types of control.
- 2439327: *Expression* of the gene for the T-cell growth hormone,

interleukin 2 (IL2), is subject to at least two types of control.

- 3918306: By assaying the supernatant fluid, IL-2 cDNA clones that *express* T-cell growth-factor (TCGF) activity were identified.

Negative:

- 2568931: The results obtained show that mature T cell growth in vivo is not accompanied by expression of high-affinity interleukin 2 (IL2) receptor in the majority of activated cells, is not abrogated by in vivo administration of anti-IL2 receptor antibodies or enhanced by the in vivo injection of recombinant IL2, and that in vivo growing T cells do **not** *produce* detectable amounts of IL2, as evaluated functionally by limiting dilution assays or the presence of IL2 mRNA, detected by Northern blots or in situ hybridization.
- 2412742: Autoimmune mice bearing the single autosomal recessive gene 1pr are **unable** to *produce* the T cell growth factor, interleukin-2 (IL-2).
- 2788724: The antigen-reactive T cell line *produces* **neither** IL-2 nor inhibiting factors such as neutralizing factors against preformed IL-2 activity and IL-2 production inhibiting factors, thus the cells are exclusive IL-2 acceptor.
- 3293056: Spleen cell populations enriched for T lymphocytes and depleted of tumor cells by density gradient centrifugation in Ficoll were **unable** to *produce* IL-2.
- PMC2526191: The conclusion that Blimp-1 represses IL-2 transcription is supported by several observations: (a) Blimp-1â€“expressing cells do **not** *express* IL-2 protein at detectable levels; (b) Blimp-1 mRNA induction correlates with IL-2 mRNA down-regulation; (c) IL-2 protein and steady-state mRNA are elevated in Blimp-1â€“deficient CD4+ T cells; and (d) endogenous Blimp-1 specifically binds to a regulatory region in the Il2 gene in activated primary CD4+ T cells.

7 Gene expression

Affirmative:

- 9767468: Thus, long-term T-lymphocyte responses and the *production* of IFN-gamma can be generated using a single inoculation of PPD-pulsed DC.
- 3142782: In the IFN-gamma-producing cells, the expression of the major histocompatibility complex class I genes was augmented; this augmentation was remarkable in T cell lines tested in this work, regardless of their poor IFN-gamma *production*.
- 12900519: We then used bone marrow chimeras and fetal liver reconstitutions to create mice with an intact gammadelta T cell repertoire but one that was specifically deficient in the capacity to *produce* IFN-gamma.
- 12900519: Moreover, genetic deficiency of gammadelta T cells resulted in impaired IFN-gamma *production* by tumor antigen-triggered alphabeta T cell upon immunization with tumor lysate.
- 11292707: IFN-gamma mRNA was also *detected* in brains of infected SCID mice depleted of NK cells by treatment with anti-asialo GM1 antibody, and such animals did not develop TE after receiving immune T cells.

Negative:

- 15893298: Nylon wool-purified "T cells", however, **failed** to *produce* IFN-

gamma in response to Con A in vitro, while the production was restored by the addition of neutrophils.

- 2512260: The C57Bl/6-derived T cell line, L12-R4, *produced* murine interferon-gamma (IFN gamma) in response to mitogenic stimulation by phorbol myristate acetate (PMA) or concanavalin A (Con A), but **not** by staphylococcal enterotoxin A (SEA).
- 12414157: Moreover, by 3 weeks post-infection splenocytes from the susceptible BALB/c mice **failed** to *produce* IFN-gamma and relied on TNF-alpha as well as CD8 T cells to control infection until the end of the plateau phase around 6 weeks post-infection when IFN-gamma production resumed and clearance began.
- 8806814: Taken together, it may be concluded that **NO** down-regulates IFN-gamma *production* mainly by inhibiting T-cell proliferation.
- 2969818: Although 21 out of 503 (4%) CD4+ T cell clones produced IL 4, but **not** IFN-gamma or IL 2, and 208 (41%) *produced* IL 2 and/or IFN-gamma, but not IL 4, a total number of 185 (37%) CD4+ clones showed the ability to produce IL 4 plus IL 2 and/or IFN-gamma.

8 Gene expression

Affirmative:

- 7699322: As previously documented in mature CD8+ alpha/beta T cells and natural killer cells, HHV-6 infection induced gamma/delta T lymphocytes to *express* de novo CD4 messenger RNA and protein, as detected by reverse transcriptase-polymerase chain reaction and fluorocytometry, respectively.
- 7699322: Whereas purified CD4- gamma/delta T cell populations were per se refractory to HIV infection, they became susceptible to productive infection by HIV-1, strain IIIB, after induction of CD4 *expression* by HHV-6.
- 12663814: With a novel intervention designed for increased potency, we have more accurately deduced the half-lives of virus-*producing* CD4(+) T cells, 0.7 day, and the generation time of HIV-1 in vivo, approximately 2 days, confirming the dynamic nature of HIV-1 replication.
- 12204972: CD4(+) CD25(+) T-cell *production* in healthy humans and in patients with thymic hypoplasia.
- 11762998: Although CD4 T cell *production* is impaired in patients infected with HIV, there is now increasing evidence that the primary basis of T cell depletion is accelerated apoptosis of CD4 and CD8 T cells.

Negative:

- 10950773: In contrast, HIV-1-specific proliferative responses were **absent** in most individuals with progressive HIV-1 infection, even though interferon-gamma-*producing* HIV-1-specific CD4(+) T cells were detectable by flow cytometry.
- 10692049: Upon administration of these mAbs to mice that *express* a human CD4 transgene, but **not** mouse CD4 (HuCD4/Tg mice), clenoliximab and keliximab exhibited similar kinetics of binding to CD4, and induced the same degree of CD4 modulation from the cell surface, although only keliximab mediated CD4+ T-cell depletion.
- PMC1828063: This is **not** unexpected, as monocytes *express* surface CD4, but clearly suggests that for optimal purities CD4+ T cells should be isolated from a monocyte-depleted sample.
- 8555467: Notably, the nuclei of reactive CD3+/CD4+ T cells nearby to and

rosetting around L&H cells in NLPHD were also strongly BCL-6+, but **lacked** CD40 ligand (CD40L) *expression*.
- 9209651: Notably, the nuclei of reactive CD3+/ CD4+ T cells near to and rosetting around L&H cells in NLPHD were also strongly bcl-6+, but **lacked** CD40 ligand (CD40L) *expression*.

9 Gene expression

Affirmative:
- 6403360: The cell *producing* IFN-gamma, both spontaneously and after UCHT1 antibody stimulation, is an OKT3+,4+,8-,HLA-DR-T lymphocyte as determined at the single cell level.
- 11548832: BACKGROUND: Gamma interferon (IFN-gamma) is *produced* by activated natural killer and T cells under pathologic circumstances.
- 11548832: BACKGROUND: Gamma interferon (IFN-gamma) is *produced* by activated natural killer and T cells under pathologic circumstances.
- 11174142: Specific cell-mediated immune responses, determined by T cell stimulation and IFN-gamma *production*, were evoked following stimulation with trichophytic antigens.
- 9393632: It was found that IFN-gamma was *produced* in response to both PPD and Leishmania stimulant by T cells in the cultures.

Negative:
- 3137203: These results indicate that the myelomonocytic HBL-38 cells, **not** a T-cell line, can also *produce* IFN-gamma identical to the product of normal human PBL.
- 9267102: Because IFN gamma is *synthesized* by activated T cells, but **not** by keratinocytes, these results suggest that Fas may only be effective in apoptosis occurring in T-cell mediated inflammatory skin diseases.
- 6434688: A parallel production of gamma interferon (IFN-gamma) is induced by recombinant IL-2 (rIL-2), and NK cells appear to be the major producer cells, whereas T cells are **unable** to *produce* IFN-gamma under these experimental conditions.
- 9420133: In vitro studies using the technique of cloning lymphocytes demonstrated that a great proportion of T-cell clones derived from bronchial mucosa of subjects with TDI-induced asthma *produced* IL-5 and interferon-gamma, but **not** IL-4, upon in vitro stimulation.
- 10352314: The proliferating T cells *produced* IFN-gamma but **not** IL-4, indicating a bias toward a type 1 immune response.

10 Gene expression

Affirmative:
- PMC2872609: Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 *expression* Combining experiments on primary T cells and mathematical modeling, we characterized the stochastic expression of the interleukin-4 cytokine gene in its physiologic context, showing that a two-step model of transcriptional regulation acting on chromatin rearrangement and RNA polymerase recruitment accounts for the level, kinetics, and population variability of expression.A rate-limiting step upstream of transcription initiation, but occurring at the level of an individual allele, controls whether the interleukin-4 gene is expressed during antigenic stimulation, suggesting that the observed stochasticity of expression is linked to the dynamics of chromatin rearrangement.The

computational analysis predicts that the probability to re-express an interleukin-4 gene that has been expressed once is transiently increased.

- PMC2872609: Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression Combining experiments on primary T cells and mathematical modeling, we characterized the stochastic *expression* of the interleukin-4 cytokine gene in its physiologic context, showing that a two-step model of transcriptional regulation acting on chromatin rearrangement and RNA polymerase recruitment accounts for the level, kinetics, and population variability of expression.A rate-limiting step upstream of transcription initiation, but occurring at the level of an individual allele, controls whether the interleukin-4 gene is expressed during antigenic stimulation, suggesting that the observed stochasticity of expression is linked to the dynamics of chromatin rearrangement.The computational analysis predicts that the probability to re-express an interleukin-4 gene that has been expressed once is transiently increased.
- PMC2872609: Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression Combining experiments on primary T cells and mathematical modeling, we characterized the stochastic expression of the interleukin-4 cytokine gene in its physiologic context, showing that a two-step model of transcriptional regulation acting on chromatin rearrangement and RNA polymerase recruitment accounts for the level, kinetics, and population variability of expression.A rate-limiting step upstream of transcription initiation, but occurring at the level of an individual allele, controls whether the interleukin-4 gene is *expressed* during antigenic stimulation, suggesting that the observed stochasticity of expression is linked to the dynamics of chromatin rearrangement.The computational analysis predicts that the probability to re-express an interleukin-4 gene that has been expressed once is transiently increased.
- PMC2872609: Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression Combining experiments on primary T cells and mathematical modeling, we characterized the stochastic expression of the interleukin-4 cytokine gene in its physiologic context, showing that a two-step model of transcriptional regulation acting on chromatin rearrangement and RNA polymerase recruitment accounts for the level, kinetics, and population variability of expression.A rate-limiting step upstream of transcription initiation, but occurring at the level of an individual allele, controls whether the interleukin-4 gene is expressed during antigenic stimulation, suggesting that the observed stochasticity of expression is linked to the dynamics of chromatin rearrangement.The computational analysis predicts that the probability to re-*express* an interleukin-4 gene that has been expressed once is transiently increased.
- PMC2872609: Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression Combining experiments on primary T cells and mathematical modeling, we characterized the stochastic expression of the interleukin-4 cytokine gene in its physiologic context, showing that a two-step model of transcriptional regulation acting on chromatin rearrangement and RNA polymerase recruitment accounts for the level, kinetics, and population variability of expression.A rate-limiting step upstream of transcription initiation, but occurring at the level of an individual allele, controls whether the interleukin-4 gene is expressed during antigenic stimulation, suggesting that the observed stochasticity of

expression is linked to the dynamics of chromatin rearrangement.The computational analysis predicts that the probability to re-express an interleukin-4 gene that has been *expressed* once is transiently increased.

Negative:

- 10511099: IL-4 was **not** *detected* in the serum of CD4+ T-cells-depleted mice.
- 11777945: This IL-4 was **not** *produced* by T cells, but soluble factors secreted by the recall Ag-activated T cells, including IL-3, triggered cells of the innate immune system, primarily mast cells, to secrete IL-4.
- 10424447: Purified CD8+ T cells from uninfected or flu infected IL-4 transgenic (tg) animals *produced* **no** detectable IL-4 or IL-5 after in vitro stimulation on anti-CD3 coated plates.
- 9120259: However, splenic T cells from SM/J and B10.SM (H-2v, neu-1a) strain mice, deficient in neu-1 sialidase activity, **failed** to *produce* IL-4 but produced normal levels of IL-2 following activation.
- 18491378: IL-4 was **not** *detected* in any sample, but IL-13 levels were also comparable between normal and T-cell-deficient mice indicating Th2-polarized T-cells are not the sole source of this cytokine.

