Data-driven Temporal Information Extraction with Applications in General and Clinical Domains

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2015

By
Michele Filannino
School of Computer Science

Contents

Word count: 35031

List of Figures

List of Tables

Acronyms

CRF Conditional Random Field. 30, 49, 50, 55, 56, 61, 65, 97, 101–105, 107–111, 149, 151, 157, 179, 182–184

CS Computer Science. 28, 163

DA Discourse Analysis. 42

DCT Document Creation Time. 45, 63, 82, 162

EE Event Extraction. 25

EHR Electronic Health Record. 95

EPSRC Engineering and Physical Sciences Research Council. 113

HeRC Health eResearch Centre. 113

HMM Hidden Markov Model. 30

i2b2 Informatics for Integrating Biology & the Bedside. 44, 113

IAA Inter Annotator Agreement. 46, 152

IE Information Extraction. 16, 27, 29, 146, 156

IR Information Retrieval. 44

ML Machine Learning. 20–23, 25, 29, 30, 34, 46–48, 54, 55, 57, 97, 128, 147, 151–153, 157, 158, 160, 166

MLN Markov Logic Network. 65

MRC Medical Research Council. 113

NER Named Entity Recognition. 22, 68, 97, 99, 148

NHLBI National Heart, Lung, and Blood Institute. 113

NIH National Institutes of Health. 113

NLM National Library of Medicine. 113

NLP Natural Language Processing. 24, 28, 29, 34, 35, 42, 48, 49, 61, 128, 148, 156

NTCIR NII Testbeds and Community for Information access Research. 44

POS Part-of-speech. 54, 99, 101

RF Random Forest. 151

SRL Semantic Role Labelling. 42

STAG Sheffield Temporal Annotation Guidelines. 39

SVM Support Vector Machine. 49, 65, 151

TE Temporal Expression. 95–97, 99, 100, 102–105, 107, 109–112

TEE Temporal Expression Extraction. 19, 23, 25, 42, 46, 48, 49, 52, 54, 57, 147–149, 158–160

TERN Temporal Expression Recognition and Normalisation. 39

TIE Temporal Information Extraction. 5, 6, 19–25, 28, 36, 43–47, 52, 54, 56, 128, 146–160

TM Text Mining. 6, 16, 24, 25, 27–30, 36

UMLS Unified Medical Language System. 95, 101

WWW World Wide Web. 15

XML eXtensible Markup Language. 17

Abstract

The automatic extraction of temporal information from written texts is pivotal for many Natural Language Processing applications such as question answering, text summarisation and information retrieval. However, Temporal Information Extraction (TIE) is a challenging task because of the amount of types of expressions (durations, frequencies, times, dates) and their high morphological variability and ambiguity. As far as the approaches are concerned, the most common among the existing ones is rule-based, while data-driven ones are under-explored.

This thesis introduces a novel domain-independent data-driven TIE strategy. The identification strategy is based on machine learning sequence labelling classifiers on features selected through an extensive exploration. Results are further optimised using an *a posteriori* label-adjustment pipeline. The normalisation strategy is rule-based and builds on a pre-existing system.

The methodology has been applied to both specific (clinical) and generic domain, and has been officially benchmarked at the i2b2/2012 and TempEval-3 challenges, ranking respectively 3rd and 1st. The results prove the TIE task to be more challenging in the clinical domain (overall accuracy 63%) rather than in the general domain (overall accuracy 69%).

Finally, this thesis also presents two applications of TIE. One of them introduces the concept of *temporal footprint* of a Wikipedia article, and uses it to mine the life span of persons. In the other case, TIE techniques are used to improve pre-existing information retrieval systems by filtering out temporally irrelevant results.

Lay abstract

The human brain has evolved to master, among the others, the ability of dealing with time. People are naturally able to interpret the temporal side of speech or text, and use this knowledge to work out new insights and discoveries. Making computers mimicking such capability has become imperative to deal with the information overload.

Automatic temporal information analysis is a challenging task in Text Mining (TM). This analysis makes knowledge extraction faster in different orders of magnitude and it enhances the exhibited intelligence of pre-existing natural language-based systems.

This thesis presents a data-driven TIE methodology which improves the state-of-the-art performance on two types of data: general and clinical. In the clinical domain TIE has proved to be crucial because of its applications, for example summarisation, visualisation of patients' clinical pathways, disease progression modelling and analysis of the effectiveness of treatments, to mention a few.

Novel applications of TIE systems are also presented. In one case, by temporally analysing people's Wikipedia pages, it is now possible to predict their life span on the time-line. In the other case, the temporal analysis has been shown to improve information retrieval systems by filtering out documents which are not temporally relevant according to the users' queries.

Declaration

No portion of the work referred to in this 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

Policy[1], in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations [2] and in The University's policy on presentation of Theses.

[1] see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487`
[2] see `http://www.manchester.ac.uk/library/aboutus/regulations`

Acknowledgements

In primis, I would like to thank my supervisors, Goran and Gavin, for having been constant and authoritative points of reference during these four years. More specifically, my biggest *hvala* goes to Goran, by whom I have been mostly supervised. I timidly started my Ph.D., full of doubts about my fit for it; I am now approaching its end with confidence and an already tremendous sense of nostalgia. I guess it's because of you two. Thanks G&G!

To Barry Cheetam for his advice and paternal scientific figure. To have been a constant and vivid representation of the kind of elegance I aspire to reach at some point in my life.

To Bijan Parsia, Ulrike Sattler and John Keane for their faith in my skills and for having openly shown that to me.

To John Latham and Gavin Brown for having shown me what teaching really looks like. Their commitment and enthusiasm towards the young generations are things I will always carry with me.

To my CDT friends with whom I shared the burden of being the first Centre for Doctoral Training cohort. May the force be always with you guys.

To my research group, the gnTEAM: Farzaneh, Martin, Daniel, Mona, Azad, Geraint, George, Chengkun, Rosyzie, Nikola and Ruth.

To my awesome super-skilled colleagues with which I have been sharing the office in the last part of my Ph.D.: Mercedes Arguello Casteleiro, Warren Read,

George Demetriou and Dmitry Tsarkov. Their advice, technical competence and human qualities made me constantly feel among kind giants. It has been an honour to share part of my Ph.D. with you guys.

A special *grazie* goes to my Italian colleagues at The University of Manchester: Simone Di Cola, Fabio Papacchini, Ignazio Palmisano, Chiara Del Vescovo and Sergio Davies. To Fabio Zennaro, particularly, since I had never experienced before an enlightened out-of-the-blue 3-hour chat about One Piece and its impact in the contemporary literature.

A very big *grazie* also goes to Angelo Coletta for his availability in sharing his point of view with me on a professionally challenging matter. Good mentors are very rare and I had the privilege to find him on my path.

To the many scientists I have met around the world and showed me their interest, support and guidance. Among all: Leon Derczynski, Angus Roberts, Hector Llorens, Naushad UzZaman, Roi Blanco, Amber Stubbs, Ozlem Uzuner, Hideo Joho, Adam Jatowt, Yoshinari Fujinuma, John Patrick, Sir Roger Penrose, Anna Lisa Gentile, Noah Smith, Mário Figueiredo, Sergio Matos, Slav Petrov, Luis Sarmento, Marteen De Rijke, Koby Crammer, Irena Spasic, Nigel Collier.

To Mavis Bracegirdle and Sean Connolly for their kind and unconditioned support. Their contribution to this quest is brighter than their modesty would ever allow them to recognise. I feel blessed for having met them on my path since they turned in two of the most important role models of my life.

To Italy. Nothing is even remotely close to your beauty. I had to leave you to truly appreciate your magnificence.

To my family, the persons who accepted my leaving without showing me any sign of suffering. I have seen my brothers growing up, my parents becoming more fragile and I have felt impotent against time. The only way I can now pay them back is to promise I will always try to be happy and make others around me

happy. I also want to thank them for somehow raising me to have confidence that is disproportionate with my look and abilities. Well done! :)

To my *nonna*. The greatest of all my desires is to see her there at the graduation ceremony. I can see her taking a flight with fierce proud for his *nipote*'s graduation. I can see her wet eyes, her mouth hardly containing thousands of words, her facial expression while looking herself around, like a child in her best dream. I can see you *nonna*, and I have learnt to make it suffice.

This period of my life has been stressful. I would have quitted after the first months if it wasn't for my fiancée (wife if you are reading this after the 12:00 CEST of Sunday, 20th December 2015), Marilena. In the darkest moments, she was the only reason I kept pushing. She has been my constant and reliable source of joy, inspiration, solace, encouragement and love. She made me strong and self-confident, humble and free. My log pose is now set up on the next island, you're now, more than ever, part of my crew.

Finally, I thank myself.

to Marilena

Author

Michele Filannino has spent the last 4 years at The University of Manchester (School of Computer Science) working on his PhD, the outcome of which is this thesis.

In 2008 he received his bachelor's degree in Computer Science from the University of Bari "A. Moro". For his dissertation he worked on making particular business process modelling components automatic via artificial intelligence and natural language processing techniques. In the same year he published his first research paper about introducing serendipity in recommender systems.

In 2010 he received his MSc in Computer Science and Computational Intelligence with full marks *cum laude* at the same university. For his dissertation, he investigated a measure of semantic similarity between words based on Wikipedia. After that, he was awarded the "Working Capital" prize from Telecom Italia s.p.a., where he did research for a year on the application of fuzzy logic to sentiment analysis algorithms.

Chapter 1

Introduction

"Begin at the beginning," the King said, gravely, "and go on till you come to an end; then stop."

– Lewis Carroll, *Alice in Wonderland*

The advent of the World Wide Web (WWW) has celebrated the beginning of a new era characterised by the abundance of digital data [160], and more specifically textual digital data (news, blogs, social media, electronic health records, research papers, encyclopaedia, dictionaries, magazines, shopping catalogues). However, the availability of such data does not provide us any good when we are not able to store, manage, filter and retrieve it efficiently [155].

The process of interpreting texts is crucial for acquiring knowledge and ensuring technological developments. For this reason, we are witnessing enormous efforts in making computers mimicking tasks such as sentiment analysis, language translation, document retrieval, where they can independently achieve a good accuracy. On the other hand, in other tasks such as writing, shopping, travelling, composing music *et cetera* they only act as support to people.

One of the human intelligent behaviour that computers are not yet able to mimic is the capacity of **interpreting facts with respect to time**. This ability

allows us to order facts on a time-line and also recognise connections among them (e.g. causalities, implications, co-occurrences and temporal contradictions). By means of it, people are able to organise, summarise and combine different pieces of information to work out new insights or deduce facts that are not explicitly mentioned.

Being able to perform such activity, would enhance the exhibited intelligence of pre-existing natural language-based systems (e.g. question answering, information retrieval, information filtering and automatic summarization), and support the advent of new natural language applications (some of which will be presented later).

The work presented in this thesis is focussed on Information Extraction (IE), which is part of Text Mining (TM). More specifically, this thesis investigates how computers can automatically extract temporal information from documents written in English. This is a challenging task since there are several different entities to recognise and specific sub-domain languages to consider.

According to the ISO-TimeML standard [133] for annotating temporal information in text, three linguistic entities are essential for temporal processing: temporal expressions, events and temporal links.

A **temporal expression**, also called *timex*, refers to any natural language phrase denoting a temporal entity such as an interval or a time point [52]. More specifically, a temporal expression may refer to:

- day times (*noon*, *3p.m.*, *the evening*, ...).
- dates at different granularity: days (*yesterday*, *Jan 8 2001*, *last Friday*, etc.), weeks (*next week*, *the second week of July*, etc.), months (*in two months*, *August 1971*), seasons or business quarters (*last spring*, *the third quarter*, etc.), years (*1978*, *the previous year*), centuries, etc.

- durations (*two months*, *five hours*).

- sets or frequencies (*every Thursday*, *the first Sunday of the month*).

Eventualities, typically called **events**, are natural language phrases which denote something that is happening [133]. More specifically, events refer to the following types of expressions:

- situations that happen or occur, which can be either punctual (*born*, *erupted*, etc.) or last for a period of time (*was evacuated*, *expecting*, etc.).

- states or circumstances in which something obtains or holds true (*shortage*, *dormant*, *attack*, etc.).

Finally, a **temporal link** represents the relationship holding between two temporal expressions, two events, or between a temporal expression and an event, and indicates how they are temporally related [133]. For example, two events can start at the same time or they can overlap. An exhaustive list of possible temporal relations is shown in Figure 1.1.