1 Localization

Affirmative:

- 11600182: Cell mixing experiments suggested that the DES-induced increase in IFN-gamma *secretion* is due to hormonal effects on T cells but not on APC.
- 17275143: IFN-gamma *secreting* T cells specific for survivin was found after temozolomide (TMZ) treatment in C57BL/6 mice intracranial (i.c.) inoculated with GL26 cells.
- 19280632: Immunogenicity studies in mice have shown that antigen-specific antibody titers and T-cell proliferative responses, as well as the *secretion* of IFN-gamma, were significantly enhanced for ovalbumin after formulation with PEG-b-PLACL-based emulsions.
- 8964609: Allergen-activated draining lymph node cells (LNC) isolated from mice exposed topically to the contact allergen oxazolone mount vigorous proliferative responses and *secrete* substantial amounts of interferon-gamma (IFN-gamma) when cultured with the T lymphocyte mitogen concanavalin A (con A).
- 8964609: Allergen-activated draining lymph node cells (LNC) isolated from mice exposed topically to the contact allergen oxazolone mount vigorous proliferative responses and *secrete* substantial amounts of interferon-gamma (IFN-gamma) when cultured with the T lymphocyte mitogen concanavalin A (con A).

Negative:

- 8932272: We demonstrate that the transplantation of polarized type 2 murine T cells (i.e., cells *secreting* IL-4 but **not** IFN-gamma) together with T-cell-depleted bone marrow results in a significant increase in survival (P<0.001) after bone marrow transplantation across minor histocompatibility barriers (B10.BR-->CBA/J).
- 9767466: Analysis by reverse transcription-polymerase chain reaction revealed that, in contrast to mouse, rat NK T cells *secrete* exclusively IFN-gamma and **not** IL-4 after anti-CD3 stimulation, and use a wider TCR-Vbeta repertoire, suggesting that rat NK T cells are not essential for the

development of Th2-type CD4+ T-cell responses.

- 10527383: In this study we report that a single, subcutaneous injection of the peptide emulsified in IFA gave rise to the development of male-specific CD8+ T cells which displayed H-Y-specific proliferative response in vitro and showed a Tc1-type pattern of cytokine production (i.e. they *secreted* IFN-gamma and IL-2, but **not** IL-4 and IL-10).
- 10678966: These primed IFN-gamma-*secreting* **LACK**-reactive T cells were not detected ex vivo after day 7 of immunization but could be recruited and detected 15 days later in the draining lymph node after an L. major footpad challenge.
- 10692034: Splenic T cells isolated from mice inoculated with pCACJ1 i.m. *secreted* interferon-gamma (IFN-gamma), but **not** interleukin (IL)-4, in vitro upon stimulation with Cry j 1 as well as with p277-288, a peptide corresponding to the T-cell epitope of Cry j 1.

2 Localization

Affirmative:

- 6403360: Stimulation of these cells with concanavalin A, phytohemagglutinin, or the UCHT1 monoclonal anti-human T cell antibody significantly increased the number of IFN-gamma-*secreting* cells.
- 6403360: Finally, cyclosporin A, a potent and selective immunosuppressive drug for T cells, strongly inhibited the *secretion* of IFN-gamma as assayed at the cell level.
- 2506311: IFN-gamma, also called immune interferon, is regarded as an important immunoregulator *secreted* by T-lymphocytes.
- 8661175: T lymphocytes (T cells) and their *secreted* lymphokine interferon-gamma (IFN-gamma) play important mediator roles in endotoxin-induced inflammation.
- 8661175: T lymphocytes (T cells) and their *secreted* lymphokine interferon-gamma (IFN-gamma) play important mediator roles in endotoxin-induced inflammation.

Negative:

- 9272363: In an investigation of cell-mediated immunity against Bordetella pertussis, we found that B. pertussis infection in infants and in mice was associated with the induction of antigen-specific T cells that *secrete* IFN-g and IL-2, but **not** IL-4 or IL-5.
- 12435401: Productively stimulated nai;ve T cells expressed IL-2 to differentiate into T helper 1 (Th1) cells, secreting interferon-gamma (IFN-gamma) upon secondary antigen stimulation; T cells primed with an APL did **not** s*ecrete* either interleukin-4 (IL-4) or IFN-gamma, but expressed TGF-beta1 and Tob, a member of the anti-proliferative gene family.
- 6436035: It is evident that IFN-gamma is **not** the only macrophage activator *released* by T lymphocytes responding to microbial antigen, and may not even be the main one to enhance antimicrobial activity in infections such as tuberculosis.
- 3096401: T lymphocytes did **not** *release* these activities in the absence of PHA with or without HuIFN gamma.
- 10358144: We conclude that in a microenvironment in which allogeneic EC are in close contact with infiltrating CD8+ T cells, such as within a graft arterial intima, CTL subsets may emerge that display EC selectivity or express CD40L and *secrete* **little** IFN-gamma after Ag contact.

3 Localization

Affirmative:

- 2531154: None of the generated T cell hybridomas exhibited antigen-specific IL-2 *secretion* when stimulated with autologous thyrocytes, although 60% of the hybridomas expressed CD3 antigen and the T cell receptor alpha/beta heterodimer.
- 2448378: Surface IL-2 epitopes were also detected on the Jurkat tumor cell line which *secretes* IL-2 upon stimulation and on another T cell tumor line MOLT 4.
- 2448378: Although it is possible that the epitopes seen were present on a distinct molecule independent of *secreted* IL-2, the distribution on a variety of T cells and regulation via cellular activation suggest that the surface expression of IL-2 epitopes is in some way related to the soluble lymphokine.
- 6213706: Mitogen stimulation led to *secretion* of equivalent amounts of IL 2 from both the major T cell subsets; in contrast, after allogeneic activation, IL 2 was produced predominantly from the T4+ subset.
- 6215193: In order to *release* Il-2, the OKT4 positive T cell requires a stimulus, such as allogeneic cells or the lectin phytohaemagglutinin A (PHA).

Negative:

- 12946104: Stimulation of T cells in the presence of CB sera increased the frequency of IL-2 producing cells ($p < 0.005$) (but **not** the amount of IL-2 *secreted*) and resulted in a higher expression of CD25 ($p < 0.05$).
- 6601235: Blood T cells from the same patients did **not** *release* interleukin-2.
- 2967331: We identified human T cell clones which *secrete* IL-4 but **not** IL-2 or IFN-gamma, and which appeared to correspond to murine Th2 clones.
- 1921263: It also revealed through the use of an in vitro assay utilizing the human IL-2 dependent cell line, Sez 627, that **none** of these T cell lines *secrete* IL-2 in detectable volumes.
- PMC2935971: That said, the failure of IL-2 to expand IL-17 producing CD4+ T cells while increasing T-reg populations may augur well for IL-2 use in auto-immunity, diseases characterized by depleted T-reg populations, and elevated IL-17 expression.

4 Localization

Affirmative:

- 9352159: In addition, HGF *excretion* tended to correlate with disease severity as higher levels were observed in patients with oliguric ATN.
- 19423096: Estradiol-induced HGF *secretion* by uterine stromal fibroblasts may have a significant effect on uterine cancer and endometriosis.
- 17765959: CONCLUSION: Irradiation-enhanced HGF *secretion* in all 3 tested glioma cell lines (up to 7 times basal levels).
- 17765959: It is tempting to associate the radiation-enhanced HGF *secretion* with an increased angiogenic potential of the tumor, which may be a factor in radioresistance.
- 10448071: Levels of HGF *released* to culture media and of HGF mRNA increased when cultures were exposed to NE, or to other adrenergic

agonists.

Negative:

- 8391878: TPA stimulated HGF *release* from the cells through an activation of C-kinase, but **not** through a formation of reactive oxygen species.
- 11145707: Hepatocyte growth factor (HGF) is a potent paracrine mediator of stromal/epithelial interactions, which is *secreted* as a matrix-associated **inactive** precursor (pro-HGF) and locally activated by tightly controlled urokinase cleavage.
- 9062512: Using immunohistochemistry, HGF *localized* to the villous core compartment with **no** localization to the trophoblast.
- 7683665: HGF is homologous to plasminogen and is first synthesized and *secreted* as an **inactive** single-chain precursor and then activated to a heterodimeric form by endoproteolytic processing.
- 11145707: Hepatocyte growth factor (HGF) is a potent paracrine mediator of stromal/epithelial interactions, which is *secreted* as a matrix-associated **inactive** precursor (pro-HGF) and locally activated by tightly controlled urokinase cleavage.

5 Localization

Affirmative:

- 2402594: Interleukin 1, but not interleukin 1 inhibitor, is *released* from human monocytes by immune complexes.
- 2402594: Interleukin 1, but not interleukin 1 inhibitor, is *released* from human monocytes by immune complexes.
- 2402594: This investigation shows that tetanus toxoid-human anti-tetanus toxoid IC induce human monocytes to *release* IL-1.
- 6334697: Furthermore, alveolar macrophages *released* significantly less IL-1 than blood monocytes (26 +/- 11 vs. 128 +/- 21 U/10(6) cells X 24 h, respectively, after stimulation with 10 micrograms/ml of LPS, P less than 0.001).
- 3875766: An intracellular monocyte derived protein possessing interleukin 1 (IL-1) activity has been compared with the *secreted* from of IL-1.

Negative:

- 2026475: The mechanisms of IL-1 beta release by carbonyl-iron or sheep red blood cells may be related to their phagocytosis, as non-phagocytic monocytes did **not** *release* IL-1 beta.
- 3496273: Peritoneal M phi preparations from women in the pre-luteal phase did **not** *release* detectable IL-1, whereas those from women in the post-luteal phase released as much as monocytes.
- 3496273: Cultured monocytes **failed** to *secrete* IL-1 and expressed less DQ than fresh monocytes.
- 8228247: Human blood monocytes synthesize but do **not** *secrete* IL-1 beta in response to low doses of bacterial cell-wall products.
- 3484775: The results demonstrate that when blood monocytes are prepared under low endotoxin conditions, they do **not** spontaneously *secrete* IL-1 activity.

1 Transcription

Affirmative:

- 18024275: The *expressions* of HGF mRNA and protein were confirmed in

the transfected BMSCs.

- 8054485: Moreover, HGF mRNA was *detected* in high HGF producers by Northern blot analysis.
- 10421795: *Transcription* of HGF and its receptor, c-met, was detected by reverse transcription-polymerase chain reaction (RT-PCR).
- 9716011: HGF and c-met mRNAs were *detected* by reverse transcriptase-polymerase chain reaction (RT-PCR) and Northern blotting.
- 9705867: Cultured adult and fetal human RPE *expressed* mRNA for HGF and c-Met by RT-PCR.

Negative:

- 10452684: The intensity of EGFR expression was consistent, and HGF mRNA was **not** *detected* during induction experiments in any cell type.
- 7629039: Although c-met protein was expressed in the cytotrophoblast, this receptor was not detectable in the syncytiotrophoblast by immunohistochemical methods. c-met mRNA was detected in placental cell line (tPA30-1) and 4 choriocarcinoma cell lines (BeWo, Jar, Jeg-3, and NUC-1), but HGF mRNA was absent in these cells.
- 7720876: The active form of HGF was **not** *detected* under our experimental conditions after these operations.
- PMC2911419: We also found that **neither** HGF mRNA nor protein was *expressed*, suggesting a ligand-independent mechanism of Met phosphorylation.
- 11768718: HGF and c-met mRNAs were clearly *detected* in HLMECs before and after treatment with IL-1beta, but **not** in HUVECs.

2 Transcription

Affirmative:

- 8844635: The age-related decline in IL-2 production has been shown to arise from a decline in IL-2 *transcription*, and a recent study suggests that the transcription factor NFAT (nuclear factor of activated T cells) may play a role in the decline in IL-2 transcription.
- 8844635: The age-related decline in IL-2 production has been shown to arise from a decline in IL-2 transcription, and a recent study suggests that the transcription factor NFAT (nuclear factor of activated T cells) may play a role in the decline in IL-2 *transcription*.
- 11897658: The normally repressed IL-2 gene is *transcribed* in nuclei from quiescent human T cells and from various non-T-cell lines.
- 3135859: CsA also prevents the constitutive secretion of IL-2 in this T-cell line by blocking *transcription* of the IL-2 gene.
- 11699390: Triptolide inhibits both Ca(2+)-dependent and Ca(2+)-independent pathways and affects T cell activation through inhibition of interleukin-2 *transcription* at a site different from the target of cyclosporin A.

Negative:

- 2785866: The following similarities in the functional biological characteristics of T cell and B cell IL-2 suggest that B cell IL-2 is not a factor which mimics IL-2 activity in the CTLL-2 assay: (i) neutralization of IL-2 by anti-IL-2 monoclonal antibody (DMS-1); (ii) elution of IL-2 following its adsorption to CTLL-2 cells; (iii) determination of the MW of IL-2 by SDS-PAGE and Western blot analysis; and (iv) ability of B cell IL-2 to support T cell proliferation and blocking of this activity by anti-tac

monoclonal antibody. cDNA probes for T cell IL-2, however, did **not** *detect* IL-2 mRNA in B cells.

- 2899602: In these cell lines, IL-2 mRNA was **not** *detectable* in RNA extracted from whole adult T cell leukemia cell populations because of dilution by other RNA species from the vast majority of cells that do not contain IL-2 mRNA.
- 9311917: In the absence of costimulation, T cells activated through their antigen receptor become unresponsive (anergic) and do **not** *transcribe* the gene encoding interleukin-2 (IL-2) when restimulated with antigen.
- 9311917: In the absence of costimulation, T cells activated through their antigen receptor become unresponsive (anergic) and do **not** *transcribe* the gene encoding interleukin-2 (IL-2) when restimulated with antigen.
- 16806475: In the absence of IL-2, human CD4(+) T cell blasts were sensitive to both FasL and Apo2L/TRAIL, but human CD8(+) T cell blasts died, with no additional effect of death receptor ligation.

3 Transcription

Affirmative:

- 2439327: To study the regulation of the murine IL2 gene in T-cell populations of differing stages of maturation, we have used a calcium ionophore in conjunction with the phorbol ester, TPA, to stimulate IL2 gene *transcription* while bypassing the requirement for triggering through a mature cell surface receptor.
- 8342142: CONCLUSIONS: These results suggest that the principal cellular abnormalities that result in altered T cell activation and IL-2 production after thermal injury lie downstream of the initiating signal transduction events and before IL-2 gene *transcription*.
- 17096403: Of the T cell cytokines assessed, there was a marked reduction in the mRNA *expression* of interleukin-2 (IL-2) in Nude mice compared with wildtype animals.
- 17096403: Of the T cell cytokines assessed, there was a marked reduction in the mRNA *expression* of interleukin-2 (IL-2) in Nude mice compared with wildtype animals.
- PMC2526191: Previous studies show that IL-2 production is tightly regulated (20, 21), and that even under optimal conditions not all T cells in a population will acquire the competence to *transcribe* the Il2 gene and synthesize IL-2 upon primary stimulation (22).

Negative:

- 15937196: These results identify a discrete new domain of IL2 regulatory sequence marked by dimethylated histone H3/K4 in expression-permissive T-cells even when they are **not** *transcribing* IL2, setting boundaries for histone H3 and H4 acetylation when the IL2 gene is transcriptionally activated.
- 8878449: We could **not** *detect* human T cell leukemia virus type I (HTLV-I) mRNA or interleukin 2 (IL-2) mRNA in either the tumor cells growing in mice or the original leukemic cells.
- 8878449: We could **not** *detect* human T cell leukemia virus type I (HTLV-I) mRNA or interleukin 2 (IL-2) mRNA in either the tumor cells growing in mice or the original leukemic cells.
- PMC1142491: These results identify a discrete new domain of IL2 regulatory sequence marked by dimethylated histone H3/K4 in expression-

permissive T-cells even when they are **not** *transcribing* IL2, setting boundaries for histone H3 and H4 acetylation when the IL2 gene is transcriptionally activated.
- 3115639: Anti-CD3-MoAb, in the absence of IL-2, induced IL-2 receptor expression on purified T cells, and anti-IL 2 receptor antibodies inhibited T cell proliferation in the presence of this growth factor.

4 Transcription

Affirmative:
- 8461246: Angiotensin-converting enzyme (ACE) inhibitors are now widely *prescribed* for the treatment of hypertension and heart failure.
- 8461246: Angiotensin-converting enzyme (ACE) inhibitors are now widely *prescribed* for the treatment of hypertension and heart failure.
- 9562936: Patterns of angiotensin-converting enzyme inhibitor *prescriptions*, educational interventions, and outcomes among hospitalized patients with heart failure.
- 9562936: BACKGROUND: Among hospitalized patients with heart failure, we describe characteristics associated with *prescription* of angiotensin-converting enzyme (ACE) inhibitors in the doses recommended by clinical practice guidelines.
- 9562936: BACKGROUND: Among hospitalized patients with heart failure, we describe characteristics associated with *prescription* of angiotensin-converting enzyme (ACE) inhibitors in the doses recommended by clinical practice guidelines.

Negative:
- 10908091: Therefore, we recommend that physicians continue to *prescribe* ACE inhibitors for patients with heart **failure** based on the target doses used in the placebo-controlled trials and not on the "high" dose target used in ATLAS.
- 11831455: Captopril, enalapril, and lisinopril are angiotensin-converting enzyme (ACE) inhibitors widely *prescribed* for hypertension and heart **failure**.
- 11052861: Although most primary care physicians stated they *prescribe* ACE inhibitors in heart **failure**, this was for only 47-62% of patients, and at doses below those identified as effective in trials.
- 9491949: BACKGROUND: Angiotensin-converting enzyme (ACE) inhibitors were *underprescribed* for patients with congestive heart **failure** (CHF) treated in the community setting in the early 1990s despite convincing evidence of benefit.
- 9491949: BACKGROUND: Angiotensin-converting enzyme (ACE) inhibitors were *underprescribed* for patients with congestive heart **failure** (CHF) treated in the community setting in the early 1990s despite convincing evidence of benefit.

5 Transcription

Affirmative:
- 3876332: Induction of interleukin 2 (IL2) mRNA *synthesis* in human tonsillar lymphocytes was studied by quantifying the relative levels of IL2 mRNA in the lymphocytes stimulated under various conditions by the dot hybridization method.
- 3876332: Induction of interleukin 2 (IL2) mRNA *synthesis* in human

tonsillar lymphocytes was studied by quantifying the relative levels of IL2 mRNA in the lymphocytes stimulated under various conditions by the dot hybridization method.

- 1964589: Further study demonstrated that the rate of degradation of 32P-labeled IL-2 mRNA, which was prepared by cell-free *transcription* of IL-2 cDNA, in the polysomal fraction obtained from PDB-stimulated lymphocytes was decreased compared with that obtained from unstimulated lymphocytes.
- 7685613: Concomitant with the inhibition of lymphocyte activation and interleukin 2 (IL-2) production, *transcription* of the IL-2 message is also reduced in a time-dependent manner.
- 1601521: In contrast, tumor-infiltrating lymphocytes (TIL) in the stroma of ovarian carcinomas or most ductal breast tumors only rarely *expressed* mRNA for TNF alpha, IL2 or IFN gamma.

Negative:

- 2788180: Patients' lymphocytes whose IL-2 responsiveness was decreased still expressed Tac antigen (low-affinity IL-2 receptors) but, in contrast to the patients' original lymphocytes, did **not** *absorb* or respond to IL-2, suggesting the loss of high-affinity IL-2 receptors (p55/p75) from these cells.
- 3928744: Furthermore, IL 2-specific mRNA was **not** *detected* in TCD-stimulated PBL, demonstrating that IL 2 was not required for TCD-induced T cell proliferation.
- 8181865: A primary tumor cell line was generated and cultured TIL were induced to *transcribe* IL-2 and IFN-gamma genes by incubation with the autologous irradiated tumor cell line, but **not** with autologous EBV-transformed cells.
- 1973607: After 2-3 weeks under immunosuppressive treatment with prednisolone and azathioprine, however, BP lymphocytes did **not** *exhibit* any IL2 receptors.
- 1356393: In contrast, LNL not adjacent to the tumor in involved LN, as well as those in tumor-uninvolved LN, did **not** *express* mRNA for cytokines or IL2 receptor.

1 Protein catabolism

Affirmative:

- 17482444: These studies thus demonstrate that betanin induces apoptosis in K562 cells through the intrinsic pathway and is mediated by the release of cytochrome c from mitochondria into the cytosol, and PARP cleavage.
- 10986477: Apoptosis is often associated with PARP cleavage and caspase activation.
- 10986477: Fibres did not cause PARP cleavage or activation of caspase 3 further confirming previous results about relatively low apoptotic potential of asbestos fibres.
- 10200351: As a logical link with DNA fragmentation analyses and TUNEL assay, cleavage of the 116 kDa PARP protein was accompanied by the appearance of a characteristic 85 kDa fragment of PARP in a population of floating cells after both treatments.
- 10726984: Apoptosis was researched by DAPI staining, annexin V-binding, electron microscopy, DNA fragmentation and PARP cleavage.

Negative:

- 15309525: Cell morphology and Western blot analyses revealed that the antibody-induced cell death **lacked** some typical features of apoptosis such as chromatin condensation or poly-ADP-ribose polymerase (PARP) cleavage.
- 12954616: Furthermore, overexpression of caspase-9 did **not** enhance PARP or caspase-7 cleavage after UV treatment.
- 18678619: ASA and NaS at 1 mM did **not** induce PARP cleavage or caspase-3 and at 5 mM, ASA but not NaS increased apoptosis.
- 11002424: The caspase inhibitors Z-VAD-FMK and Z-DEVD-FMK blocked apoptosis induced by CD437 in DU145 and LNCaP cells, in which increased caspase-3 activity and PARP cleavage were observed, but not in PC-3 cells, in which CD437 did **not** induce caspase-3 activation and PARP cleavage.
- PMC1665652: Synchronized cells, in the absence of Vpr expression, did **not** display PARP cleavage after release (unpublished data).

2 Protein catabolism

Affirmative:

- 10589695: The cytosolic proteins of the parathyroids were used to study PTH mRNA protein binding by ultraviolet cross-linking and the *degradation* of the PTH transcript in vitro.
- 1001258: The similarities found in the sites of hormone proteolysis and in the kinetics of hormone metabolism in the rat and dog, coupled with the less direct evidence indicating that similar cleavages are also present in man and bovine, are consistent with the view that *proteolysis* of parathyroid hormone is peripheral tissues is specific, at least in mammalian species, and may be a critical step in controlling the availability of biologically active hormone.
- 6994557: These studies, however, coupled with (1) further investigations of intracellular *degradation* of parathyroid hormone, if this indeed operates in vivo; (2) the proteolytic conversion of secreted hormone in peripheral tissues; and (3) analysis of transcriptional control of biosynthesis of parathyroid hormone, using radioactive cDNA for hybridization studies of mRNA production and turnover, hold great promise for further understanding of critical regulatory factors central to expression of the actions of parathyroid hormone.
- 8467581: Particular emphasis is given to the calcium-stimulated *degradation* of parathyroid hormone within the parathyroid as one of the major pathways by which circulating levels of bioactive hormone are controlled.
- 9608898: The pathogenesis of renal osteodystrophy is related to phosphate retention, and its effect on calcium and calcitriol metabolism, in addition to roles played by metabolic acidosis, cytokines, and *degradation* of parathyroid hormone.

Negative:

- 10589695: With uremic parathyroid proteins, the PTH mRNA was **not** *degraded* at all at 120 min and was moderately decreased at 180 min.
- 7428706: These membranes were **unable** to *degrade* parathyroid hormone (PTH), bovine PTH-(1-84) [bPTH-(1-84)], or bPTH-(1-34).
- 1236605: Biologically active, labelled parathyroid hormone was *degraded* to fragments by rat kidney membranes, but **not** by chick kidney. 4.

- 7428706: These membranes were **unable** to *degrade* parathyroid hormone (PTH), bovine PTH-(1-84) [bPTH-(1-84)], or bPTH-(1-34).

3 Protein catabolism

Affirmative:

- 5796358: Finally, a three pool model was formulated to describe the kinetics of plasma insulin disappearance in man, representing plasma (pool 1), interstitial fluid (pool 2), and all tissues in which insulin is utilized and *degraded* (pool 3).
- 3532665: Insulin *degradation* is unaffected by pregnancy and the proinsulin share of the total plasma insulin immunoreactivity does not increase in pregnancy.
- 1102319: The fact that MCR did not fall in the OC group with increasing plasma insulin concentrations whereas it did in normal subjects, suggested that OC leads to the loss of saturable component of insulin *degradation* that is present in normal subjects.
- 11217151: Analysis of plasma 123I-insulin immunoreactivity and trichloroacetic acid precipitate showed that insulin *degradation* did not occur as in normal controls.
- 3888744: Insulin *degradation* is unaffected by human pregnancy and the proinsulin share of the total plasma insulin immunoreactivity does not increase in pregnancy.

Negative:

- 403392: With the use of a specific enzyme which *degrades* insulin but **not** proinsulin, postprandial plasma proinsulin values have been measured in a large number of subjects under a variety of physiologic and pathologic conditions.
- 403392: With the use of a specific enzyme which *degrades* insulin but **not** proinsulin, postprandial plasma proinsulin values have been measured in a large number of subjects under a variety of physiologic and pathologic conditions.
- 6445191: The plasma membranes do **not** *degrade* insulin significantly in the absence of reduced glutathione, and over 99% of the cellular degradative capacity is found in the postmicrosomal supernatant (cytosol).
- 3283938: Epstein-Barr virus-transformed lymphocytes from this patient synthesize an insulin receptor precursor that is normally glycosylated and inserted into the plasma membrane but is **not** *cleaved* to mature alpha and beta subunits.

4 Protein catabolism

Affirmative:

- 9493967: PARP cleavage in the apoptotic pathway in S2 cells from Drosophila melanogaster.
- 9493967: Further experiments must be conducted and the peptide fragments must be sequenced to relate protease activities with PARP cleavage.
- 9987014: We conclude that HI injury or traumatic injury to the developing rat forebrain leads to PARP cleavage in directly affected areas and in sites distant from the primary injury that precedes the appearance of cells with apoptotic morphology.
- 15196974: In the 'moderately' exposed neurons, ATP depletion to 59+/-6%

of control was associated with a decrease in the cell counts, apoptotic morphology, and cleavage of PARP.