In a sentence like *The Prime Minister said yesterday that the reform promoted three months ago has been very successful.*, the phrases *yesterday* and *three months ago* are temporal expressions, where *said* and *promoted* are events. Also, *said* is temporally connected to *yesterday*, and *promoted* is temporally connected with *three months ago* and *said*. The ISO-TimeML representation of the annotated sentence is presented in Figure 1.2, where temporal expressions are annotated with the eXtensible Markup Language (XML) tag TIMEX3, events with EVENT tags and temporal relations with TLINK tags. Each of them includes attributes, which express some of their semantic properties: type and value for temporal expressions, class for events, and relType for temporal relations.

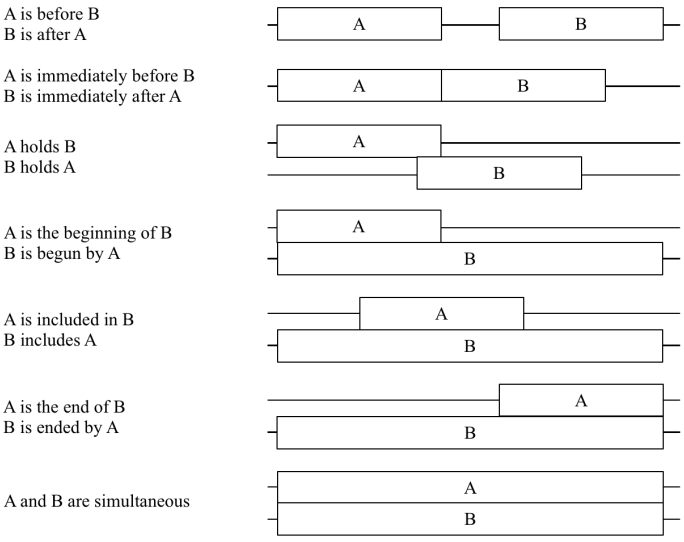


Figure 1.1: Types of temporal relations.

```
<?xml version="1.0" ?>
<TimeML xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:noNamespaceSchemaLocation="http://timeml.org/timeMLdocs/TimeML_1.2.1.xsd">
    <DOCID>Example_document</DOCID>
    <DCT>Apr 17, 2012</DCT>
    <TITLE>Example document</TITLE>
    <TEXT>
        The Prime Minister said
        <EVENT eid="e1" class="OCCURRENCE">said</EVENT>
        <TIMEX3 tid="t1" type="DATE" value="2012-04-16">yesterday</TIMEX3>
        that the reform
        <EVENT eid="e2" class="OCCURRENCE">promoted</EVENT>
        <TIMEX3 tid="t2" type="DATE" value="2012-01-16">three months ago</TIMEX3>
        has been very successful.
    </TEXT>
    <MAKEINSTANCE eiid="ei1" eventID="e1" pos="VERB" tense="PAST" aspect="NONE" />
    <MAKEINSTANCE eiid="ei2" eventID="e2" pos="VERB" tense="PAST" aspect="NONE" />
    <TLINK eventInstanceID="ei1" relatedToTime="t1" relType="DURING" />
    <TLINK eventInstanceID="ei2" relatedToTime="t2" relType="DURING" />
    <TLINK eventInstanceID="ei1" relatedToEventInstance="ei2" relType="AFTER" />
</TimeML>
```

Figure 1.2: TimeML annotation for the sentence *"The Prime Minister said yesterday that the reform promoted three months ago has been very successful."*. The annotation contains: (I) two temporal expressions (*"yesterday"* and *"three months ago"*), (II) two events (*"said"* and *"promoted"*), and (III) three temporal relations (*"said"* $\xrightarrow{during}$ *"yesterday"*, *"promoted"* $\xrightarrow{during}$ *"three months ago"* and *"said"* $\xrightarrow{after}$ *"promoted"*).

A Temporal Information Extraction (TIE) system is a software that, given a piece of text in input, can automatically provide annotations of temporal expressions, events and temporal relations (see Figure 1.2 for an example).

This thesis mostly focusses on Temporal Expression Extraction (TEE) which is typically divided in two main sub-tasks: identification and normalisation [4]. The former aims at detecting the correct span of expressions, whereas the latter aims at predicting the semantic properties of pre-identified entities.

1.1 Motivation

The idea of extracting temporal information from texts is not new. The scientific literature already includes annotation standards [132, 66, 133], extraction approaches and ad-hoc evaluation metrics [184, 180], which have been supported and fostered by the organisation of several shared tasks [184, 185, 181, 169]. In spite of the research carried out so far, TIE is still an open question, with numerous shared tasks organized at TempEval [184, 185, 181], Temporalia [78], i2b2 [169], Clinical TempEval [16] and many other scientific conferences.

The research on TIE started in the general domain, where the availability of news, their easiness to be gathered, the relative lack of typos or misspellings, and the almost strict adherence to the English grammar made them a suitable candidate for TIE research.

Furthermore, a growing research interest has been developing around TIE on clinical data [3, 199]. The automatic temporal analysis of such narratives takes on a great importance, since they describe patients' history through clinical events which are not necessarily chronologically presented in the text. Being able to automatically analyse them has the potential of making clinical audits easier, enhance time efficiency, reduce clinical errors and improve patients' quality of

care [64]. It also makes all those data machine processable, enabling investigations such as disease progression modelling, analysis of the effectiveness of treatments and visualisation of patients' clinical pathways. However, despite of the growing research interest, there are no publicly available TIE systems for clinical data yet and the ones tailored for the general domain perform poorly because of the specificity of the clinical sub-language [24, 32].

In several TIE shared tasks, rule-based approaches have been performing better than the data-driven counter parts, suggesting that the latter approaches are not suited for the task. However, the possibility that the poor performance of data-driven systems depends on the size of the training data sets used so far, being too small to allow automatic learning, must be taken into account. On the other hand, rule-based systems might not suffer from that because the data sets are large enough to allow linguistic experts to generalise their rules, which typically have the form of regular expressions patterns.

The work presented in this thesis focuses on exploring the Machine Learning (ML)-based approaches. Such choice is justified by the fact that those systems solve or mitigate some problems typically exhibited by rule-based systems:

- they are difficult and expensive to develop, since they require linguistic expertise.
- they are not designed to be easily adapted to other languages, sub-languages, genres and sub-domains.
- the contribution of a rule becomes negligible as the number of rules grows (long-tail problem).

The methodology presented in Chapter 3 introduces a new TIE strategy, fully data-driven, which includes a post-processing component aiming at improving the results.

1.2 Research hypotheses and questions

The main research hypothesis of this thesis is that it is possible to extract the temporal flow of events from narratives written in English, and that such task can be, to a certain extent, automatically learned by computers from data, in addition to linguistic rules carefully coded by experts. In the ability of computers to automatically learn new tasks lies the possibility of drastically reducing development costs (time and effort) of such systems, and making them easier to be adapted to different sub-languages, domains, etc.

More specifically, the research questions addressed in this thesis are:

- Can we automatically extract temporal information from documents by using *ML* techniques?
- Can we *reliably* assess what are the linguistic attributes which contribute to the extraction?
- Do automatically annotated corpora help to train better data-driven systems?
- Can we further improve data-driven systems' predictions without any human intervention?
- How can a TIE system be automatically adapted to a different domain?
- What could be interesting applications of TIE other than temporal ordering of events?

1.3 Aim and objectives

The main aim of this thesis is to design, develop and evaluate a data-driven framework for the extraction of temporal information from general and clinical data written in English.

More specifically, the objectives of this research are:

1. Design, implement and evaluate a fully data-driven strategy for TIE, which has a novel architecture and performs competitively well with respect to the pre-existing strategies proposed in the literature.

2. Harvest the TIE literature to identify all the commonly used types of attributes and analyse whether their contribution is beneficial or detrimental for the task. The common ML-based measures of error will be used to assess the performance.

3. Investigate how the proposed strategy can be used with a different genre of text: clinical narratives. Assess what resources need to be adapted, what components need to be replaced, implement the system and analyse the errors.

4. Investigate and measure the efficacy of using TIE strategies to support two applications: temporal orientation classification and temporal footprint prediction. For the former task, the common classification accuracy measures will be used, against a shared official benchmark test set. For the former task, an error measure mutuated from the field of temporal algebra will be used.

1.4 Research contributions

This thesis provides the following research contributions:

1. An extensive analysis of the significance of features with respect to all the feature types previously used in the literature. The results show that the use of morphological features is statistically equal or better than other more complex models. Such result is not generalisable to Named Entity

Recognition (NER) tasks other than TEE and is limited by the way features have been previously used in the literature.

2. An hybrid strategy for temporal expression extraction which uses both rules, for the normalisation phase, and ML-based approaches, for the identification phase. Such strategy, when benchmarked on the TempEval-3 test set, proved to perform competitively well with respect to the rule-based systems and better than the other presented ML-based systems.

3. An instantiation of the previously described strategy tailored for the clinical domain. The normalisation component has been adapted by adding new rules, specifically designed for clinical narratives. The system, presented at i2b2/2012, achieves state-of-the-art extraction performance.

4. A novel methodology to predict the temporal orientation of search engines' queries, which relies on features derived from TIE techniques. Such methodology improves the search engines' accuracy by filtering-out temporally irrelevant results. The research shows that TIE-based features are crucial for the temporal orientation classification task.

5. The concept of *temporal footprint*, which expresses entities life-span on the time-line, along with a comparison of mining methodologies to be used with persons' Wikipedia pages. We found that the use of TIE systems is justified for long texts, rather than short ones, where a simple regular expression-based approach is effective.

1.5 Research outcomes

The work presented in this thesis has produced the following research outcomes:

1.5.1 Publications (chronologically ordered)

- Michele Filannino. Temporal expression normalisation in natural language texts. *CoRR*, abs/1206.2010, 2012

- Michele Filannino, Gavin Brown, and Goran Nenadic. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics

- Aleksandar Kovačević, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, 20(5):859–866, 2013

- Michele Filannino and Goran Nenadic. Mining temporal footprints from Wikipedia. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 7–13, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University

- Michele Filannino and Goran Nenadic. Using machine learning to predict temporal orientation of search engines' queries in the temporalia challenge. In *NTCIR-11, EVIA 2014 (NII Testbeds and Community for Information Access Research)*, 2014

- Michele Filannino and Goran Nenadic. Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. *Data & Knowledge Engineering*, 100:19–33, 2015

1.5.2 Resources and availability

The resources presented in this thesis (code, datasets and results) can all be found on my academic web page[1] or GitHub page[2]. Each chapter will provide links to the specific resources.

1.6 Thesis structure

The next chapter (Chapter 2) introduces the background work for this thesis. It presents the definition of temporal information in linguistics, the main Natural Language Processing (NLP) techniques, the evolution of the proposed annotation schemas and the literature on TIE. It also presents the main scientific challenges for temporal TM, the evaluation metrics used to benchmark the proposed strategies and, finally, the main applications of TIE.

Chapter 3 and Chapter 4 form the core methodological contribution of this thesis, and are based on peer-reviewed journal papers [92, 58]. Chapter 3, in particular, is based on previous publications [55, 54] (see Appendix A).

Chapter 3 presents a novel TIE methodology designed on general domain documents. The proposed ML-based methodology uses an optimal set of features, which have been collected by harvesting the TIE literature and then performing model selection. The system has been evaluated at TempEval-3 [181] and ranked 5th out of 21 submitted runs, as the best performing ML-based system.

Chapter 4 introduces a TM pipeline for TIE on clinical documents. The strategy hybrids the machine-learning technique used in the general domain (see Chapter 3) with rule-based components. The methodology proposed has been tested on the i2b2/2012 [169] data which is a collection of clinical discharge notes. The system

[1] `http://www.cs.man.ac.uk/~filannim`
[2] `https://github.com/filannim`

ranked 1st in the TEE task and 5th in the Event Extraction (EE) task. Chapter 5 and 6 present two novel applications of TIE techniques.

Chapter 5 presents a strategy to improve information retrieval system. It shows that by analysing users' queries with TIE methods, it is possible to predict queries' temporal orientations (present, past, future, atemporal) and filter the results, which ultimately leads to an accuracy improvement.

Chapter 6 introduces the idea of *temporal footprint*, persons' life span on the time-line, and how it is possible to predict them by analysing Wikipedia textual content. The experiments compare different methodologies with respect to the length of the pages. They prove that simple approaches are effective on short Wikipedia pages, whereas TEE methods provides better predictions in case of long Wikipedia pages.

Finally, Chapter 7 wraps up the main contributions of this thesis along with the challenges still left open and the new ones opened by this work. The thesis is concluded by Chapter 8.

This thesis does not have to be read linearly since some chapters are independent from each other. The transition diagram in Figure 1.3 suggests the possible reading paths.