- 11120353: Serum-starved SHE cells were compared to the etoposide-treated HL60 cell line as a control for typical apoptosis-related PARP cleavage.

Negative:

- 10497198: In vitro experiments showed that PARP cleavage by caspase-7, but **not** by caspase-3, was stimulated by its automodification by long and branched poly(ADP-ribose).
- 9781697: Taken together, in vivo phosphorylation of PARP might be **independent** of the activation or cleavage of PARP.
- 14559808: However, E1A did **not** induce PARP cleavage but rather suppressed PARP expression at the transcriptional level.
- PMC2880028: In our analysis, cleavage of PARP is the key output; because the process is all-or-**none**, if >50% of PARP is *cleaved*, it is eventually all cleaved and thus a simulated cell is deemed dead at 50% cleaved PARP (see Methods).
- 9973225: These drugs also did **not** increase caspase-3 or PARP cleavage when combined with TRAIL.

5 Protein catabolism

Affirmative:

- 11809529: Treatment of KB cells with an apoptosis-inducing concentration of [6]-dehydroparadol caused induction of proteolytic cleavage of pro-caspase-3.
- PMC2949386: TP187 decreases the number of proliferating cells and induces caspase-3 cleavage in tumor xenografts
- 10049561: This inhibition correlated with the absence of the Gas2 peptide and pro-caspase-3 cleavage.
- 11872639: Topical genistein completely inhibited cleavage of PARP and caspase-3.
- 16014577: Apoptosis was preceded by cleavage of caspase-3 (4-6 h) and caspase-8 (6-8 h) and their respective substrates, alpha-fodrin and Bid.

Negative:

- PMC2775411: Indeed, there were no increase of DEVDase (effector caspase) activity, **no** proteolytic cleavage of caspase-3, no DNA fragmentation, and no apoptotic bodies on histological examination.
- 9708735: Moreover, the cytochrome c-mediated cleavage of Casp3 is **absent** in the cytosolic extracts of Casp9-deficient cells but is restored after addition of in vitro-translated Casp9.
- 11470470: Western blot analysis confirmed that in cardiomyopathies **no** cleavage of caspase-3 and caspase-7 occurred.
- PMC2858442: The inhibitor may bind caspase-8 only after the so-called substrate switch that is necessary for caspase-3 and Bid cleavage but **not** for p43/p41 and p43-FLIP formation (Hughes et al, 2009).

1 Phosphorylation

Affirmative:

- 8945479: Activated Lck *phosphorylates* T-cell receptor zeta-chains, which then recruit the ZAP70 kinase to promote T-cell activation.

- 17100647: We report the successful expression and detection of a *phosphorylated* form of human T cell tyrosine kinase, Lck, in Saccharomyes cerevisiae, which leads to growth suppression of the yeast cells.
- 2481585: Within minutes after activation of a human T cell-derived line (Jurkat) via stimulation of either the TcR-CD3 complex or the CD2 glycoprotein, we observed a *hyperphorphosylation* of p56lck.
- 16107303: Furthermore in Jurkat T cell extracts, a recombinant intron B plus SH3 p56lck domain fails to interact with some TCR-induced tyrosine *phosphorylated* polypeptides and known p56lck partners such as Sam68 and c-Cbl.
- 1535787: Triggering of T cells with a combination of anti-CD3 mAbs which activate T cells but not p56lck and gp160 greatly potentiated the increase of p56lck *autophosphorylation* and kinase activity.

  Negative:
- 8183556: Recently, we found that p50csk specifically *phosphorylates* the **negative** regulatory Tyr-505 of the T cell-specific src-family kinase p56lck, and thereby suppresses its catalytic activity.
- 9581568: Thus CD45 is intrinsically a much more active phosphatase than RPTPalpha, which provides one reason why RPTPalpha **cannot** effectively *dephosphorylate* p56(lck) and substitute for CD45 in T-cells.
- 8663155: In contrast to the T cell protein tyrosine kinase, Lck, ZAP-70 did **not** *phosphorylate* the cytoplasmic portion of the TCRzeta chain or short peptides corresponding to the CD3epsilon or the TCRzeta immunoreceptor tyrosine-based activation motifs.
- PMC2994893: As in primary T cells, in Hut-78 cells, Lck Tyr505 was basally *phosphorylated* and **not** dephosphorylated upon TCR stimulation.
- 8663155: In contrast to the T cell protein tyrosine kinase, Lck, ZAP-70 did **not** *phosphorylate* the cytoplasmic portion of the TCRzeta chain or short peptides corresponding to the CD3epsilon or the TCRzeta immunoreceptor tyrosine-based activation motifs.

2 Phosphorylation

Affirmative:

- 2108026: When a non-platelet aggregatory deoxyphorbol (12-deoxyphorbol 13-phenylacetate 20-acetate) was combined with a subthreshold dose of the Ca2+ ionophore, A23187, a large increase in *phosphorylation* of p47 and a fourfold decrease in Ka was observed.
- 7559410: Phosphatidylinositol (3,4,5)-trisphosphate stimulates *phosphorylation* of pleckstrin in human platelets.
- 7559410: In such platelets, serine- and threonine-directed *phosphorylation* of pleckstrin also occurs and has been attributed to protein kinase C activation.
- 7559410: Pleckstrin *phosphorylation* in response to thrombin receptor stimulation is progressively susceptible to inhibition by wortmannin, a potent and specific inhibitor of platelet PI 3-kinases.
- 7559410: Synthetic PtdIns(3,4,5)P3, when added to saponin-permeabilized (but not intact) platelets, causes wortmannin-insensitive *phosphorylation* of pleckstrin.

  Negative:
- 2849942: Human platelet smg p21 was **not** *phosphorylated* by protein

kinase C.
- 7608115: Thus, human platelet tubulin is **not** *phosphorylated* either in unstimulated platelets or in thrombin-stimulated platelets.
- 2547793: However, a crude platelet kinase preparation *phosphorylated* ABP in the presence of cAMP, but **not** in the presence of Ca2+/phosphatidylserine.
- 2391768: Furthermore, Con A was shown to stimulate the protein kinase C activity of platelets, which *phosphorylates* a 40-kDa platelet protein; the Con A effects were antagonized by alpha-methyl-D-mannoside, staurosporine and K-252a, but **not** by KT5720.

3 Phosphorylation

Affirmative:
- 1546747: Immunocytochemical studies with antibodies against alpha tubulin, tau, and *phosphorylated* subunits of neurofilament polypeptides did not disclose differences in the staining of neurons with fragmented or normal Golgi apparatus, suggesting that the alteration of the organelle is not secondary to a gross lesion of the cytoskeleton.
- PMC2228393: Neurofibrillary tangles consist of *hyperphosphorylated* Tau proteins that aggregate inside neurons along neurites â€“ observed as neuropil threads â€“ and finally in the soma.
- 8336148: Okadaic acid induces early changes in microtubule-associated protein 2 and tau *phosphorylation* prior to neurodegeneration in cultured cortical neurons.
- 8395566: Microtubule-associated protein tau is known to be *hyperphosphorylated* in Alzheimer disease brain and this abnormal hyperphosphorylation is associated with an inability of tau to promote the assembly of microtubule in the affected neurons.
- 7689658: We show here that tau can be *phosphorylated* in cultured hippocampal neurons by the MAP kinase p44mpk, and phosphorylation of tau compromises its functional ability to assemble microtubules.

Negative:
- 17028556: At the same time, Tau-P301LxGSK-3B mice have dramatic forebrain tauopathy, with "tangles in almost all neurons", although **without** *hyper-phosphorylation* of Tau.
- 19190923: Importantly, we detected a few neurons that contained abundant truncated tau but were **lacking** *hyperphosphorylation*, and these neurons exhibited remarkable nuclear condensation.
- 8317268: Furthermore, the pretangle neurons can readily be immunolabeled for abnormally *phosphorylated* tau but **not** for ubiquitin.

4 Phosphorylation

Affirmative:
- 12688680: *Phosphorylation*, but not overexpression, of epidermal growth factor receptor is associated with poor prognosis of non-small cell lung cancer patients.
- 14734462: Because the ErbB receptors play an important role in lung cancer progression, we analyzed the expression of epidermal growth factor receptor (EGFR), *phosphorylated* EGFR, transforming growth factor alpha (TGFalpha), and HER2-neu as potential prognostic factors in stage I NSCLC.

- 14734462: Because the ErbB receptors play an important role in lung cancer progression, we analyzed the expression of epidermal growth factor receptor (EGFR), *phosphorylated* EGFR, transforming growth factor alpha (TGFalpha), and HER2-neu as potential prognostic factors in stage I NSCLC.
- 14734462: Because the ErbB receptors play an important role in lung cancer progression, we analyzed the expression of epidermal growth factor receptor (EGFR), *phosphorylated* EGFR, transforming growth factor alpha (TGFalpha), and HER2-neu as potential prognostic factors in stage I NSCLC.
- 19235531: *Phosphorylated* epidermal growth factor receptor and cyclooxygenase-2 expression in localized non-small cell lung cancer.

Negative:

- 18585821: Although antibodies against phophorylated EGFR have been used in vitro, *phosphorylated* EGFR has yet **not** been examined well in resected non-small cell lung cancers (NSCLCs).
- 16373402: EGFR and its downstream proteins were constitutively *phosphorylated* in the PC-9 cells **without** any ligand stimulation as compared with A549 lung cancer cells expressing wild-type EGFR.
- 16505275: High predictive value of epidermal growth factor receptor *phosphorylation* but **not** of EGFRvIII mutation in resected stage I non-small cell lung cancer (NSCLC).

5 Phosphorylation

Affirmative:

- 17334396: However, it has been proposed that tyrosine *phosphorylation* of p120 may contribute to cadherin-dependent junction disassembly during invasion.
- 16904204: p120 catenin and *phosphorylation*: Mechanisms and traits of an unresolved issue.
- 15684660: This role is probably regulated by signaling events that induce p120 *phosphorylation*, but monitoring individual phosphorylation events and their consequences is technically challenging.
- 15684660: Previously, we used phospho-tryptic peptide mapping to identify eight major sites of p120 serine and threonine *phosphorylation*.
- 17719574: In contrast to growth factor-stimulated tyrosine *phosphorylation* of p120, its relatively constitutive serine/threonine phosphorylation is not well understood.

Negative:

- 17334396: In contrast, p120 knockdown impairs epidermal growth factor-induced A431 invasion into three-dimensional matrix gels or in organotypic culture, whereas re-expression of siRNA-resistant p120, or a p120 isoform that **cannot** be *phosphorylated* on tyrosine, restores the collective mode of invasion employed by A431 cells in vitro.
- 11382764: Changing all of these sites to phenylalanine resulted in a p120 mutant, p120-8F, that could not be efficiently *phosphorylated* by Src and **failed** to interact with SHP-1, a tyrosine phosphatase shown previously to interact selectively with tyrosine-phosphorylated p120 in cells stimulated with epidermal growth factor.
- 15107817: In addition, p120(ctn) connected with N-cadherin was phosphorylated at tyrosine residues, whereas the isoform linked to E-

cadherin was **not** *phosphorylated*.

1 Binding

Affirmative:

- 10433099: Apoptotic cell death mediated by an *interaction* of Fas with Fas ligand (FasL) could be a mechanism by which MHC class II-negative pancreatic beta-cells are destroyed by CD4+ T lymphocytes.
- 15814689: *Interaction* of Fas with Fas ligand (FasL) is known to play a role in peripheral tolerance mediated by clonal deletion of Ag-specific T cells.
- 7512035: In addition, activated T cells from gld/gld homozygous animals are not capable of *binding* to a Fas.Fc fusion protein at high levels, whereas activated T cells from normal and lpr/lpr animals bind Fas.Fc efficiently.
- 15153474: However, in the context of an extensive tumor burden, chronic stimulation of such CD4(+) T cells often leads to the up-regulation of both Fas and Fas *ligand*, and coexpression of these molecules can potentially result in activation-induced cell death and the subsequent loss of effector activity.
- 12920696: Apoptotic cell death may be induced by either cytotoxic T cells through the release of proteases, such as perforin and granzyme B, or the *interaction* of Fas ligand (FasL/CD95L), expressed by T lymphocytes, with Fas (Apo-1/CD95) on epithelial cells.