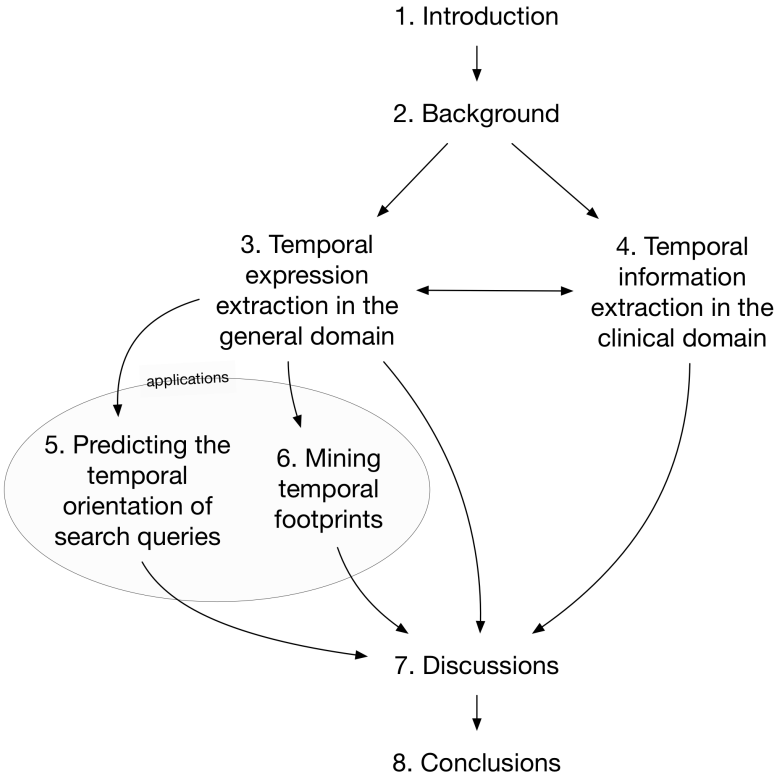


Figure 1.3: How to read this thesis.

Chapter 2

Background

"Il passato è un'indispensabile guida per chi vuol visitare il presente o immaginarsi il futuro."

– Tiziano Terzani, *La porta proibita*

This thesis focuses on the field of Information Extraction (IE) in the area of Text Mining (TM). Feldman [50] defines the discipline as:

> " ... a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalised database records, but in the unstructured textual data in the documents in these collections."

In TM it is assumed that data are presented in text format, written in a specific natural language. The text is seen as an unstructured source of information, as opposed to structured ones such as databases, files etc. The goal is to analyse

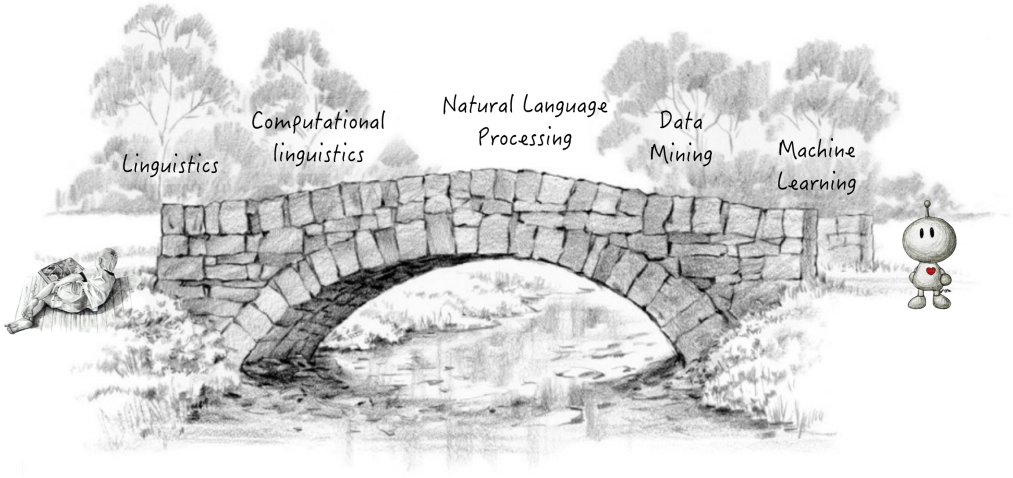


Figure 2.1: The text mining bridge.

and structure the information conveyed by the text, therefore making the data interpretable and further processable by a computer, and highlighting the relevant bits of texts from the rest.

For this reason, TM represents the interface between natural language texts and computers. In bridging such gap, TM draws on advances in disciplines such as Linguistics, Computational Linguistics, Natural Language Processing (NLP), Data mining and Machine Learning (see Figure 2.1).

This Chapter provides an overview of the background relevant to this thesis. It includes an introduction to temporal information in the context of Computer Science (CS) and NLP and the point-of-view in linguistics. It also provides an introduction to the current annotation schema for temporal information. It will present the main community challenges and the specific evaluation metrics used. Eventually, the field of Temporal Information Extraction (TIE) is presented, along with the historical reasons, the most important sub-tasks and the main scientific contributions to the field.

2.1 Natural Language Processing

A typical high-level data-driven TM functional architecture is composed of the following parts:

- Data collection: Data are gathered or collected by using automatic or manual techniques. The automatic ones range from simple document retrieval from a data source to web crawling and filtering [19].

- Pre-processing: documents are analysed and divided in their fine-grained sub-components: sections, sentences, words. If the data are annotated using some annotation schema, such annotations are analysed and incorporated in the data model. Linguistic features, at different levels of granularity are extracted using several NLP techniques: sentence splitting, chunking, part-of-speech tagging, constituency parsing, dependency parsing and reference resolution [106, 81].

- Data mining and pattern analysis: Machine Learning (ML) techniques are applied to the pre-processed data representation in order to train models to detect patterns in the data [50, 7]. The most suitable type of pattern we are searching for depends on the task, which in turn influences the choice of the ML algorithm. In the IE case, manually written rules can be also used in addition to ML-based components or in alternative of them [81].

- Post-processing: when different data mining algorithms are used, this is the stage when the predictions are merged to finally provide the structured information [7].

- Presentation: structured results are stored according to specific annotation schema suitable for the information [135]. Such data are then presented by using visualisation techniques or browsing interfaces [183]. In some cases,

data are not meant to be presented to the stakeholders, but rather to be used by other automatic systems. In this case, the presentation layers correspond to storage.

The **sentence splitting** phase (also called *sentence segmentation*) takes a text and divides it in a list of sentences (see Figure 2.2). The process is not trivial due to the complexity of the natural language. Different methodologies have been proposed, from simple ones typically rule-based [39, 115], to more elaborate ones [26, 172] that use ML sequence labelling models (such as Hidden Markov Model (HMM) [137] and Conditional Random Field (CRF) [94]).

Word segmentation is the task of decomposing sentences in lists of words (see Figure 2.3). In most of the TM research, words are considered the most atomic component of texts. Since, from an abstract point of view, word segmentation is a task similar to the sentence splitting, the evolution of approaches adopted resembles the ones mentioned for that task: from rule-based to ML-based [123, 164, 86].

In languages such as English, Italian, Russian, French and Spanish the boundary is often signalled by a white space, and the task is nowadays considered solved up to a satisfactory level. Conversely, in the case of Chinese or Arabic, where the white space character is not a unique signal, the research activity is still intense and the problem is far from solved [34, 33].

Part-of-speech taggers classify words into their parts-of-speech and labels them accordingly (see Figure 2.4). Parts-of-speech (PoS) are categories of words which have similar grammatical properties. In English, for example, frequent PoS are noun, verb, pronoun, adjective, adverb, preposition, conjunction, interjection and determiner.

Automatic part-of-speech taggers have been extensively studied in English and other languages, since they provide a fundamental piece of linguistic information. Generally speaking, the current performance of part-of-speech taggers has reached

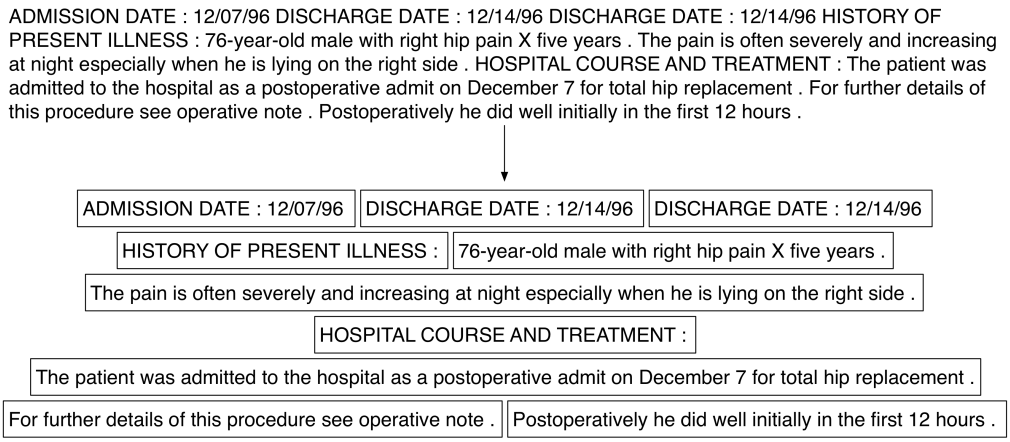


Figure 2.2: Example of sentence splitting.

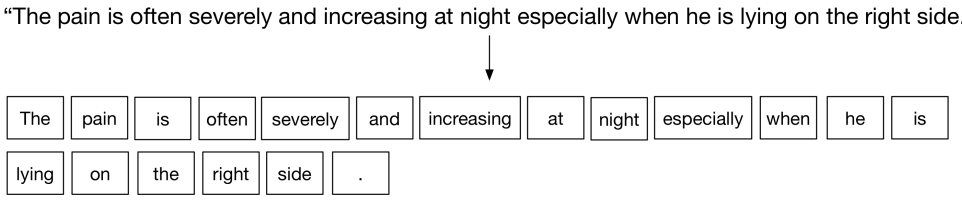


Figure 2.3: Example of tokenisation.

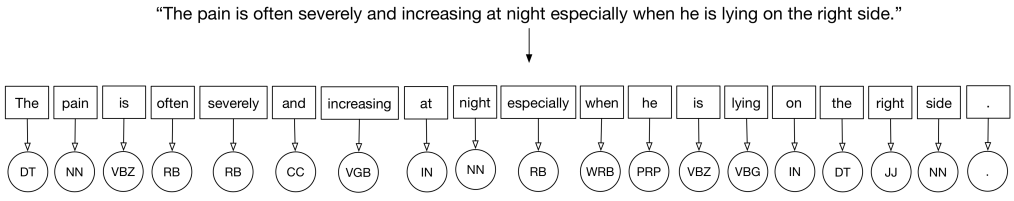


Figure 2.4: Example of part-of-speech tagging.

a high level of performance and is nowadays universally considered a solved problem [157, 105, 127, 144]. The performance is generally lower in specific domains than in the general domain [173, 140, 153].

Constituency parsers are concerned with how words group together in phrases: noun phrases, verb phrases, adjective phrase and others (see Figure 2.5) [106, 81]. The aim of constituency parsing is to predict the phrase tree structure for a particular sentence, once it has been word segmented. This task is intimately related with the existence of a formal language grammar, which disciplines how correct sentences are produced [36, 37, 38]. Each node in the tree is called *constituent* and every level of it can be seen as the application of a specific grammar rule [85, 195, 47, 60, 109].

For a particular sequence of words, multiple constituency trees may be valid according to the grammar. Some of the trees will be more likely to predict the true structure, which ultimately lead to the meaning of the sentence.

Dependency parsing is concerned with how words in a sentence relate to each other according to a set of grammatical functions which holds between the head and dependent word (see Figure 2.6) [106, 81]. A dependency representation is a labelled directed graph, where the nodes are the words and the labelled arcs are dependency relations [48, 120, 112].

Reference resolution is the process of automatically connecting different expressions to the entity they actually refer to. A reference, in fact, is the process by which a speaker uses expressions like *Maria* and *she* in the same passage to refer to the same entity (a person called Maria). The expression *she* is called 'referring expression', and *Maria* is called 'referent'. *Maria* and *she* co-refer since they refer to the same entity.