Negative:

- 12153509: Taken together with the fact that DN T cells massively express Fas ligand (FasL), this study implied that FasL overexpressed on DN cells may be involved in the accumulation of DN T cells in LN, LN atrophy and wasting syndrome, and that lprcg Fas, which can *bind* to Fas ligand but **not** transduce apoptosis signal into cells, may modulate these pathological conditions by interfering with the binding of FasL to Fas.
- 16133864: In this model, islet grafts from C3H mice that carry the lpr mutation, and therefore **lack** the ability to undergo apoptosis through CD95-CD95L *interaction*, were completely protected when grafted in autoimmune diabetic mice despite periinsulitis (infiltration of T cells) which however did not progress to islet destruction.
- 16133864: In this model, islet grafts from C3H mice that carry the lpr mutation, and therefore **lack** the ability to undergo apoptosis through CD95-CD95L *interaction*, were completely protected when grafted in autoimmune diabetic mice despite periinsulitis (infiltration of T cells) which however did not progress to islet destruction.
- 12232795: T cell apoptosis was observed at relatively low concentrations of kynurenines, did **not** require Fas/Fas ligand *interactions*, and was associated with the activation of caspase-8 and the release of cytochrome c from mitochondria.
- 12232795: T cell apoptosis was observed at relatively low concentrations of kynurenines, did **not** require Fas/Fas ligand *interactions*, and was associated with the activation of caspase-8 and the release of cytochrome c from mitochondria.

2 Binding

Affirmative:

- 2522840: Rather, the *binding* of CD4 MoAb to CD4+ T cells interferes with

a late event because it is capable of abolishing the proliferative activity of fully activated CD4+ T cells.

- 8102120: Surface Ig levels of CD4+ T cells were closely *associated* with the CD4 cell number in HIV-infected patients of all stages of disease (r = -0.67, P = 0.00005).
- 9630922: The exocyclic derived from CDR3 (residues 82-89) of human CD4, which specifically *associated* with CD4 on the T cell surface to create a heteromeric CD4 complex, blocked IL-2 production and antagonized the normal function of the CD4 receptor.
- 1460417: CD4+ CTL were shown to *recognize* not only the infected cells within these acutely infected cultures but also noninfected CD4+ T cells that had passively taken up gp120 shed from infected cells and/or free virions.
- 1701034: In normal T cells, surface *association* of CD4 molecules with other CD4 molecules or other T-cell surface proteins, such as the T-cell antigen receptor, stimulates the activity of the p56lck tyrosine kinase, resulting in the phosphorylation of various cellular proteins at tyrosine residues.

Negative:

- 1699999: In contrast to these results on class II-dependent T cell proliferation, MHC-independent T cell activation (via CD3 antibodies) was largely resistant to inhibition with the same dose range of CD4 mAb (provided that CD3 and CD4 reagents could **not** *compete* for the same class of FcR).
- PMC2405786: We also did **not** observe any significant *association* between CD4+ and CD8+ T-cell proliferation to anti-CD3 and SEB with CD4 count change (Table 1).
- PMC2718810: In line with this, only the *interaction* of CD8 depleted (CD4+), but **not** CD4 depleted (CD8+) T cells with monocytes resulted in the secretion of IL-2 and IL-10 into the cell culture supernatant (Figure 7E), consistent with MHC class II dependency of these two cytokines as shown in Figure 6B.
- PMC2959493: Hence, zanolimumab exerts its action through inhibition of CD4+ T cell signaling in concert with the induction of Fc-dependent ADCC and CD4 down-modulation (Fig. 2).

3 Binding

Affirmative:

- 17097690: The aim of this study is to investigate the effect of three cucurbitacins (Cuc) E, D and I on the bilirubin-albumin *binding*, both in human serum albumin (HSA) and in plasma.
- 10365540: The purpose of this study was to examine the displacement effect of valproate (VPA) on bilirubin-albumin *binding* in human serum albumin (HSA) and human plasma.
- 17364966: Peroxynitrite mild nitration of albumin and LDL-albumin *complex* naturally present in plasma and tyrosine nitration rate-albumin impairs LDL nitration.
- 9425126: Here we present evidence that in blood plasma peroxynitrite induces the formation of a disulphide cross-linked protein identified by immunological (anti-albumin antibodies) and biochemical criteria (peptide mapping) as a *dimer* of serum albumin.

- 18948018: It was shown that albumin derived from a chromatographic process, which had a *bilirubin*:albumin ratio similar to that observed in plasma, had a vibrant yellow appearance.

  Negative:
- 7415075: When albumin is prepared from human blood plasma by the cold ethanol method, nonesterified long chain fatty acids present in the plasma do **not** strictly *copurify* with albumin.
- 2794522: The antibody, raised in mice immunized with nonenzymatically glycated albumin isolated from human plasma, recognizes glycated epitopes residing in albumin but not in other plasma proteins, and does **not** *react* with unglycated albumin.
- 6164224: It is concluded that loss of albumin through the gut does **not** *account* for the depressed plasma albumin concentration in these patients.

4 Binding

Affirmative:
- 11602735: The anti-gp120(CD4BD) MAbs *complexed* with gp120 suppressed gamma interferon production as well as proliferation of gp120-specific CD4 T cells.
- 17157668: Gp120 *binding* to human CD4+ T cells was analyzed by flow cytometry.
- 17157668: Physiologically relevant concentrations of EGCG (0.2 micromol/L) inhibited *binding* of gp120 to isolated human CD4+ T cells.
- 17157668: CONCLUSION: We have demonstrated clear evidence of high-affinity binding of EGCG to the CD4 molecule with a Kd of approximately 10 nmol/L and inhibition of gp120 *binding* to human CD4+ T cells.
- 1614536: A large number of antibodies have been raised against CD4, the receptor on T cells for the envelope glycoprotein gp120 of the human immunodeficiency virus, and the site at which gp120 *binds* to CD4 has been delineated.

  Negative:
- 10779509: In this report, coreceptor functions of mutant human CD4 molecules, which have **no** or reduced *affinity* to an HIV envelope protein, gp120, were assessed in a murine T cell receptor/class II MHC recognition system.
- 1701034: **No** T-cell tyrosine protein kinase signalling or calcium mobilization after CD4 *association* with HIV-1 or HIV-1 gp120.
- 9443108: Intact Fn and Fn-CTHBD strongly **inhibit** the *interaction* of gp120/160 with soluble CD4 and, under low serum conditions, are capable of neutralizing the infectivity of HIV-1 for CD4-positive T cells.
- 9719434: Reagents which **block** the *interaction* of HIV-1 gp120 with CD4+ T cell are of therapeutic interest.

5 Binding

Affirmative:
- 14580119: These include reorganization of the actin and microtubule cytoskeleton, dynamic *interactions* between microtubules and actin filaments, effects of axon guidance molecules, actions of actin regulatory proteins, and dynamic changes in intracellular calcium signaling.
- 2775532: It is believed generally that actin filaments are attached to the cell membrane through an *interaction* with membranous actin-binding

proteins.

- 2901417: The results suggest that ADP-ribosylated actin acts as a capping protein which *binds* to the barbed ends of actin filaments to inhibit polymerization.
- 6893988: From these results it appears that the formation of bundles of actin filaments in microvilli and in cones is a two-step process, involving actin *polymerization* to form filaments, randomly oriented but in most cases having one end in contact with the plasma membrane, followed by the zippering together of the filaments by macromolecular bridges.
- 2402158: Examples include the formation of spindle fibers from tubulin during cell division and the *polymerization* of actin into the actin filaments of the pseudopod in chemotaxis.

  Negative:
- 11889122: Our measurements show that the slow spatial homogenization of the actin filament network, **not** actin *polymerization* or the formation of polymer overlaps, is the rate-limiting step in the establishment of an elastic actin network and suggest that a new activity of F-actin binding proteins may be required for the rapid formation of a homogeneous stiff gel.
- 8281943: The protease-treated actin was, however, neither able to spontaneously assemble into filaments **nor** to *copolymerize* with intact actin unless its tightly bound Ca2+ was replaced with Mg2+.
- 9147130: The enzymatically modified actin could *form* actin filaments after treatment with ADP-ribosylhydrolase but **not** after treatment with phosphodiesterase.
- 7107709: Hence, 95K protein is a rod-shaped, dimeric, Ca++- and pH-regulated actin binding protein that cross-links but does **not** *sever* actin filaments.
- 2162826: A 74-kDa protein (adseverin) derived from adrenal medulla severs actin filaments and nucleates actin polymerization in a Ca2(+)-dependent manner but does **not** form an EGTA-resistant complex with actin monomers, which is different from the gelsolin-actin *interaction*.

1 Positive regulation

Affirmative:

- 2138708: Human interleukin 4 (IL-4) *upregulates* Fc epsilon R2/CD23 expression on the surface of B lymphocytes.
- 2138708: Human interleukin 4 (IL-4) *upregulates* Fc epsilon R2/CD23 expression on the surface of B lymphocytes.
- 2136716: Furthermore, IL-4 *induces* Fc epsilon-receptor (CD23) expression on 30% of unstimulated human B cells, whereas BCAF-containing supernatants from clone P2, that do not contain detectable amounts of IL-4, promote B cell proliferation without inducing CD23 expression.
- 9509417: While CD23(a) is constitutively expressed in B cells, the expression of CD23(b) is specifically *induced* by interleukin-4 (IL-4) or selected mitogenic stimuli.
- 9509417: While CD23(a) is constitutively expressed in B cells, the expression of CD23(b) is specifically *induced* by interleukin-4 (IL-4) or selected mitogenic stimuli.

  Negative:
- 8218946: However, in contrast to normal B cells, IL-4 did **not** *increase*

CD23 membrane expression on RPMI-8226 cells.

- 1709049: Thirdly, CD19 monoclonal antibody, which inhibits B cell proliferation in response to IL-4 plus anti-Ig, was found to inhibit IL-4-*induced* CD23 but **not** sIgM expression.
- 2136716: Furthermore, IL-4 induces Fc epsilon-receptor (CD23) expression on 30% of unstimulated human B cells, whereas BCAF-containing supernatants from clone P2, that do not contain detectable amounts of IL-4, promote B cell proliferation **without** *inducing* CD23 expression.
- 1388135: An antibody to CD72 (BU40) has been found to mimic interleukin-4 (IL-4) both in its ability to activate resting B cells into the early G1 phase of cell cycle and to augment the expression of major histocompatibility complex (MHC) class II antigen; unlike IL-4, the CD72-clustered antibody **fails** to *induce* the expression of CD23.

2 Positive regulation

Affirmative:

- 2789139: However, a combination of IL4, IL5 and IL6 (with or without IL1) at optimal concentrations could not induce IgE synthesis by purified normal B cells, indicating that cytokine-mediated signals, although essential, are not sufficient for the IL4-dependent *induction* of IgE synthesis.
- 10887336: IL-4 is *important* for B-cell production of IgE, and the human IL-4 receptor alpha chain (hIL-4Ralpha) is crucial for the binding and signal transduction of IL-4, so hIL-4Ralpha may be a candidate gene related to atopy.
- 7722171: In contrast, terminally differentiated, IgE-producing B cells no longer express functional IL-4R because DAB389IL-4 only modestly inhibited ongoing IgE synthesis by B cells from patients with hyper-IgE states and only minimally affected IL-4-*induced* IgE synthesis in normal B cells when the toxin was added at day 7.
- 2172384: We demonstrate here that EBV and IL-4 *induced* the synthesis of IgE by surface IgE-negative B cell precursors isolated by cell sorting.
- 2967330: Like IL-4-containing SUP, rIL-4 also showed the ability to *induce* IgE production in B cells from both atopic and nonatopic donors.

Negative:

- 2172384: IL-4 **failed** to *induce* IgE synthesis in established EBV B cell lines and failed to induce 2.0-kb mature C epsilon transcripts but induced 1.8-kb germ-line C epsilon transcripts.
- 2789139: Recombinant IL4 could *induce* IgE synthesis by peripheral blood mononuclear cells and autologous T/B cell mixtures, but **not** by highly purified B cells.
- 1383379: In contrast to these observations with MNC, IL-4 **failed** to *induce* IgE and IgG4 production by purified B cells.
- 1382870: IgE production was **not** *induced* by IL-4 in purified B cells.
- 1904400: Similarly, DSCG did not enhance IgG2, IgG3 or IgG4 production from sIgG2-, sIgG3- or sIgG4- B cells, respectively, Interleukin-4 (IL-4) or interleukin-6 (IL-6) also *enhanced* Ig production **except** IgG4 from large activated B cells.

3 Positive regulation

Affirmative:

- 6434688: A parallel production of gamma interferon (IFN-gamma) is *induced* by recombinant IL-2 (rIL-2), and NK cells appear to be the major producer cells, whereas T cells are unable to produce IFN-gamma under these experimental conditions.
- 6429853: IL-2 is *required* for the optimum expression of IL-2 receptors on activated T lymphocytes and for maximum synthesis of IFN-gamma in vitro.
- 6434688: A parallel production of gamma interferon (IFN-gamma) is *induced* by recombinant IL-2 (rIL-2), and NK cells appear to be the major producer cells, whereas T cells are unable to produce IFN-gamma under these experimental conditions.
- 15271977: IL-2 is another key immunoregulatory cytokine that is involved in T helper differentiation and is known to *induce* IFN-gamma expression in natural killer (NK) and T cells.
- 2147201: The lymphokines IL-2 and IL-4 promoted the growth of human PHA-triggered T cells, but only IL-2 *induced* the production of IFN-gamma and TNF.

Negative:

- 3086435: **Neither** IL 1 nor IL 2 alone *induced* IFN-gamma production in purified T lymphocyte cultures.
- 1907764: The addition of interleukin-2 (IL-2) or phorbol-12-myristate-13-acetate to T cells stimulated with PHA, IL-1 and IL-6 did **not** *restore* the production of IFN-gamma to an extent comparable to that produced by T cells stimulated in the presence of accessory cells.
- 2147201: IL-6 did **not** influence IFN-gamma or TNF production or T cell proliferation *induced* by PHA-IL-2 and did not modulate IL-1-induced IFN-gamma production.

4 Positive regulation

Affirmative:

- 6292303: These results indicate that whereas classical polyclonal B cell activators (PWM, EBV) fail to induce IgE synthesis by normal B cells, IgE synthesis is readily *induced* by an IgE-specific helper factor released by T cells from patients with hyper-IgE states.
- 3485112: IgE-binding factors from three of four HIE patients *enhanced* IgE synthesis by B cells from patients with perennial allergic rhinitis, or with seasonal allergic rhinitis (SAR) and recent pollen exposure, but did not enhance IgE synthesis by B cells from nonatopic donors or from SAR patients with no recent pollen exposure.
- 2533179: Recombinant 37-kD IgE-BFs *increase* the IL4-induced synthesis of IgE by peripheral blood lymphocytes, as well as the IL4-independent, ongoing synthesis of IgE by either in vivo activated B cells from allergic patients or by in vitro IL4-preactivated B cells.
- 2533179: Recombinant 37-kD IgE-BFs *increase* the IL4-induced synthesis of IgE by peripheral blood lymphocytes, as well as the IL4-independent, ongoing synthesis of IgE by either in vivo activated B cells from allergic patients or by in vitro IL4-preactivated B cells.
- 6237918: These results suggest that a subset of human T cells bearing an Fc epsilon R secretes an IgE-binding glycoprotein which selectively *enhances* IgE synthesis by IgE-bearing B cells.

Negative:

- 3485112: IgE-binding factors from three of four HIE patients enhanced IgE synthesis by B cells from patients with perennial allergic rhinitis, or with seasonal allergic rhinitis (SAR) and recent pollen exposure, but did **not** *enhance* IgE synthesis by B cells from nonatopic donors or from SAR patients with no recent pollen exposure.
- 2976801: These cells secrete IgE binding factors which enhance IgE synthesis but **not** IgG synthesis by preactivated IgE bearing B cells from allergic subjects but not resting B cells from normal donors.

5 Positive regulation

Affirmative:

- 10944420: IFN-gamma alone or combined with LPS *induced* iNOS expression and increased nitrite production in iNOS(+/+) macrophages, but not in iNOS(-/-) macrophages.
- 12417260: UVB light, which is used therapeutically to treat inflammatory dermatosis, was found to suppress IFN-gamma-*induced* expression of NOS2 mRNA and protein, and nitric oxide production in both keratinocytes and macrophages.
- 9753237: IFN-gamma-*induced* iNOS mRNA expression is inhibited by rebamipide in murine macrophage RAW264.7 cells.
- 16883061: JAK inhibitors AG-490 and WHI-P154 decrease IFN-gamma-*induced* iNOS expression and NO production in macrophages.
- 17689680: BACKGROUND: We hypothesized that acetylation of the Stat1 regulates interferon-gamma (IFN-gamma) *mediated* macrophage expression of inducible nitric oxide synthase (iNOS).

Negative:

- 9193655: Murine macrophages possess the capacity to express the inducible NO synthase (iNOS) which is **not** constitutively expressed but *induced* at the transcriptional level by interferon gamma (IFN-gamma) alone or synergistically with LPS.
- 9193655: Murine macrophages possess the capacity to express the inducible NO synthase (iNOS) which is **not** constitutively expressed but *induced* at the transcriptional level by interferon gamma (IFN-gamma) alone or synergistically with LPS.
- 11694524: In contrast to islets, dsRNA + IFN-gamma **fails** to *induce* iNOS expression or nitric oxide production by macrophages isolated from IRF-1(-/-) mice; however, dsRNA + IFN-gamma induces similar levels of IL-1 release by macrophages isolated from both IRF-1(-/-) and IRF-1(+/+) mice.
- 17035338: IFN-gamma *induces* **NO** production, inducible NO synthase (iNOS) protein, and promoter expression in mouse macrophage cells.

1 Regulation

Affirmative:

- 9568685: Whether insulin acutely *regulates* plasma leptin in humans is controversial.
- 9398728: In animal models, insulin and agents that increase intracellular cAMP have been shown to similarly *affect* plasma leptin in vivo.

Negative:

- 8954052: These results suggest that insulin does **not** acutely *regulate* plasma leptin concentrations in humans.

- 11832440: Insulin and pentagastrin did **not** *modify* plasma leptin, whatever HSV status.
- 10856891: Adrenaline, insulin and glucagon do **not** have acute *effects* on plasma leptin levels in sheep: development and characterisation of an ovine leptin ELISA.

2 Regulation

Affirmative:

- 2946861: The autologous mixed lymphocyte reaction (AMLR) and in vitro *influence* of interleukin 1 (IL-1) and interleukin 2 (IL-2) on the AMLR were also studied.
- 3106694: *Role* of interleukin-2 (IL-2) and IL-2 receptor expression in the proliferative defect observed in mitogen-stimulated lymphocytes from patients with gliomas.
- 7493771: Co-operative *effect* between insulin-like growth factor-1 and interleukin-2 on DNA synthesis and interleukin-2 receptor-alpha chain expression in human lymphocytes.
- 8839133: In addition, we studied the proliferative *response* of lymphocytes to mitogens or interleukin-2 (IL-2) alone or in combination with immunomodulating drugs or interleukin-4 (IL-4).

Negative:

- 2874115: The culture supernatant from these cell lines showed no IL-2 activity toward Con-A-stimulated human peripheral blood lymphocytes, and their growth was **not** *affected* by additional IL-2 in cultures.
- 2474592: Normal lymphocytes stimulated with Df expressed Tac antigen (low-affinity IL-2 receptor) but, in contrast to the patients' lymphocytes, did not absorb **nor** *respond* to IL-2.

3 Regulation

Affirmative:

- 10828498: The aim of this work was to study the *effect* of centrally applied ANF or CNP on plasma ANF.
- 10841438: The current project sought to study the *effect* of the NOS inhibitor Nomega-nitro-L-arginine methyl ester (L-NAME, 10 microg x min(-1), sc for 7 days) on plasma volume, plasma atrial natriuretic factor (ANF), plasma endothelin-1 (ET), and plasma renin activity (PRA) during gestation in conscious rats.
- 2960617: Infusion of ANF at doses expected to *change* plasma ANF levels minimally decreased arterial pressure in hypertensive rats over 7 days.
- 9112384: Neither ANP nor CNP infusion had any *effect* on plasma IR-NT-ANP levels under basal conditions.

Negative:

- 7848625: YT-146 had **no** *effect* on plasma renin activity (PRA), plasma aldosterone, vasopressin (ADH), and atrial natriuretic peptide (ANP) in the acute study.
- 1826031: **Neither** SQ 28,603 nor C-ANF(4-23) *affected* MAP or plasma ANF in the normotensive rats.

4 Regulation

Affirmative:

- 1636698: Proinsulin had a significantly weaker *effect* than insulin, at the

lowest infusion dose, in percent suppression of plasma nonesterified fatty acids, blood glycerol, and beta-hydroxybutyrate levels (all P less than 0.05).

- 11194244: RESULTS: Insulin aspart and buffered regular human insulin were both *effective* in controlling average daily blood glucose levels (8.2 +/- 1.9 and 8.5 +/- 2.1 mmol/l, respectively) (mean +/- SD) and maintaining serum fructosamine (343 +/- 25.7 and 336 +/- 27.4 micromol/l) and HbA1c (6.9 +/- 0.6 and 7.1 +/- 0.6%) levels.
- 2692943: The *effect* of human biosynthetic proinsulin (PRO) on blood glucose (BG) control and glucose excursions was studied in a nonrandomized design in eight patients with unstable type 1 diabetes mellitus and compared with that of human NPH insulin.
- PMC2993798: A reduced ability of insulin to activate glucose transport into cells, i.e., insulin resistance, is marked by high glucose and high insulin levels in circulating blood, and aberrant insulin-*regulated* gene functions [10].

  Negative:
- 10078556: Blood glucose and serum insulin levels were **not** *affected* by intranasal insulin.

5 Regulation

Affirmative:

- 16845238: Basal-prandial insulin regimens that use a long-acting insulin analogue to *control* the fasting plasma glucose level and a short-acting insulin analogue for post-meal glucose excursions replace insulin in a manner that most closely approximates normal physiologic patterns.
- 4351804: Various lines of evidence indicate that the insulin receptor on the plasma membrane, in addition to the insulin coupled to the agarose, was *responsible* for the observed binding.
- 1100459: *Effect* of intracisternal insulin on plasma glucose and insulin in the dog.
- 7522843: Circulating insulin-like growth factor II/mannose-6-phosphate receptor and insulin-like growth factor binding proteins in fetal sheep plasma are *regulated* by glucose and insulin.

  Negative:
- 6999134: Concentration of plasma insulin was elevated 15 minutes following the 6 mU insulin treatment for the concentrate and forage ration, while 1 mU insulin did **not** *affect* plasma insulin.
- 668977: Glucose and insulin injections given during lethargy showed **no** *effects* on plasma insulin and glucose respectively but confirmed a very low plasma clearance of glucose and insulin.
- 1874929: Although ambient plasma TG and FFA concentrations fell significantly, plasma glucose, insulin, HGP, concentrations fell significantly, plasma glucose, insulin, HGP, and glucose MCR did **not** *change*.
- 16554011: High nutrient intake resulted in significantly elevated maternal plasma concentrations of insulin, leptin, prolactin and glucose, **no** significant *changes* in fetal insulin, leptin or non-esterified fatty acids and attenuated fetal prolactin concentrations.

1 Negative regulation

Affirmative:

- 1328039: The molecular mechanisms by which human interleukin-4 (IL-4) *down-regulates* tumour necrosis factor-alpha (TNF-alpha) production by monocytes remain unknown.
- 1328039: The molecular mechanisms by which human interleukin-4 (IL-4) *down-regulates* tumour necrosis factor-alpha (TNF-alpha) production by monocytes remain unknown.
- 1328039: IL-4 *reduced* TNF-alpha production by monocytes when IL-4 and lipopolysaccharide (LPS) were added concomitantly, or upon subsequent activation by LPS 16 hr after first exposure to IL-4.
- 2785566: IL-4 *down-regulates* IL-1 and TNF gene expression in human monocytes.
- 11991671: With addition of IFNalpha-neutralizing antibodies, the ability of IL-4 to *suppress* LPS-induced TNFalpha production with prolonged monocyte culture was restored.

Negative:

- 9558115: Studies with Abs to gammac and an IL-4 mutant that is unable to bind to gammac showed that IL-4 can *suppress* IL-1beta but **not** TNF-alpha production by LPS-stimulated monocytes in the presence of little or no functioning gammac.
- 11991671: Like MDMac, interferon alpha (IFNalpha)-treated monocytes expressed less IL-4 receptor gamma c chain, reduced levels of IL-4-activated STAT6 and IL-4 could **not** *suppress* LPS-induced TNFalpha production.
- 7690805: In contrast to the response by blood monocytes, the response to IL-4 by synovial fluid cells was selective; IL-4 did **not** significantly *suppress* LPS-induced TNF-alpha production, but decreased CD14 expression to a similar extent in the two cell populations.

2 Negative regulation

Affirmative:

- 2754338: When added to murine adipocytes in culture, tumor necrosis factor (TNF) *decreases* the levels of lipoprotein lipase (LPL).
- 2754338: When added to murine adipocytes in culture, tumor necrosis factor (TNF) *decreases* the levels of lipoprotein lipase (LPL).
- 2754338: When added to murine adipocytes in culture, tumor necrosis factor (TNF) *decreases* the levels of lipoprotein lipase (LPL).
- 2754338: When added to murine adipocytes in culture, tumor necrosis factor (TNF) *decreases* the levels of lipoprotein lipase (LPL).
- 12526099: TNF also *suppressed* the lipoprotein lipase (LPL) activity of 3T3-L1 adipocytes.

Negative:

- 2198021: TNF did **not** *decrease* LPL activity in isolated adipocytes.

3 Negative regulation

Affirmative:

- 12765949: Analysis using phospho-specific antibodies revealed that insulin *decreases* phosphorylation of sites 3a + 3b in human muscle, and this was accompanied by activation of Akt and inhibition of glycogen synthase kinase-3alpha.