When the referent refers to an entity which is previously/successively mentioned in the text, the reference is called **anaphora**/*cataphora* and the referring expression is called anaphoric/cataphoric. Anaphora resolution is a special case

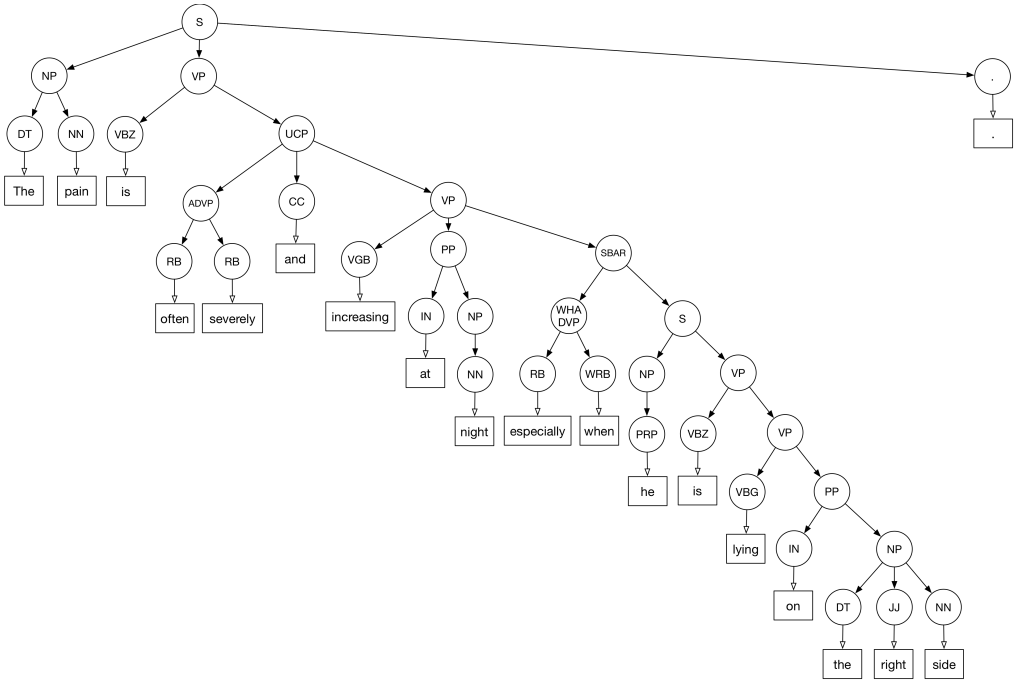


Figure 2.5: Example of constituency parsing.

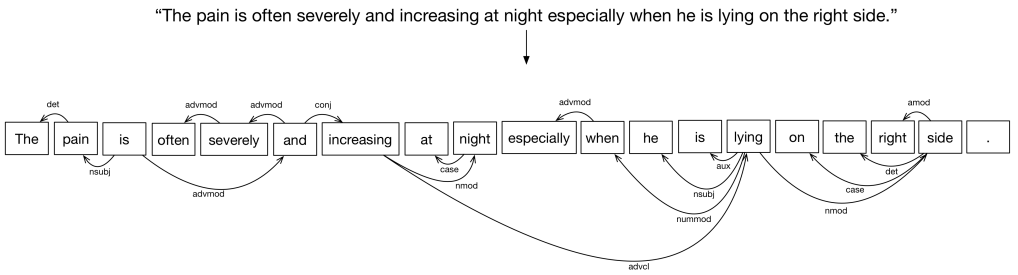


Figure 2.6: Example of dependency parsing.

of reference resolution. The literature presents a wide range of methodologies to tackle the problem [25, 63, 117] and those tasks are far from being considered solved in NLP.

At the end of the pre-processing steps, a sentence is ideally represented in a structure that resembles the one depicted in Figure 2.7. It is a directed graph in which the sub-tree structure is the constituency tree, the leaves are the segmented words and the arrows between the pairs of words are the dependency relations. The dependency relations reveal that *'pain'* is the subject of the verb *'is'* and that the same relation holds between the verb *'lying'* and the pronoun *'he'*.

This enriched structure is afterwards used in the **Data Mining and Pattern Analysis phase**. In the case of Named Entity Recognition tasks (such as temporal expression and events identification) the phase is carried out using ML or rule-based systems. Both types of system are based on features computed from the previously mentioned data structure at different linguistic levels: morphological (related to the internal structure of words), syntactic (related to the structure of sentences), semantic (related to the meaning). Features are selected according to their expected contribution in the learning process, since those not related to the classification task negatively affect the performance. In the case of rule-based systems, such models are expressed in the form of explicit rules formulated by experts in the domain. Predictions are then computed according to the fact that one or multiple rules are activated on a set of words. Rules typically take into account fewer features than ML systems.

Sometimes predictions provided in the Data Mining and Pattern Analysis phase can be further improved, typically by discarding or fixing consistently wrong ones. This is done in the **post-processing phase**. Often post-processing techniques are designed as a set of precision-optimised rules [92], in some other cases they are designed as a ML-based system which acts on top of the previous phases [55].

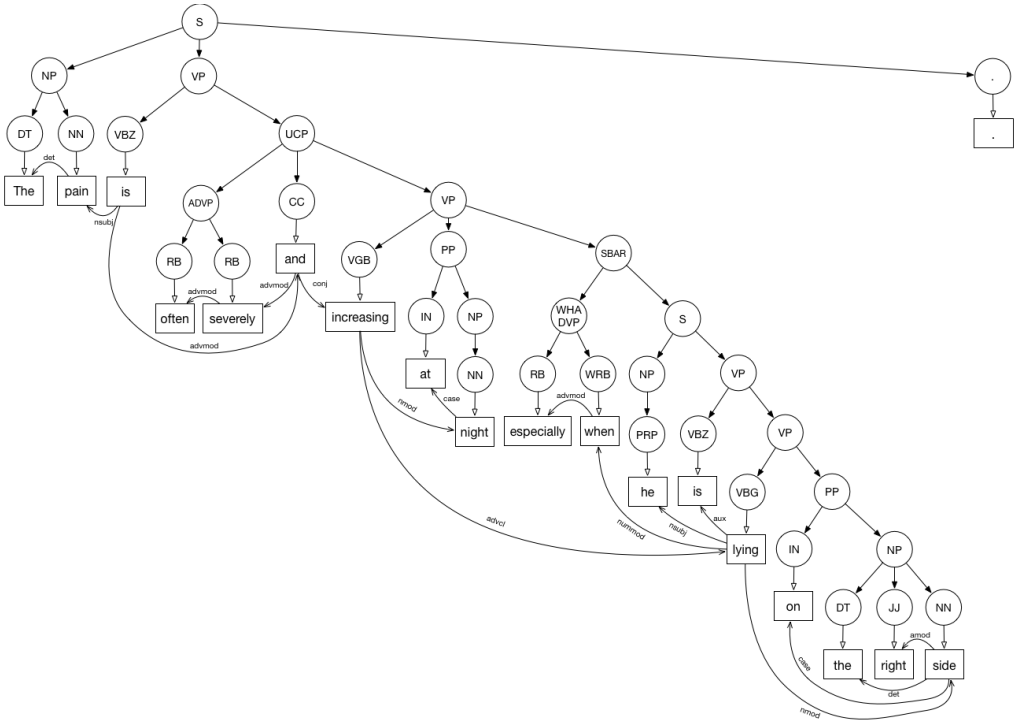


Figure 2.7: Typical NLP pre-processing data structure for the example sentence *The pain is often severely and increasing at night especially when he is lying on the right side.*.

The last step of a typical TM pipeline is the **presentation** of the results: storage and/or visualisation. The data are usually represented according to a suitable format which is tailored for their use in the next stage. The annotation schemas used for TIE will be presented in the following section. The use of standard annotation schemas facilitates the visualisation (see Figure 2.8).

2.2 Temporal information

The expression *'temporal information'* refers to the existence of linguistic structures, shared by almost[1] every modern language, to anchor facts to temporal frames. Those structures can be found in texts in the form of phrases. In linguistics such phrases are called *temporal transitions*: expressions used to convey frequencies (e.g. *every morning*), durations (e.g. *for two months*), precise time points (e.g. *at that time*, *next July*), beginnings (e.g. *before*, *then*, *since*), endings (e.g. *finally*, *in the end*) and contemporaneities (e.g. *meanwhile*, *at the same time*).

Although the obligatory temporal expression in English is the tense [161], Hans

[1] In Kuuk Thaayorre, the language spoken by Pormpuraaw, an Aboriginal community situated on the west cost of Cape York Peninsula, time flows according to cardinal directions [61]. Finally, Amondawa tribe in Amazon, seems not to have the abstract concept of time at all, therefore they do not have words like 'year' or 'now' [149].

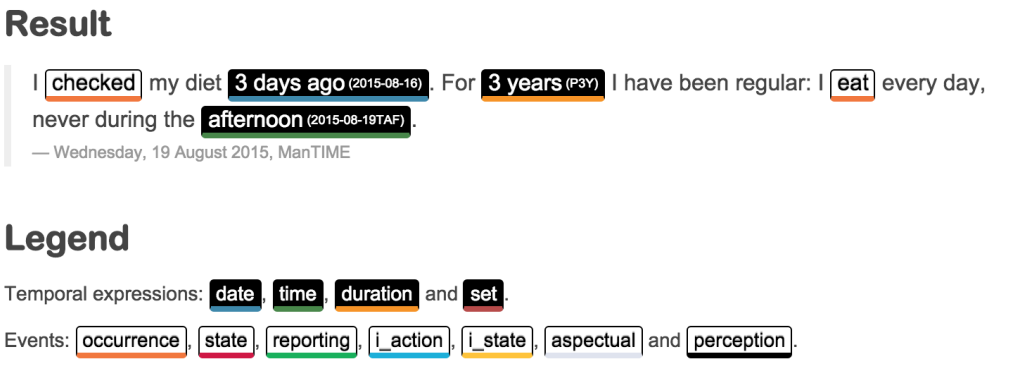


Figure 2.8: Example of visualisation of an annotated sentence

Reichenbach argued [143] that the simplest sentence is understood in terms of three temporal notions: speech time, reference time and event time. The temporal values and their function can vary according to the syntactic use of such temporal transitions, but their relational values are consistent. This means that the expression *now* in a text can refer to different time points, can have different part-of-speech tags but it will be always anchored to the same time (speech time in this case). This consistency is a characteristic that English shares with other languages.

In a study conducted by Carlota Smith on the interpretation of temporal expressions, she highlighted that "the domain of temporal specification is shown to be larger than a sentence" [161], which has the consequence of making the context crucial for the interpretation of temporal expressions [162].

Temporal expressions elicit a binding between the natural language domain and the time domain because it is always possible to represent such expressions as an exact time point (e.g. *21/07/1985*), interval (e.g. *the last 5 days*) or set (e.g. *every two days*) using the ISO 8601 standard[2] (see Section 2.3 for more details).

Temporal expressions can mainly be divided into three different types [4]: *fully-qualified*, *deictic* and *anaphoric*.

Fully-qualified A temporal expression is fully-qualified with respect to the binding when all the information required to infer a point in the time domain are fully included inside the expression. In this category, the following patterns fall: *3-5 August 2001*, *21st July 1985* or *31/04/2011*. Fully-qualified expressions are the easiest to detect because of their rigid syntactic form, although some complex variations can be found, such as *on the 7th day of June*, *five years after the end of the Second World War II*.

Deictic A temporal expression is deictic when inferring the binding with the time domain necessarily requires to take into account the time of utterance

[2]http://www.w3.org/TR/NOTE-datetime

(i.e. when the document has been written or when the speech has been given). Deictic expressions could not be properly associated to a precise time without that piece of information. Examples of deictic temporal expressions are *today*, *yesterday*, *last Sunday* or *two months ago*.

Anaphoric A temporal expression is anaphoric when it can be mapped to a precise point in time only by taking into account temporal expressions previously mentioned in the text. Examples of this category are *March 15*, *the following week*, *Saturday*. The only difference between deictic and anaphoric expressions is the location of the temporal reference: for deictic expressions it is the time of utterance or publication, while for anaphoric expressions it is a time previously evoked in the text.

There are other kinds of temporal expression categorisations. Pustejovsky et al. [130], for example, identify the possible shapes of temporal expressions with respect to their semantics and differentiate the following types:

- time or date references (*1:20am, July 26th, 1999*),
- time references that anchor on another time (*three ours after noon*),
- durations (*a few days, several weeks*),
- recurring times (*every third month, twice in the hour*),
- context-dependent times (*today, last year*),
- vague references (*somewhere in the middle of June, the near future*),
- times indicated by an event (*the day S. Berlusconi resigned*).

In this thesis I will adopt the first categorisation.

2.3 TIE annotation schemas

The temporal expression identification task was born as a named-entity recognition task. In 2004, with the Automatic Content Extraction program, it became a separate task called Temporal Expression Recognition and Normalisation (TERN). The new name highlighted an important dichotomy in the task: recognition (or identification) and normalisation (see Section 2.6).

The first appearance of a temporal expression annotation standard was in 1995 during the Sixth Message Understanding Conference (MUC-6) when the tag TIMEX was proposed. Its aim was to separate and highlight temporal expressions from the rest of the text. In the following edition of the same conference, in 1998, the general annotation guidelines extended the definition of temporal expression to named holidays (e.g. "Christmas" or "Easter") and time-zone mentions (e.g. "{1:30 p.m. Chicago time}$_{\text{TIMEX}}$" rather than "{1:30 p.m.}$_{\text{TIMEX}}$ Chicago time"), and additionally, eliminated determiners introducing temporal expressions (e.g. "around the {4th of May}$_{\text{TIMEX}}$" rather than "{around the 4th of May}$_{\text{TIMEX}}$").