Negative:

- 12765949: Insulin did **not** *decrease* phosphorylation of sites 2 + 2a in healthy human muscle, whereas in diabetic muscle insulin infusion in fact caused a marked increase in the phosphorylation of sites 2 + 2a.

4 Negative regulation

Affirmative:

- 15337697: Although serum levels of MIP-1alpha were not suppressed by TRX-1 until day 21, both an in vitro chemotaxis chamber assay and an in vivo air pouch model showed that TRX-1 significantly *suppressed* MIP-1alpha- or MIP-2-induced leukocyte chemotaxis.

Negative:

- 15337697: Although serum levels of MIP-1alpha were **not** *suppressed* by TRX-1 until day 21, both an in vitro chemotaxis chamber assay and an in vivo air pouch model showed that TRX-1 significantly suppressed MIP-1alpha- or MIP-2-induced leukocyte chemotaxis.

5 Negative regulation

Affirmative:

- 11673527: IFN-gamma- and IFN-beta-mediated *inhibition* of MMP-9 gene expression is dependent on the transcription factor STAT-1alpha, since IFN-gamma and IFN-beta fail to suppress MMP-9 expression in STAT-1alpha-deficient primary astrocytes and human fibrosarcoma cells.

Negative:

- 11673527: IFN-gamma- and IFN-beta-mediated inhibition of MMP-9 gene expression is dependent on the transcription factor STAT-1alpha, since IFN-gamma and IFN-beta **fail** to *suppress* MMP-9 expression in STAT-1alpha-deficient primary astrocytes and human fibrosarcoma cells.
- 11673527: IFN-gamma- and IFN-beta-mediated inhibition of MMP-9 gene expression is dependent on the transcription factor STAT-1alpha, since IFN-gamma and IFN-beta **fail** to *suppress* MMP-9 expression in STAT-1alpha-deficient primary astrocytes and human fibrosarcoma cells.

Appendix D
BioContext data and code availability

We provide the data produced as the output of BioContext, as well as all the intermediary data freely available. It is accessible on the web, and also available through the supplementary materials of this thesis available at www.cs.man.ac.uk/~sarafraf/thesis-supplementary.html. All the code written for the BioContext, the wrappers for several tools, and the tools that we developed will be available at http://www.biocontext.org/ .

Here we list the data and code from each stage. For more details about the size of each data set, see Chapter 6 .

NER

The gene, protein, and protein complexes, species, and anatomy NER from GeneTUKit, GNAT, BANNER, and LINNAEUS, provided for MEDLINE and the open access part of PMC.

A majority of these named entities are normalised to their database identifiers. We also provide the union and intersection sets on the outputs of the above tools.

Parses

The dependency parse trees and constituency parse trees of all the sentences in MEDLINE and open access PMC, computed by McClosky parser, GDep, and Enju.

Events

Biomedical events extracted by TEES and EventMiner, and the union and intersection sets of their outputs.

Context extractors

The python implementation of Negmole is standalone and can be downloaded from the Google Code project page (http://code.google.com/p/negmole/). The anatomical association is a part of the GETM system and is available on SourceForge (http://getm-project.sourceforge.net/).

The results of context association to the events are provided in tables, with rows referencing other tables. To effectively use this data, large database joins are required. Alternatively, the denormalised table can be used.

Denormalised event data

The denormalised table, containing stand-alone event information including every mention of the text mined events. Each row contains complete event information as described in 5.2.1 including the normalised named entity references, negation and speculation information, anatomical location, confidence, HTML formatted sentences highlighting the event attributes, etc. The filtered events do not appear in this table.

Collapsed event data

This data contains distinct event mentions, ignoring the repeated mentions across the literature. It also provides the number of mentions for each distinct event, and the sum of their document-level confidences. The document-level confidence is the maximum confidence of a specific event in a document.

Conflicting pairs

This data contains pairs of contrasting hashes. The two hashes in a conflicting pair represent two distinct events that are contrasting in a strict sense: They are asserted (not speculated), they have complete context extracted, and they are similar in all contextual attributes except for negation.

References

Agarwal, S. & Yu, H., 2010. Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Informatics Association*, 17(6), pp.696–701.

Aho, A.V. & Ullman, J.D., 1972. *The theory of parsing, translation, and compiling*, Upper Saddle River, NJ, USA: Prentice-Hall, Inc. Available at: http://portal.acm.org/citation.cfm?id=578789.

Albert, S., Gaudan, S., Knigge, H., Raetsch, A., Delgado, A., Huhse, B., Kirsch, H., Albers, M., Rebholz-Schuhmann, D. & Koegl, M., 2003. Computer-Assisted Generation of a Protein-Interaction Database for Nuclear Receptors. *Mol Endocrinol*, 17(8), pp.1555–1567.

Ananiadou, S. & Mcnaught, J. eds., 2005. *Text Mining for Biology And Biomedicine*, Artech House Publishers. Available at: http://www.worldcat.org/isbn/158053984X.

Ananiadou, S. & Nenadic, G., 2006. Automatic Terminology Management in Biomedicine. In S. Ananiadou & J. McNaught, eds. *Text Mining for Biology and Biomedicine*. 685 Canton Street Norwood MA 02062: Artech House, Inc.

Ballesteros, M., Francisco, V., Díaz, A., Herrera, J. & Gervás, P., 2011. *Inferring the Scope of Negation and Speculation Via Dependency Analysis*,

Barker, C. & Pullum, G.K., 1990. A Theory of Command Relations. *Linguistics and Philosophy*, 13(1), pp.1–34.

Béchet, F., 2011. Named Entity Recognition. In G. Tur & R. De Mori, eds. *Spoken Language Understanding: Systems for Extracting Semantic Information*. John Wiley & Sons, Ltd, pp. 257–290. Available at: http://dx.doi.org/10.1002/9781119992691.ch10.

Bikel, D.M., 2004. A Distributional Analysis of a Lexicalized Statistical Parsing Model. In D. Lin & D. Wu, eds. *Proceedings of EMNLP 2004*. Barcelona, Spain: Association for Computational Linguistics, pp. 182–189.

Björne, J., Ginter, F., Pyysalo, S., Tsujii, J. & Salakoski, T., 2010. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12), pp.i382-90.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T. & Salakoski, T., 2009. Extracting Complex Biological Events with Rich Graph-Based Feature

Sets. In *Proceedings of the BioNLP'09 Shared Task on Event Extraction*. pp. 10–18.

Burges, C.J.C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2, pp.121–167.

Butler, P., 2009. Publication Bias in Clinical Trials due to Statistical Significance or Direction of Trial Results: RHL commentary. *The WHO Reproductive Health Library*.

Ceusters, W., Elkin, P. & Smith, B., 2007. Negative findings in electronic health records and biomedical ontologies: a realist approach. *International journal of medical informatics*, 76 Suppl 3, p.S326–S333.

Chapman, N.P., 1988. *LR Parsing: Theory and Practice (Cambridge Studies in Cultural Systems)*, Cambridge University Press. Available at: http://www.amazon.com/exe%bidos/redirect?tag=citeulike07-20&path=ASIN/052130413X.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F. & Buchanan, B.G., 2001a. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5), pp.301–310.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F. & Buchanan, B.G., 2001b. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp*, pp.105–109.

Chen, L., Liu, H. & Friedman, C., 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2), pp.248–256.

Chiang, J.-H. & Yu, H.-C., 2003. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19(11), pp.1417–1422.

Cohen, A.M. & Hersh, W.R., 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1), pp.57–71.

Cohen, K.B. & Hunter, L., 2008. Getting Started in Text Mining. *PLoS Comput Biol*, 4(1), p.e20+.

Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C. & Hunter, L.E., 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11, p.492.

Collins, M.J., 1999. *Head-driven statistical models for natural language parsing*. Available at: http://repository.upenn.edu/dissertations/AAI9926110/.

Culotta, A. & Sorensen, J., 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. ACL '04. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 423+. Available at:

http://dx.doi.org/10.3115/1218955.1219009.

Cunningham, H., Maynard, D. & Bontcheva, K., 2011. *Processing with GATE*, University of Sheffield Department of Computer Science.

Dean, J. & Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51(1), pp.107–113.

Dickersin, K., Chan, S., Chalmers, T.C., Sacks, H.S. & Smith, H., 1987. Publication bias and clinical trials. *Controlled clinical trials*, 8(4), pp.343–353.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T. & Hogue, C.W., 2003. PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4, p.11.

Duchier, D., 1999. Axiomatizing dependency parsing using set constraints. In *6th Meeting on Mathematics of Language*. Orlando/FL.

Duchier, D., 2000. Constraint Programming for Natural Language Processing.

Easterbrook, P.J., Gopalan, R., Berlin, J.A. & Matthews, D.R., 1991. Publication bias in clinical research. *The Lancet*, 337(8746), pp.867–872.

Elkin, P.L., Brown, S.H., Bauer, B.A., Husser, C.S., Carruth, W., Bergstrom, L.R. & Wahner-Roedler, D.L., 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1), p.13+.

Emms, M., 2008. Tree Distance and Some Other Variants of Evalb. In D. T. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, ed. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Farkas, R., Vincze, V., Móra, G., Csirik, J. & Szarvas, G., 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Uppsala, Sweden: Association for Computational Linguistics, pp. 1–12. Available at: http://www.aclweb.org/anthology-new/W/W10/W10-3001.bib.

Fellbaum, C., 1998. *WordNet: An Electronic Lexical Database*, Bradford Books.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N. & Welty, C., 2010. Building Watson: An Overview of the DeepQA Project | Ferrucci | AI Magazine. *AI MAGAZINE*, 31(3), pp.59–79.

Fleischhacker, D., 2011. Enriching Ontologies by Learned Negation The Semanic

Web: Research and Applications. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, J. Pan, G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. Leenheer, & J. Pan, eds. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin / Heidelberg, pp. 508–512. Available at: http://dx.doi.org/10.1007/978-3-642-21064-8_44.

Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A., 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(suppl 1), p.S74–S82.

Fu, W., Sanders-Beer, B.E., Katz, K.S., Maglott, D.R., Pruitt, K.D. & Ptak, R.G., 2009. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic acids research*, 37(Database issue), p.gkn708+.

Georgescul, M., 2010. A hedgehop over a max-margin framework using hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*. CoNLL '10: Shared Task. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 26–31. Available at: http://dl.acm.org/citation.cfm?id=1870535.1870539.

Gerner, M., 2011. *Integrating text mining approaches to identify entities and extract events from the biomedical literature*. Faculty of Life Sciences, University of Manchester.

Gerner, M., Nenadic, G. & Bergman, C.M., 2010a. An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In *Proceedings of the BioNLP workshop*.

Gerner, M., Nenadic, G. & Bergman, C.M., 2010b. LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11, p.85.

Gindl, S., Kaiser, K. & Miksch, S., 2008. Syntactical negation detection in clinical practice guidelines. *Studies in health technology and informatics*, 136, pp.187–192.

Hagege, C., 2011. Linguistically Motivated Negation Processing: an Application for the Detection of Risk Indicators in Unstructured Discharge Summaries. *Proceedings of the CICLing 2011*, 43, pp.101–106.

Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., Gonzalez, G., Nenadic, G. & Bergman, C.M., 2011. The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27(19), pp.2769-71.

Hakenberg, J., Plake, C., Royer, L., Strobelt, H., Leser, U. & Schroeder, M., 2008. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*, 9(Suppl 2), p.S14.

Hakenberg, J., Solt, I., Tikk, D., Tari, L., Rheinländer, A., Ngyuen, Q.L., Gonzalez,

G. & Leser, U., 2009. Molecular event extraction from link grammar parse trees. In *BioNLP '09: Proceedings of the Workshop on BioNLP*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 86–94. Available at: http://portal.acm.org/citation.cfm?id=1572353.

Hara, T., Miyao, Y. & Tsujii, J., 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In R. Dale, K.-F. Wong, J. Su, & O. Y. Kwong, eds. *Natural Language Processing – IJCNLP 2005*. Lecture Notes in Artificial Intelligence. Jeju Island, Korea: Springer-Verlag, pp. 199–210.

Hearst, M., 2003. What Is Text Mining? Available at: http://people.ischool.berkeley.edu/~hearst/text-mining.html.

Huang, M., Liu, J. & Zhu, X., 2011. GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, 27(7), pp.1032-3.

Huang, Y. & Lowe, H.J., 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3), pp.304–311.

Hunter, L. & Cohen, B.K., 2006. Biomedical Language Processing: Perspective What's Beyond PubMed? *Molecular cell*, 21(5), pp.589–594.

Jaeger, S., Gaudan, S., Leser, U. & Rebholz-Schuhmann, D., 2008. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics*, 9(Suppl 8), p.S2.

Joachims, T., 1999. Making large-scale support vector machine learning practical. In Cambridge, MA, USA: MIT Press, pp. 169–184. Available at: http://dl.acm.org/citation.cfm?id=299094.299104.

Joachims, T., 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '06. New York, NY, USA: ACM, pp. 217–226. Available at: http://doi.acm.org/10.1145/1150402.1150429.