The first specific guidelines for temporal expression annotation were suggested in 2001, after the publication of the Sheffield Temporal Annotation Guidelines (STAG) [158]. The TIMEX tag was extended to an annotation schema with the purpose of identifying chunks related to temporal aspects, events and temporal relations among them. The schema consisted of a set of tags: TIMEX, EVENT and SIGNAL. The last tag is used to annotate temporal expression triggers such as *on*, *during* or *for*. Temporal expressions were limited to dates and times, but it was possible to express the relations among events and temporal expressions. However, at this stage, the meaning of temporal expressions was not represented in a machine understandable way.

In the same year, the STAG were reviewed and extended as part of a DARPA program by introducing more expressive power to the previous annotation [59].

The new extension allowed to represent each temporal expression by a shared non-ambiguous notation. TIMEX2 introduced in particular one new attribute, called VALUE, which is filled with the ISO 8601 standard representation for dates and times [75]. The purpose of this standard is to provide an unambiguous and well-defined method of representing dates and times, in order to avoid misinterpretation of numeric representations of dates and times, particularly when data are transferred between countries with different conventions for writing numeric dates and times.

In 2002, for the Question Answering workshop, a first XML-markup language for a formal specification of events, temporal expressions and relations was created and named *TimeML* [132]. It used TIMEX3 to denote temporal expressions and, at the same time, provided tags to annotate events, temporal triggers (called 'signals') and event-event, temporal-event and temporal-temporal relationships. The framework introduced a new set of tags in order to represent different aspects of the temporal information: SLINK or subordinate links between verbs (e.g. "John saw Mike singing that night"), ALINK or aspectual links between events and their arguments (e.g. "John stopped talking"), and CONFIDENCE tag to allow annotators to express certainty about their annotations.

In Figure 2.9 the evolution of annotation standard is illustrated.

Five years after the creation of TimeML annotation framework, it was presented to the ISO for consideration as standard and approved in March 2009. ISO-TimeML inherits the framework from TimeML 1.2.1 (the last TimeML version) adding some useful documentation for the annotation on different languages (namely Chinese, Italian and Korean). The framework is not language specific and allows to represent multiple aspects of temporal information.

A customisation of the ISO-TimeML is the i2b2/12 Temporal Annotation schema [169] which has been tailored specifically for clinical data. In this case, the definition of event has been extended to cover clinically relevant events (e.g.

```
John <EVENT eid="e1" class="OCCURRENCE" tense="PAST"
   relatedToTime="t1" timeRelType="AFTER" signalID="
   s1">left</EVENT> <TIMEX tid="t1" type="COMPLEX"
   eid="e2" signalID="s1" relType="after">2 days</
   TIMEX> <SIGNAL sid="s1">before</SIGNAL> the <EVENT
    eid="e2" class="OCCURRENCE">attack</EVENT>
```

(a) TIMEX

```
John <EVENT eid="e1" class="OCCURRENCE" tense="PAST"
   relatedToTime="t1" timeRelType="AFTER" signalID="
   s1">left</EVENT> <TIMEX tid="t1" type="DURATION"
   value="P2D" mod="after">2 days</TIMEX> <SIGNAL sid
   ="s1">before</SIGNAL> the <EVENT eid="e2" class="
   OCCURRENCE">attack</EVENT>
```

(b) TIMEX2

```
John <EVENT eid="e1" eiid="ei1" class="OCCURRENCE"
   pos="VERB" tense="PAST" aspect="NONE" polarity="
   POS">left</EVENT> <TIMEX3 tid="t1" type="DURATION"
    value="P2D" temporalFunction="false">2 days</
   TIMEX3> <SIGNAL sid="s1">before</SIGNAL> the <
   EVENT eid="e2" eiid="ei2" class="OCCURRENCE" pos="
   NOUN" tense="NONE" aspect="NONE">attack</EVENT>
<TLINK eventInstanceID="ei1" signalID="s1"
   relatedToEvent="ei2" relType="BEFORE" magnitude="
   t1"/>
```

(c) TimeML

Figure 2.9: Evolution of the annotation standards. The main difference between TIMEX [158] and TIMEX2 [52] is the presence of the VALUE and MOD attributes. In TimeML [132] annotation there is a new tag, TLINK, that is used to elicit the temporal link among events and temporal expressions.

problems, treatments, tests, clinical department names) and the temporal relations have been simplified by reducing the number of relation types.

Finally, temporal aspects are relevant for other NLP sub-fields, where their presence plays an important linguistic role. For this reason, temporal aspects can be found in many different annotation schemas. An example is the field of Discourse Analysis (DA) where the class TEMPORAL is used to highlight the fact that two arguments are temporally related [131]. In the Penn Discourse Tree Bank, two types of temporal relations are annotated depending on whether the relation is temporally ordered or overlapping: synchronous and asynchronous respectively. In the case of Semantic Role Labelling (SRL), the PropBank corpus creators recognised the importance of temporal aspects by introducing *temporal markers*, which indicate when an action takes place [20].

2.4 Community challenges for temporal text mining

Since the beginning of the research in this field, different challenges and conferences have been organised, stimulating new ideas and resources (tools, corpora, software, gazetteers). The aim of these events has been to investigate the ways of tackling the problem and assess the state-of-the-art. All these efforts have been initially spent to cover the general domain.

The most important challenge for Temporal Expression Extraction (TEE) is TempEval. Its first edition[3] was hosted by SemEval-2007 (International Workshop on Semantic Evaluation). It introduced TimeBank, a corpus of news temporally annotated using TimeML. The challenge proposed three tasks focussed on temporal relations between:

- an event and a temporal expression in the same sentence.

[3] http://www.timeml.org/tempeval/

- an event and the document creation time.
- the main events of two consecutive sentences.

The second edition of TempEval[4] was hosted by SemEval-2010. It was based on TimeML and proposed six tasks:

- determine the extent of the time expressions in a text, along with the value of the features TYPE and VALUE.
- determine the extent of the events in a text along with the value of the features TENSE, ASPECT, POLARITY, and MODALITY.
- determine the temporal relation between an event and a time expression in the same sentence.
- determine the temporal relation between an event and the document creation time.
- determine the temporal relation between two main events in consecutive sentences.
- determine the temporal relation between two events where one event syntactically dominates the other event.

TempEval-3[5] has been hosted by SemEval-2013 and provided a revised version of the TimeBank corpus along with a 1-million-word corpus annotated using the top three TIE systems as benchmarked at TempEval-2. Chapter 3 discusses whether that corpus can be useful for training temporal expressions.

Another series of related conferences is the TIME International Symposium on Temporal Representation and Reasoning[6] which started in 1994. This are more

[4]http://timeml.org/tempeval2/
[5]http://www.cs.york.ac.uk/semeval-2013/task1/
[6]http://time.di.unimi.it/TIME_Home.html

related to interval temporal logic, verification, reasoning and ontologies of time and space-time, although some space is given to temporal information extraction.

In 2014, the NII Testbeds and Community for Information access Research (NTCIR) organized Temporalia, a pilot task in the area of Information Retrieval (IR) to foster research in temporal information access. The fact that time plays a crucial role in estimating information relevance and validity requires search engines to be able to consider temporal aspects of information in greater detail. Temporalia proposed a challenge that establishes common grounds for designing and analysing temporally-aware information access systems.

In the clinical domain, the Informatics for Integrating Biology & the Bedside (i2b2) series of conferences has originally proposed the temporal information extraction task and provided a gold corpus to the community. The temporal aspect has been the subject of two different editions: 2012 and 2014. The former was exclusively focussed on the temporal information extraction task. It proposed an ad-hoc annotation schema based on TimeML, but specifically tailored for clinical data, and required participants to build systems able to automatically identify clinical temporal expressions (Task A), clinical events (Task B), temporal relations from gold annotated entities (Task C) and temporal relations from not annotated data (Task A+B+C). The 2014 edition hosted the second task which aimed at identifying risk factors for heart disease over time. Longitudinal patient records are provided to participants and they are asked to automatically track risk progression over the records.

Clinical TempEval [16], the most recently organised evaluation exercise in TIE, proposed six tasks in line with the TempEval tradition. The tasks are organized in the following way:

- identification of temporal expressions spans.
- identification of event spans.

- normalisation of temporal expressions (TYPE and VALUE attributes).
- normalisation of events (TYPE, POLARITY, DEGREE and MODALITY attributes).
- extraction of temporal relations between events and the Document Creation Time (DCT).
- extraction of temporal relations among narrative containers [134, 116].

The data are annotated according to the THYME guidelines, which have been previously used to craft the i2b2/2012 annotations guidelines. The data set includes 600 de-identified clinical notes from 200 different patients.

2.5 Evaluation metrics

The evaluation metrics used in TIE follow those proposed at TempEval-2.

The identification phase, for both the events and the temporal expressions, is evaluated with respect to *precision*, *recall* and *F1-measure*. These are applied according to two different criteria: (I) strict, all partial annotations are incorrect; (II) lenient, all partial annotations are correct.

For example, if the gold standard annotation contains temporal expression *"tomorrow afternoon"* and the system annotates just a part of it (e.g. *afternoon*), then this is considered a wrong annotation for the strict criteria, and a correct one for the lenient. The normalisation phase for temporal expressions and events is performed by checking that any attribute value equals the expected ones. Attribute values are considered incorrect otherwise (Section 8.2.1 illustrates the limitations of such metric).

For temporal relations among events and temporal expressions, the evaluation metric is the accuracy: the percentage of correct relations compared to all the gold

relations. A relation is considered correctly predicted if connects the expected entities with the expected temporal relation. Low Inter Annotator Agreement (IAA) and the functional dependencies among relations [182] have justified the investigation of more appropriate evaluation metrics. UzZaman and Allen [180] proposed the use of *temporal closure*: an automatic reasoning mechanism that derives new relations starting from an initial set by using the known properties of temporal relations (e.g. transitivity). The final extended set is then used to compute precision and recall measures.

2.6 Temporal information extraction

The task of temporal information extraction involves the extraction of three main entities, which corresponds to those recognised by linguists and formalised in the annotation standards. The entities are: temporal expressions, events and temporal relations. Figure 2.10 depicts the TIE task.

This thesis focuses on the first two tasks, whose definitions and background will be presented in the following sections.

2.6.1 Temporal expression extraction

The first system for automatic temporal expression annotation appeared in 1998 [15]. For several years this topic has been approached only from a theoretical perspective. It aroused an increasing interest with the proposal of a temporal annotation schemas and an ad-hoc system for TEE: a monolithic rule-based system that merged identification and normalisation phase, by using TIMEX1 as grounding standard [104]. There was already a ML component that, using C4.5 algorithm [136], tried to resolve some ambiguities in the text. The original aim of the field was to make the annotation phase easier with respect to the previous schemas

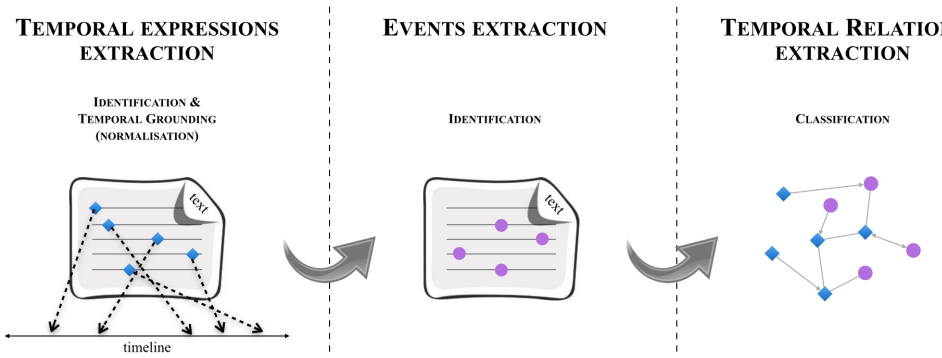


Figure 2.10: Temporal information extraction.

in order to collect annotated data and use the temporal information to enhance performances of question answering systems.

Ahn et al. [4] were the first analysing the problem of TIE from an engineering perspective. Their approach was to divide the task and for each of them design separate components with separate functions: identification and normalisation (see Figure 2.11). By showing that "decoupling recognition from normalisation can improve overall performance" they made the dichotomy universally accepted in the community to the extent that it has been adopted by almost all the recent systems [184, 187, 188, 90].

In the identification phase, the effort is concentrated on how to detect the boundaries of temporal expressions in natural language texts. Although the best performing systems are rule-based, the major part of the approaches so far explored the use ML techniques. In the normalisation phase, the main goal is to interpret the expression and represent it in a proper pre-defined format. The task is particularly challenging due to the presence of non-fully qualified temporal expressions: deictic and anaphoric ones (see Section 2.2). Approaching the problem by using hand-crafted rules turned out to be the best method to accomplish the task, at least in terms of accuracy.

```
Italian prime minister Mario Monti \event{said} \
   timex{yesterday} that the reform has been very
   successful.
```

(a) Example of the identification phase.

```
Italian prime minister Mario Monti <EVENT eid="e1"
   class="OCCURRENCE">said</EVENT> <TIMEX3 tid="t1"
   type="DATE" value="2012-04-16">yesterday</TIMEX3>
   that the reform has been very successful.
```

(b) Example of the normalisation phase.

Figure 2.11: Example of temporal information in text.

2.6.1.1 Identification

The state-of-the-art system, according to the TempEval-3 benchmark (see Section 2.4), is HeidelTime [165, 168]. It is based on the UIMA framework [53] and implements both recognition and normalisation in a monolithic set of rules. HeidelTime is composed of 43 rules which are expressed at lexical level by using expression patterns (regular expression-based) and integrated normalisation functions. Since the same rules perform both identification and normalisation, HeidelTime identifies only expressions for which a normalisation rule is known. The same architecture inspired other works in the field [44, 198].

An analogous approach has been followed by Grover et al. [67] and their LT-TTT2 system. It is built on top of their internal pipeline (called LT-XML2) that integrates several NLP tools. This system has been adapted to the TEE problem with an ad-hoc layer that outputs the internal format in TimeML. The system uses mainly rules, although some pipeline components are based on ML (Maximum Entropy taggers).

The TETI system [27] falls in the category of the contribution on languages different from English. TETI is a system developed to recognise temporal expressions

in Italian texts. Although it is a classical rule-based system, it relies on WordNet semantic relations among temporal expressions.

Llorens et al. [100] presented TIPSem, a temporal expression recognition system that uses semantic roles to better represent the connection between events and temporal expressions. The system works with English and Spanish, and uses WordNet to automatically detect expressions related to time by exploring hypernyms of concepts such as *time_period*, *time_unit* and *time*. The authors successively extended their work by integrating CRFs with semantic roles and a new rule-based normalisation component [101], whereas other experiments have been carried out using a minimal set of features [17].

Mazur and Dale [107], instead, focussed on the use of the dependency tree to identify the extent of temporal expressions. The main limitation of this research was that it entirely relies on the dependency relations within a sentence. Unfortunately, as stated in Section 2.1, the state-of-the-art dependency taggers are still far from providing reliable results. Therefore, errors in the dependency relations impacted negatively on the final performance.

The system mentioned so far are exclusively based on linguistic rules. On the opposite side, data-driven methods have also been used to perform TEE.

Ahn et al. [5] experimented with a rich feature set and a linear kernel-based *Support Vector Machine* [21]. The identification was carried out by classifying tokens in temporal types: *recurrence*, *vague duration*, *duration*, *vague point* and *point*.

Poveda et al. [129] used advanced NLP techniques to extract different types of features: lexical, morphological, syntactic and contextual. They used Support Vector Machine (SVM) with polynomial kernel over tokens experimenting with different polynomial degrees and different feature sets. In a follow-up work, they introduced a sophisticated *Bootstrapping technique* [130] enhancing the recognition

of temporal expressions in a semi-supervised fashion (see Figure 2.12).

UzZaman and Allen [179] produced a complete framework for identification and normalisation of events and temporal expressions called TRIOS (see Figure 2.13). The identification phase is carried out using CRFs trained on a set of morpho-lexical features. In some additional experiments, they showed that adding syntactic features lead to a worse performance.

Adafre and de Rijke [2] pushed the feature engineering further by focussing on the same technique previously used (CRFs) and introducing the first non rule-based post-processing pipeline with the aim of boosting the performance. The proposed system is based on the reclassification of expressions expanding the prediction to the near tokens in an iterative process. The authors shed light on the possibility of using a data-driven post-processing pipeline, which ultimately proved successful. This paper inspired the work on the a posteriori label adjustment pipeline which will be presented in the Chapter 3.

Rigo et al. [145] proposed a system for Italian, Spanish and English based on CRFs and morpho-syntactic features and examined the contribution of each feature in an incremental fashion. Although the experiments were not cross-validated, they set the ground for a systematic feature exploration in the community (see Chapter 3 and 4). A similar approach was later adopted by Jung et al. [80].

Recently, the results from the last temporal information extraction challenge, TempEval-3 [181], show the identification performance ranges from 0.81 and 0.90 in terms of lenient $F_{\beta=1}$ measure (from 0.70 to 0.83 for strict matching).

Brucato et al. [23] introduced a narrower class of temporal expressions, called *named temporal expressions* (e.g. “Autumn Holiday” or “Liberation Day”), which are typically harder to detect since they are not composed by time-related words. They proposed a way of identifying and normalising them by using Linked-Data resources on different languages.

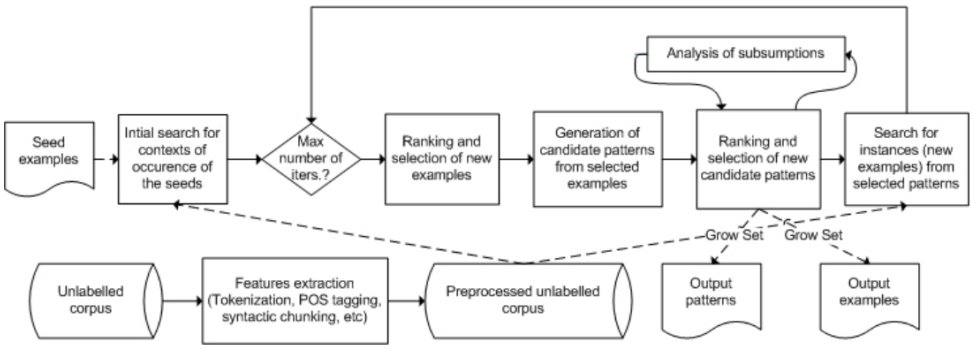


Figure 2.12: Temporal expression identification bootstrapping architecture. Taken from [129].

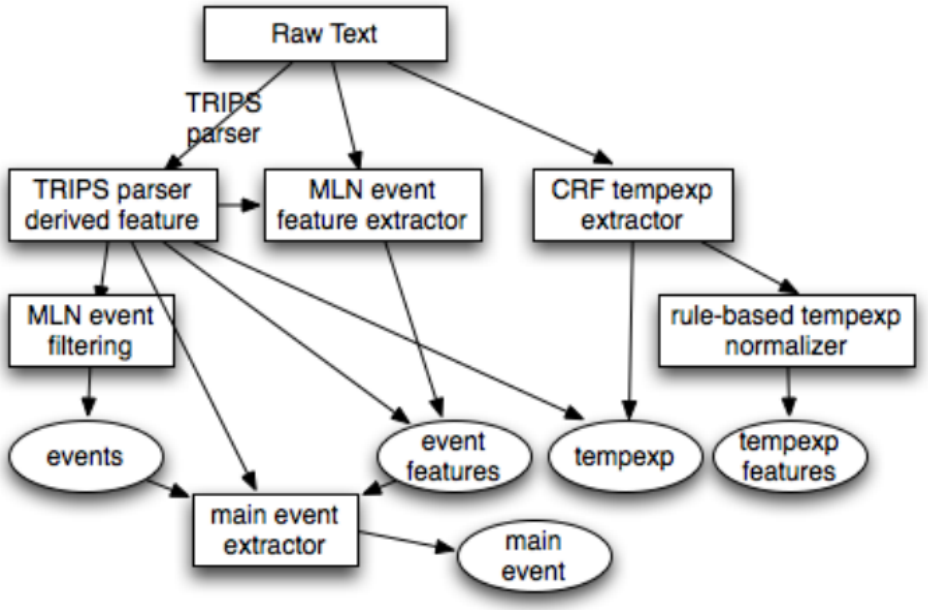


Figure 2.13: TRIOS architecture (from [179]).

The organisation of i2b2 2012 Shared Task on TIE [169] (see Section 2.4) raised the interest in the portability of the pre-existing TIE methodologies to other domains [166], in particular to the clinical one [74], where temporal analysis has the potential of revolutionising the way correlation between symptoms and diseases are studied and discovered. The i2b2 series of shared tasks has fostered the emergence of corpora [62], annotation guidelines [118, 169] and the first generation of clinical TIE systems [96, 142, 92, 170, 193, 65], many of which have been built around pre-existing general domain systems, especially for the TEE phase. The identification performance in the clinical domain, expressed with the $F_{\beta=1}$ measure) ranges from 0.84 to 0.90 (lenient matching).

The legal domain has also been explored with respect to the TIE. For example, Schilder and McCulloh conducted a study on TIE in different types of legal documents [156].

2.6.1.2 Normalisation

The normalisation task gets temporal expressions previously identified in text and predicts their ISO 8601 representation, which expresses their temporal meaning in an inter-operable way (see Section 2.3 for more details). This is a crucial step in the TIE since simply highlighting temporal expressions does not make them interpretable or usable for reasoning purposes further on. The currently available annotated data do not provide enough information to enable the learning (see Section 8.2.1). For this reason, none of the approaches proposed so far in the literature is data-driven.

The normalisation component in TRIOS is a rule-based system that is focussed on predicting TYPE and VALUE attributes of the TIMEX3 tag. Rules are expressed as simple regular expressions over tokens' morphology (e.g. "[0-9][st|nd|rd|th] ?[Jan|Feb|Mar|...|Dec]\\.?" or "[0-9]2:[0-9]2[ap\\.]?[m\\.]?"). This component was

used by the authors to participate to TempEval-2 challenge where it proved to achieve the second best performance. The normalisation system presented in this thesis takes inspiration from it, by extending its functionalities (see Chapter 3) and porting it to the clinical domain (see Chapter 4). The architecture of TRIOS also resembles the ones adopted by many other rule-based normalisation systems.

Chang et al. introduced SUTime [30, 29] which uses a rule-based approach over the pre-existing Stanford CoreNLP toolkit. The main strength of this work is the division of the normalisation step into two different phases: representation and resolution. The first one represents temporal expressions as temporal objects easier to map to their logical representation. The next step consists of the application of temporal objects to the specific reference times.

The normalisation problem has also been tackled by using *Probabilistic Context Free Grammars* [8]. Temporal expressions are parsed according to a grammar and a temporal meaning is then associated. This work introduces new types of temporal expressions that partially overlap with the TimeML standard. Using TimEM, the inference component based on the CKY algorithm, it tunes the parameter of the grammar according to a training set. The performance does not make this approach the state-of-the-art, but its results are promising. Its main limitation is the amount of data required to train the grammar parameters properly [9].

Community-driven approaches have also been explored [99]. Llorens et al. built a community-driven tool for gathering hand written rules and accurately evaluate their acceptance. The same authors studied an orthogonal problem: how to automatically generate correctly annotated data from a seed data set [45].

The results for the normalisation phase in TempEval-3 [181] show accuracies ranging from 0.68 to 0.81 (for the VALUE attribute) and 0.86 to 0.94 (for TYPE attribute), which become sensibly lower in the case of clinical data: from 0.54 to 0.73 (for the VALUE attribute) and 0.72 to 0.89 (for TYPE attribute).

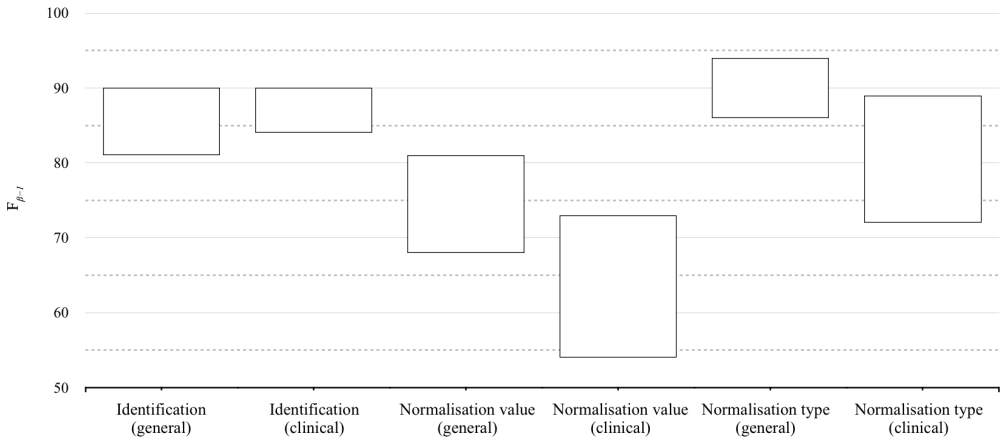


Figure 2.14: Temporal Expression Extraction performance comparison.

The Figure 2.14 summaries the differences in terms of performance for the TEE task.