Kilicoglu, H. & Bergler, S., 2009. Syntactic dependency based heuristics for biological event extraction. In *BioNLP '09: Proceedings of the Workshop on BioNLP*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 119–127. Available at: http://portal.acm.org/citation.cfm?id=1572340.1572361.

Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y. & Tsujii, J., 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. BioNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–9. Available at: http://portal.acm.org/citation.cfm?id=1572342.

Kim, J.-D., Ohta, T. & Tsujii, J., 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1), p.10.

Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y. & Collier, N., 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. JNLPBA '04. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 70–75. Available at: http://portal.acm.org/citation.cfm?id=1567610.

Kim, J.-jae, Zhang, Z., Park, J.C. & Ng, S.-K., 2006. BioContrasts: extracting and exploiting proteinâ protein contrastive relations from biomedical literature. *Bioinformatics*, 22(5), pp.597–605.

Klima, E., 1964. Negation in English. In J. Fodor & J. J. Katz, eds. *The Structure of Language*. Englewood Cliffs: Prentice-Hall, pp. 246–323.

Krallinger, M., Vazquez, M., Leitner, F. & Valencia, A., 2010. Results of the BioCreative III (interaction) article classification task. In *Proceedings of the BioCreative III*. pp. 17–23.

Lafferty, J., Mccallum, A. & Pereira, F., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann.

Lakoff, G., 1973. Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, 2(4), pp.458–508.

Van Landeghem, S., Ginter, F., Van de Peer, Y. & Salakoski, T., 2011. EVEX: A PubMed-scale resource for homology-based generalization of text mining predictions. In *BioNLP 2011*. pp. 28-37.

Van Landeghem, S., Saeys, Y., De Baets, B. & Van de Peer, Y., 2008. Extracting Protein-Protein Interactions from Text using Rich Feature Vectors and Feature Selection. In T. Salakoski, D. R. Schuhmann, & S. Pyysalo, eds. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*. Turku Centre for Computer Science (TUCS), pp. 77–84.

Langacker, R., 1969. On Pronominalization and the Chain of Command. In D. Reibel & S. Schane, eds. *Modern Studies in English*. Englewood, Cliffs, NJ: Prentice-Hall, pp. 160-186.

Lasnik, H., 1976. Remarks on Coreference. *Linguistic Analysis*, 2(1), pp.1–22.

Lawler, J., 2010. Negation and Negative Polarity. In P. Colm Hohgan, ed. *The Cambridge Encyclopedia of the Language Sciences*. Cambridge, UK: Cambridge University Press.

Leaman, R. & Gonzalez, G., 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp.652–663.

Leroy, G., Chen, H. & Martinez, J.D., 2003. A shallow parser based on closed-

class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, pp.145–158.

Light, M., Qiu, X.Y. & Srinivasan, P., 2004. The Language of Bioscience: Facts, Speculations, and Statements in Between. In BIOLINK 2004. Available at: http://que.info-science.uiowa.edu/~light/research/mypapers/lightHLTworkshop.pdf.

Lindsay, R.K. & Gordon, M.D., 1999. Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci.*, 50(7), pp.574–587.

Lovins, J.B., 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11, pp.22-31.

Lu, Z. & Wilbur, W.J., 2010. Overview of BioCreative III gene mention normalization. In *BioCreative III*.

MacKinlay, A., Martinez, D. & Baldwin, T., 2009. Biomedical event annotation with CRFs and precision grammars. In *BioNLP '09: Proceedings of the Workshop on BioNLP*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 77–85. Available at: http://portal.acm.org/citation.cfm?id=1572340.1572351.

Manning, C.D., Raghavan, P. & Schütze, H., 2008. *Introduction to Information Retrieval* 1st ed., Cambridge University Press. Available at: http://www.amazon.com/exe%bidos/redirect?tag=citeulike07-20&path=ASIN/0521865719.

Manning, C.D. & Schuetze, H., 1999. *Foundations of Statistical Natural Language Processing* 1st ed., The MIT Press. Available at: http://www.amazon.com/exe%bidos/redirect?tag=citeulike07-20&path=ASIN/0262133601.

McCawley, J.D., 1993. *Everything that Linguists have Always Wanted to Know about Logic . . . But Were Ashamed to Ask* 2nd ed., 1427 E. 60th Street Chicago, IL 60637 USA: University of Chicago Press.

McClosky, D., Charniak, E. & Johnson, M., 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 28–36. Available at: http://portal.acm.org/citation.cfm?id=1858003.

McClosky, D., Charniak, E. & Johnson, M., 2006. Effective Self-Training for Parsing. In *HLT-NAACL*.

McClosky, D., Surdeanu, M. & Manning, C.D., 2011. Event Extraction as Dependency Parsing. In *Proceedings of the Association for Computational Linguistics - Human Language Technologies 2011 Conference (ACL-HLT 2011)*.

Medlock, B. & Briscoe, T., 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.6724.

Memisevic, R., 2007. Monte - machine learning in Python. Available at: http://montepython.sourceforge.net/ [Accessed October 1, 2011].

Mendenhall, W., Beaver, R. J., & Beaver, B. M., 2008. *Introduction to Probability and Statistics*, Duxbury Press, 13 edition.

Mitsumori, T., Murata, M., Fukuda, Y., Doi, K. & Doi, H., 2006. Extracting Protein-Protein Interaction Information from Biomedical Text with SVM. *IEICE - Trans. Inf. Syst.*, E89-D(8), pp.2464–2466.

Miwa, M., Sætre, R., Kim, J.D., & Tsujii, J., 2010. Event extraction with complex event classification using rich features. *Journal of Bioinformatics and Computational Biology,* 8, pp.131-146.

Miyao, Y., Sætre, R., Sagae, K., Matsuzaki, T. & Tsujii, J., 2008. Task-oriented Evaluation of Syntactic Parsers and Their Representations. *Proceedings of ACL08 HLT*, (June), pp.46–54.

Morante, R., Van Asch, V. & Daelemans, W., 2010a. Memory-Based resolution of in-sentence scopes of hedge cues. In *Fourteenth Conference on Computational Natural Language Learning: Shared Task*. Uppsala, Sweden: Association for Computational Linguistics. Available at: http://aclweb.org/anthology-new/W/W10/W10-3006.pdf.

Morante, R. & Daelemans, W., 2009. A metalearning approach to processing the scope of negation. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 21–29. Available at: http://portal.acm.org/citation.cfm?id=1596381.

Morante, R., Schrauwen, S. & Daelemans, W., 2011. Corpus-based approaches to processing the scope of negation cues: an evaluation of the state of the art. *Proc of IWCS 2011*, (Section 4), pp.350–354.

Morante, R. & Sporleder, C. eds., 2010b. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden: University of Antwerp. Available at: http://www.aclweb.org/anthology/W10-3100.

Morgan, A.A., Wellner, B., Colombe, J.B., Arens, R., Colosimo, M.E. & Hirschman, L., 2007. Evaluating the automatic mapping of human gene and protein mentions to unique identifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp.281–291.

Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H., Torres, R., Krauthammer, M., Lau, W., Liu, H., Hsu, C., Schuemie, M., Cohen, K. & Hirschman, L., 2008.

Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2), p.S3.

Mutalik, P.G., Deshpande, A. & Nadkarni, P.M., 2001. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the American Medical Informatics Association : JAMIA*, 8(6), pp.598–609.

Nivre, J., 2009. Dependency Grammar and Dependency Parsing. In S. Kübler, R. McDonald, & J. Nivre, eds. Morgan & Claypool Publishers. Available at: http://www.morganclaypool.com/doi/abs/10.2200/S00169ED1V01Y200901HLT002.

Ohta, T., Kim, J.-D. & Tsujii, J., 2007. *GENIA Guidelines for event annotation*,

Ohta, T., Tateisi, Y. & Kim, J.D., 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 82–86. Available at: http://portal.acm.org/citation.cfm?id=1289260.

Özgür, A. & Radev, D.R., 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1398–1407. Available at: http://portal.acm.org/citation.cfm?id=1699686.

Patrick, J., Wang, Y. & Budd, P., 2007. An automated system for conversion of clinical notes into SNOMED clinical terminology. In *Proceedings of the fifth Australasian symposium on ACSW frontiers - Volume 68*. ACSW '07. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., pp. 219–226. Available at: http://portal.acm.org/citation.cfm?id=1274559.

Penagos, C.R., Salgado, H., Flores, I.M. & Vides, J.C., 2007. Automatic reconstruction of a bacterial regulatory network using Natural Language Processing. *BMC Bioinformatics*, 8(1), p.293+.

Porter, M.F., 1980. An algorithm for suffix stripping. *Program*, 14(3), pp.130–137.

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J. & Salakoski, T., 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8, p.50.

Raychaudhuri, S., Schütze, H. & Altman, R.B., 2002. Using Text Analysis to Identify Functionally Coherent Gene Groups — Genome Research. *Genome Research*, 12, pp.1582–1590.

Rinaldi, F., Dowdall, J., Hess, M., Kaljurand, K., Koit, M., Vider, K. & Kahusk, N., 2002. Terminology as Knowledge in Answer Extraction. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering,* pp. 107–112.

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M. & Romacker, M., 2006. An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3), p.S3+.

Sagae, K., Miyao, Y. & Tsujii, J., 2007a. HPSG Parsing with Shallow Dependency Constraints. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 624–631. Available at: http://www.aclweb.org/anthology/P/P07/P07-1079.

Sagae, K. & Tsujii, J., 2007b. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Prague, Czech Republic: Association for Computational Linguistics, pp. 1044–1050. Available at: http://www.aclweb.org/anthology-new/D/D07/D07-1111.bib.

Sanchez, O., 2007. *Text mining applied to biological texts: beyond the extraction of protein-protein interactions*. Department of Computing and Electronic Systems, University of Essex.

Sarafraz, F. & Nenadic, G., 2010. Using SVMs with the Command Relation Features to Identify Negated Events in Biomedical Literature. In *The Workshop on Negation and Speculation in Natural Language Processing*.

Shatkay, H., Pan, F., Rzhetsky, A. & Wilbur, W.J., 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18), pp.2086–2093.

Solt, I., Gerner, M., Thomas, P., Nenadic, G., Bergman, C.M., Leser, U. & Hakenberg, J., 2010. Gene mention normalization in full texts using GNAT and LINNAEUS. In *Proceedings of the BioCreative III Workshop*.

Spasic, I., Sarafraz, F., Keane, J.A. & Nenadic, G., 2010. Medication information extraction with linguistic pattern matching and semantic rules. *J Am Med Inform Assoc*, 17(5), pp.532-5.

Swaminathan, R., Sharma, A. & Yang, H., 2010. Opinion Mining for Biomedical Text Data: Feature Space Design and Feature Selection. In *the Nineth International Workshop on Data Mining in Bioinformatics BIOKDD 2010*. Available at: http://cs.sfsu.edu/ huiyang/publications.htm.

Szarvas, G., Vincze, V., Farkas, R. & Csirik, J., 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *BioNLP '08: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 38–45. Available at: http://portal.acm.org/citation.cfm?id=1572306.1572314.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J. & von Mering, C., 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database

issue), pp.D561-8.

Tang, B., Wang, X., Wang, X., Yuan, B. & Fan, S., 2010. A Cascade Method for Detecting Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Uppsala, Sweden: Association for Computational Linguistics, pp. 13–17. Available at: http://www.aclweb.org/anthology/W10-3002.

Tolentino, H., Matters, M., Walop, W., Law, B., Tong, W., Liu, F., Fontelo, P., Kohl, K. & Payne, D., 2006. Concept negation in free text components of vaccine safety reports. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*. Available at: http://view.ncbi.nlm.nih.gov/pubmed/17238741.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J., 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In P. Bozanis & E. N. Houstis, eds. *Advances in Informatics*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 382–392. Available at: http://dx.doi.org/10.1007/11573036_36.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R. & Zhang, H., 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*, 37(Database issue), pp.D555-9.

Uzuner, O., 2008. Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp.1252–1253.

Velldal, E., 2011. Predicting speculation: a simple disambiguation approach to hedge detection in biomedical literature. *Journal of Biomedical Semantics*, 2(Suppl 5), p.S7.

Vincze, V., Szarvas, G., Mora, G., Ohta, T. & Farkas, R., 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5), p.S8+.

Yakushiji, A., Miyao, Y., Ohta, T., Tateisi, Y. & Tsujii, J., 2006. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 284–292. Available at: http://portal.acm.org/citation.cfm?id=1610116.

Yang, H., Nenadic, G. & Keane, J.A., 2008. Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics*, 9(Suppl 3):S11. Available at: http://biocreative.sourceforge.net/w_song.pdf.

Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A. & Wilbur, J.W., 2002.

Automatic extraction of gene and protein synonyms from medline and journal articles. In *Proc. AMIA Symp*. pp. 919–923. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.6455.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K.B., 2007. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5), pp.358–375.