2.6.2 Event extraction

An event is conventionally referred as an expression in the text that describe *eventuality*. In the general domain case, they are typically referred as inflected verbs and nominal forms (e.g. "[**killed**]$_{\text{EVENT}}$ by the [**crash**]$_{\text{EVENT}}$").

The event extraction phase consists of identifying event mentions in text along with predicting their TimeML attributes (type, polarity, modality, aspect).

The research on event extraction is sensibly smaller than that in TIE, since events are usually easier to identify. The ISO-TimeML standard defines them mostly as verbs. The use of the Part-of-speech (POS) tag is enough to achieve good performance.

The task has been tackled with several different methodologies which follow the same trends introduced in the previous section. Mainly two types of strategies have been followed: rule-based and data-driven.

EVITA [152], part of the TARSQI toolkit [186], is an hybrid architecture. Events represented by verbs are identified using rules, which relying on contextual parsing. Events represented by nouns are instead identified using ML. The grammatical features of the identified events are computed using 140 linguistic rules.

FSS-TimEx [196] uses a finite-state grammar engine with 90 rules based on regular expressions, which is able to predict the event boundaries along with the attribute values. Kolya et al. [91] extended the monolithic approach by splitting the set of rules: one for the boundaries prediction, and one for each attribute to be predicted. The same approach was originally proposed by UzZaman et al. [178].

Data-driven approaches are the ones commonly used to tackle the event extraction, the strategies on the other hand are different. The TIPSem system [101], the best performing system at TempEval-2 challenge, predicts events using ML classifiers. It uses CRFs for the boundaries prediction with an extensive set of linguistic token-level features. On the other hand, the attributes are predicted using features defined at token-level. The idea of using CRFs for event extraction was originally proposed by Bethard [18]. Maximum Entropy classifiers [28, 80], Support Vector Machines [17, 101] and Logistic Regression strategies [17, 89] have also been tested.

Motivated by the fact that events in text are always linked to their arguments, McClosky et al. [111, 110] proposed the use of dependency parsing relations. The identification of events is carried out looking for particular dependency relations between entities in the text, where one of the linked entities is identified as an event. They tested the strategy on the BioNLP‘09 data [87] where the events are defined as biological entities: proteins, transcriptions, gene expressions and related. The dependency relations were used also with Italian [146] and Bulgarian [22] texts.

In the clinical domain, the definition of event proposed as part of the i2b2/12

Shared Task [169] is broader and more articulated. According to the annotation schema proposed, there are different types of event which cover names of clinical departments, treatments, tests, problems and occurrences. Only the latter corresponds to the definition of event in the general domain (according to the ISO-TimeML standard).

The shared task stimulated the research on the event extraction on clinical data and a multitude of different strategies have been examined and benchmarked so far. Interestingly, the problem has never been tackled using just rules [169].

Kovačević et al. [92] proposed an hybrid strategy which tackles each type of event separately. They used a data-driven approach (CRF-based) for every type, except for the clinical departments, which were identified using an ad-hoc gazetteer. The idea of separating the learning task according to the type of event proves to be effective [169, 193, 163]. Grouin et al. [65] experimented with different settings. They tried CRF models by using syntactic and semantic features, with and without the use of an a posteriori filtering component (see Chapter 3).

The Figure 2.15 summaries the differences in terms of performance for the event extraction task.

2.7 Conclusions

TIE is composed of three main sub-tasks: temporal expression, events and temporal relation extraction. This thesis focusses on the first one. In both cases, three main approaches have emerged: rule-based, data-driven and hybrids. The normalisation of temporal expression is the only phase where the use of hand-written or manually curated rules had never find an alternative.

Data-driven approaches, which are those used in this thesis, typically consist of ML components which are able to learn patterns from a pre-processed representa-

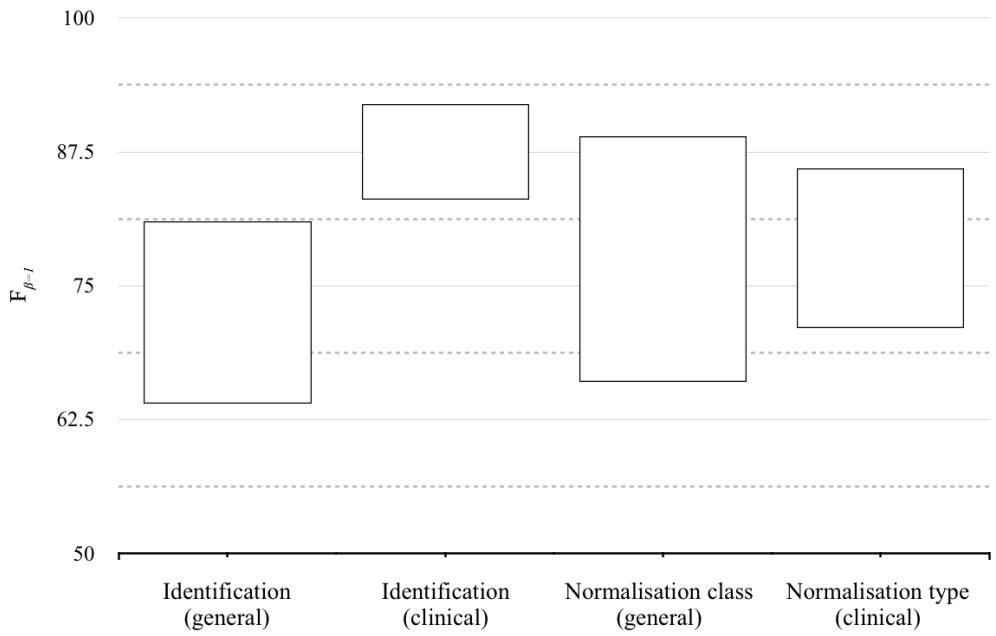


Figure 2.15: Event Extraction performance comparison.

tion of text. The performance of such systems mainly depends on the linguistic features on which the learning phase is based. The literature, as illustrated before, presents a *plethora* of different feature types, but no extensive feature type selection study which discriminates between beneficial and detrimental feature types.

The limit of data-driven approaches when ported to other languages or sub-languages, and how they can be further improved are two important subjects of this thesis. The next chapter (Chapter 3) will present a methodology for TEE in the general domain which enhances the performance of ML-based systems, whereas Chapter 4 will present a similar system tailored on clinical data.

Chapter 3

Temporal expression extraction with extensive feature type selection and a posteriori label adjustment

> *"The Imagination merely enables us to wander into the darkness of the unknown where, by the dim light of the knowledge we carry, we may glimpse something that seems of interest. But when we bring it out and examine it more closely it usually proves to be only trash whose glitter had caught our attention. Imagination is at once the source of all hope and inspiration but also of frustration. To forget this is to court despair."*
> – William Ian Beardmore Beveridge, *The art of scientific investigation*

This chapter is directly adapted from the following paper:

- Michele Filannino and Goran Nenadic. Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. *Data & Knowledge Engineering*, 100:19–33, 2015

which itself builds-on the following papers:

- Michele Filannino, Gavin Brown, and Goran Nenadic. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics

- Michele Filannino. Temporal expression normalisation in natural language texts. *CoRR*, abs/1206.2010, 2012

3.1 Abstract

The automatic extraction of temporal information from written texts is pivotal for many NLP applications such as question answering, text summarisation and information retrieval. It allows filtering information and inferring temporal flows of events. This chapter presents ManTIME, a general domain temporal expression identification and normalisation system. The identification phase combines the use of CRF along with a novel a posteriori label adjustment pipeline, whereas the normalisation phase is carried out using a rule-based approach. Following an extensive review of the feature space, we investigate the performance variation with respect to different models and feature types.

We evaluate six combinations of training data and the a posteriori label adjustment pipeline with respect to the TempEval-3 benchmark test set. The best setting achieved 0.95 precision, 0.85 recall and 0.90 $F_{\beta=1}$ in the identification phase with normalisation accuracies of 0.86 (for type attribute) and 0.77 (for value attribute). Specifically, we show that the use of WordNet-based features in the identification task negatively affects the overall performance, and that there is no statistically significant difference in the results based on gazetteers, shallow parsing and prepositional noun phrase labels used on top of the morpho-lexical features. We also show that the use of silver annotated data (alone or in addition to the human-annotated ones) does not improve the performance.

3.2 Introduction

A temporal expression, also called *timex*, refers to any natural language phrase denoting a temporal entity such as an interval or a time point [52]. For example, in a sentence like *"The Prime Minister said yesterday that the reform promoted three months ago has been very successful."*, the phrases *"yesterday"* and *"three months ago"* are temporal expressions.

```
<?xml version="1.0" ?>
<TimeML xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:noNamespaceSchemaLocation="http://timeml.org/timeMLdocs/TimeML_1.2.1.xsd">
    <DOCID>Example_document</DOCID>
    <DCT>Apr 17, 2012</DCT>
    <TITLE>Example document</TITLE>
    <TEXT>
        The Prime Minister said
        <EVENT eid="e1" class="OCCURRENCE">said</EVENT>
        <TIMEX3 tid="t1" type="DATE" value="2012-04-16">yesterday</TIMEX3>
        that the reform
        <EVENT eid="e2" class="OCCURRENCE">promoted</EVENT>
        <TIMEX3 tid="t2" type="DATE" value="2012-01-16">three months ago</TIMEX3>
        has been very successful.
    </TEXT>
    <MAKEINSTANCE eiid="ei1" eventID="e1" pos="VERB" tense="PAST" aspect="NONE" />
    <MAKEINSTANCE eiid="ei2" eventID="e2" pos="VERB" tense="PAST" aspect="NONE" />
    <TLINK eventInstanceID="ei1" relatedToTime="t1" relType="DURING" />
    <TLINK eventInstanceID="ei2" relatedToTime="t2" relType="DURING" />
    <TLINK eventInstanceID="ei1" relatedToEventInstance="ei2" relType="AFTER" />
</TimeML>
```

Figure 3.1: TimeML annotation of the sentence *"The Prime Minister said yesterday that the reform promoted three months ago has been very successful."* in the TimeML format. The annotation contains: (I) two temporal expressions (*"yesterday"* and *"three months ago"*), (II) two events (*"said"* and *"promoted"*), and (III) three temporal relations (*"said"* $\xrightarrow{during}$ *"yesterday"*, *"promoted"* $\xrightarrow{during}$ *"three months ago"* and *"said"* $\xrightarrow{after}$ *"promoted"*).

Timexes elicit a natural binding between the language and the time domain, making it possible to represent such language expressions as a time point, interval or set.

Temporal expressions can be of three different types [4]: *fully-qualified*, *deictic* and *anaphoric*. A timex is fully-qualified when it unambiguously refers to a precise interval or point in the time domain. For example, the following expressions fall in this category: *"21st July 1985"*, *"31/04/2011 at 12 o'clock"* or *"Martin Luther King's day 2013"*. In the case of deictic expressions, inferring the binding with the time domain necessarily requires to take into account the time of utterance (when the document was written or when the speech was given, often referred to as DCT). Typical deictic temporal expressions include *"today"*, *"yesterday"*, *"last Sunday"* and *"two months ago"*. Finally, anaphoric expressions are a particular case of deictic expressions for which the utterance time varies according to the temporal expressions previously mentioned in the text. Examples of this category are *"that year"*, *"the same week"* or *"the previous month"*.

Research in temporal expression extraction aims at investigating novel and effective approaches to extraction of temporal information from texts. Several scientific challenges [184, 188, 181] have been organized over the years, providing human-annotated data as gold standard to evaluate performance of the state-of-the-art systems.

Early attempts of automatically annotating temporal expressions in texts started in late 1990's [15], and aroused an increasing interest with the proposal of a temporal annotation schemas [104], mainly aiming at enhancing performance of question answering systems. Following the work of Ahn et al. [4], the temporal expression extraction task is now conventionally divided into two main steps: identification and normalisation. In the former step, the effort is concentrated on how to detect the right boundary of temporal expressions in the text. In the

normalisation step, the aim is to interpret and represent the temporal meaning of each pre-identified expression often using the TimeML format [132]. It provides a specification for representing temporal expressions, events and temporal relations (see an example in Figure 3.1). The normalisation task is usually focussed on predicting the two main temporal expressions attributes: TYPE of the temporal expression (e.g. SET, DURATION, DATE or TIME) and its full VALUE according to the ISO 8601 format [75].

In this chapter we introduce ManTIME, a temporal expressions extraction system, where the identification uses machine learning on an extensive set of features and an a posteriori label adjustment pipeline, which further improves the performance. The normalisation phase is carried out by using a set of rules. We evaluated ManTIME on the latest TempEval-3 official benchmark data, achieving 0.95 precision, 0.85 recall and 0.90 $F_\beta = 1$ in the identification phase with normalisation accuracies of 0.86 (for type attribute) and 0.77 (for value attribute).

ManTIME uses 93 features of 4 types, which have been engineered following a systematic review of the scientific literature in temporal information extraction. We explore what categories of feature provides the best performance.

We also investigate the role that *silver training data* have on the performance. Such resources are large automatically generated datasets, which have been created by merging the annotations provided by three state-of-the-art temporal extraction systems [177]. We consider different training scenarios: silver data alone or in combination with gold data, using or not using the a posteriori label adjustment pipeline.

3.3 Related work

The identification step in temporal expression extraction is usually tackled by using machine learning-based approaches. A variety of features have been used such as morphological and dictionary-based. Ahn et al. [4] used morphological features with SVM [21] and CRF [94] showing a notable improvement in performance [5]. Llorens et al. [101, 102] successively added semantic features using a similar architecture. Poveda et al. [130] introduced a sophisticated semi-supervised approach which particularly helped to improve the recall, while Mani et al. [104] used rules learned by a decision tree classifier. Ling and Weld [98] tried Markov Logic Network (MLN) in order to extract temporal relations. Recently, the results from the last temporal information extraction challenge, TempEval-3 [181], show the identification performance ranges from 0.81 to 0.90 in terms of lenient $F_{\beta=1}$ measure (from 0.70 to 0.83 for strict matching).

The second step in temporal expression extraction is the normalisation, which is typically accomplished using rule-based approaches. Grover et al. [67], for example, used regular expression-based rules on top of a pre-existing identification system. UzZaman and Allen [179] developed TRIOS, an open-source rule-based normaliser, focussing on TYPE and VALUE attributes prediction. Llorens et al. [103] extended this architecture making it community-driven: Internet users are allowed to candidate new rules to be integrated in a central rule repository. Angeli et al. [8] proposed a method to learn interpreting temporal representations through the use of a compositional grammar for temporal expressions. To the best of our knowledge, their system is the only piece of research that diverges from rule-based approaches, although the performance is noticeably lower. Recent TempEval-3 normalisation accuracies ranged from 0.68 to 0.86 (for VALUE) and 0.86 to 0.94 (for TYPE attribute) [181].

There are also monolithic temporal expression extraction systems, in which

there is no separation between identification and normalisation. Saquete et al. [151], for example, produced a seminal work proposing a multi-lingual dictionary-based architecture for event ordering, which successively extended into a non-monolithic system [150]. More recently, NavyTime [28] and HeidelTime [165] proposed a set of hand-crafted rules combined with an ad-hoc rule selection algorithm, whereas SUTime [30] used a deterministic rule-based system built on top of the Stanford Core NLP pipeline.

Recently, temporal information extraction aroused increasing interest in the medical domain [176, 169, 92, 16], where temporal information can be used to automatically extract patient clinical histories or temporal cause-effect relations with respect to particular treatments. In the medical domain, the normalisation phase proved to be harder than in the general domain. More specifically, the results from i2b2 2012 [169] show that the identification accuracies range from 0.84 to 0.90, whereas normalisation accuracies rage from 0.54 to 0.73 (for VALUE) and 0.72 to 0.89 for (for TYPE attribute).

While a number of architectures, features and datasets are used for temporal expression extraction, we are not aware of any systematic studies on the types of features that are beneficial for temporal expression extraction, as the effect of different types of training data.

3.4 System architecture

The approach proposed in this paper adopts the dichotomy between identification and normalisation [4], and therefore it consists of two components. The general system architecture is depicted in Figure 3.2. Each step of the architecture will be illustrated in detail in the next sections. For training and testing we mainly used the TempEval-3 datasets as explained in Section 3.5.1.

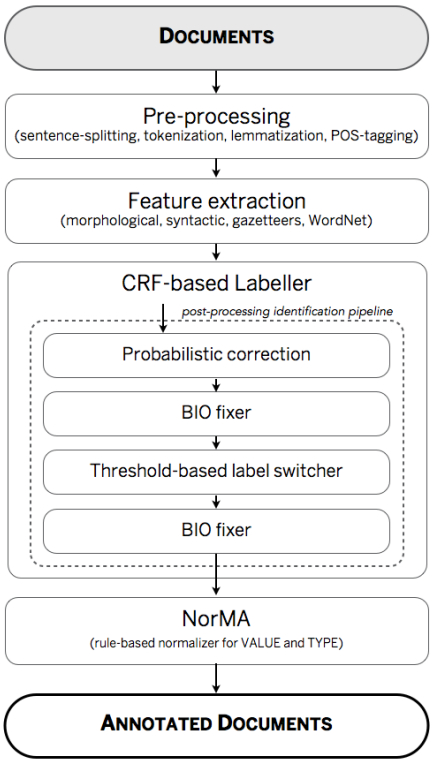


Figure 3.2: ManTIME architecture. Documents are pre-processed using TreeTagger [157], which provides tokens, lemmas and POS-tags. The remaining features are extracted in order to build the token-feature matrix. The machine-learning based labeller predicts a label (B, I or O) for each token and the identification post-processing pipeline is applied. The annotations are finally exported in the TimeML format and for each annotated expression the normalisation component (NorMA) is run.

3.4.1 Temporal expression identification

The identification phase concerns the detection of temporal expressions in the text and the effort is concentrated on predicting their correct boundary or span.

We tackled the identification problem as a sequencing labelling task leading to the choice of CRFs. We trained the system using both human-annotated data and silver data (see Section 3.5.1) in order to investigate the potential contribution of different types of annotated data.

Although the silver data has the advantage of being far larger than the human-annotated data (666K words vs. 95K, see Table 3.6 in Section3.5.1), our hypothesis is that manually-annotated corpora are more accurate (i.e. less noisy), and for this reason are still important in the training phase. Because of this trade-off, we developed a post-processing pipeline on top of the CRFs sequence labeller to boost the identification performance, similarly to the approach proposed by Adafre and de Rijke [2].

Below we describe the CRF-based labeller, the model selection and the post-processing pipeline components in detail.

3.4.1.1 Feature engineering

Temporal expression identification can be seen as a Named Entity Recognition (NER) problem. From this perspective, it is naturally approached as a sequence labelling task, for which we decided to use the Linear Chain Conditional Random Fields (LC-CRFs).

LC-CRFs is a machine learning technique that defines a conditional probability distribution over sequences of input samples taking the following form:

$$P(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y, \mathbf{x})\right) \tag{3.1}$$

where $Z(\mathbf{x})$ is the normalisation factor, K is the number of features, $\mathbf{x}$ represents the sequence of tokens, y represents the sequences of predicted labels, and f_k and λ_k represent the feature function and its weight respectively.

We used the *BIO* format (each token is labelled as being at the *(B)eginning*, *(I)nside* or *(O)utside* of a temporal expression entity) in all the experiments presented here. The factor graph has been generated using the following topology:

$$\begin{aligned} T = \big\{&(w_0), (w_{-1}), (w_{-2}), (w_{+1}), (w_{+2}), (w_{-2} \wedge w_{-1}), (w_{-1} \wedge w_0), \\ &(w_0 \wedge w_{+1}), (w_{-1} \wedge w_0 \wedge w_{+1}), (w_0 \wedge w_{+1} \wedge w_{+2}), (w_{+1} \wedge w_{+2}), \\ &(w_{-2} \wedge w_{-1} \wedge w_0), (w_{-1} \wedge w_{+1}), (w_{-2} \wedge w_{+2})\big\} \end{aligned} \tag{3.2}$$

where w_0 represents the current token, w_{+k} the following and w_{-k} the previous tokens.

In addition to the labelling (or tagging) schemas (*BI*, *BIO*, *BIOE* or *BIOEU*[1]) and the topology of the factor graph, the effectiveness of using CRFs mainly depends on the quality of features.

ManTIME relies on 93 features, which have been collected as a result of a systematic review of the literature in temporal information extraction which we conducted with the aim of exploring the features' contribution. These features belong to the following four disjoint categories.

Morpho-lexical: This set includes the token, its lemma, stem, character pattern (e.g., *"Jan-2003"* is represented as 'Ccc-dddd'), collapsed pattern (e.g., *"Jan-2003"*: 'Cc-d'), first three characters, last three characters, upper first character, word without letters, word without letters or numbers, verb tense and word polarity[2]. For lemma and POS tags we use TreeTagger [157].

[1]The *E* symbol is used with the last annotated token (End), whereas the *U* is used for annotated expressions which contain just one token (Unique).

[2]Opinion Lexicon collected by Hu and Liu: `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

Further, boolean features are included, indicating if the word is lower-case, alphabetic, alphanumeric, titled, capitalized, an acronym (capitalized with dots), number, decimal number, stop-word or has an 's' as last character. Additionally, there are phonetic features and ones specifically crafted to handle temporal expressions in the form of regular expression matching: cardinal and ordinal numbers, times (e.g., *"10:54am"*, *"1 o'clock"*, *"15:19"*), temporal periods (e.g., *"morning"*, *"noon"*, *"nightfall"*), day of the week, seasons, past references (e.g., *"ago"*, *"recent"*, *"before"*), present references (e.g., *"current"*, *"now"*), future references (e.g., *"tomorrow"*, *"later"*, *"ahead"*), temporal signals (e.g., *"since"*, *"during"*), fuzzy quantifiers (e.g., *"about"*, *"few"*, *"some"*), modifiers (e.g., *"approximately"*, *"in the middle"*, *"at the end"*), temporal adverbs (e.g., *"daily"*, *"earlier"*), adjectives, conjunctions and prepositions. A total of 61 morpho-lexical features have been engineered.

Syntactic: Chunks and prepositional noun phrases belong to this category. Both are extracted using the shallow parsing software MBSP[3] and represented in the BIO format.

Gazetteers: The matching of sub-expressions with gazetteer entries is also represented in the BIO format because gazetteers include multi-token entries. We used the following gazetteers: male and female names[4] along with world festivity names[5]. We also used U.S. cities, nationalities and country names from the NLTK[6] corpora. A total of seven gazetteer-based features have been engineered.

[3] http://www.clips.ua.ac.be/software/mbsp-for-python

[4] http://www.cs.man.ac.uk/~filannim/public/male_female_gazetteers.tar.gz

[5] http://www.cs.man.ac.uk/~filannim/public/world_festivals_gazetteer.tar.gz

[6] http://nltk.org/

WordNet: For each token we use the number of senses associated to the word, the first two most common senses, the first four lemmas, the first four entailments for verbs, antonyms, hypernyms and hyponyms. Each of them is defined as a separate feature. A total of 23 WordNet-based features have been engineered. We note that this group of features constitutes an extension of those previously used in the field [71, 100]. In particular, we note that temporal signals (which typically indicate the presence of temporal expressions nearby in text, e.g. *'She slept for just [4 hours]*$_{timex}$.') are known in linguistics to be characterised by having antonyms, whereas the rest of temporal expression words typically do not [138]. We hypothesized that such piece of information should have been integrated to help the machine learning model to highlight temporal expressions.

All the features used in the experiments are presented in Table 3.1 and Table 3.2 with details.

All the experiments have been carried out using CRF++[7] with parameters $C = 1$, $\eta = 0.0001$ and ℓ_2-regularization function.

3.4.1.2 Model selection

The 93 features mentioned above have been combined in four different models combining the following types of features:

- **Model 1**: Morpho-lexical only
- **Model 2**: Morpho-lexical + syntactic
- **Model 3**: Morpho-lexical + gazetteers
- **Model 4**: Morpho-lexical + gazetteers + WordNet

[7] `https://code.google.com/p/crfpp/`

#	Type	Feature	Example
1	M	token (original form)	"Obama" → 'Obama'
2	M	stop-word	"which","he","believes" → 'B', 'B', 'O'
3	M	WordNet lemma	"share","prices" → 'share', 'price'
4	M	TreeTagger lemma	"was" → 'be'
5	M	TreeTagger POS tag	"it","claims" → 'PP', 'VBZ'
6	M	lexical pattern	"12:00Pm" → 'dd:ddCc'
7	M	collapsed lexical pattern	"12:00Pm" → 'd:dCc'
8	M	fist character upper-case	"Manchester" → 'True'
9	M	with digits	"i2b2" → 'True'
10	M	with punctuation symbols	"p.m." → 'True'
11	M	all capital letters and dots	"I.E.E.E.","IEEE" → 'True', 'False'
12	M	all digits and dots	"20.5" → 'True'
13	M	with alpha-numeric characters only	"at","2:00","p.m." → 'True', 'False', 'False'
14	M	with alphabetic characters only	"now" → 'True'
15	M	with decimal characters only	"20" → 'True'
16	M	with digits only (Unicode)	"\u00B2" → 'True'
17	M	lower-case	"car" → 'True'
18	M	numeric	"10" → 'True'
19	M	space(s)	" " → 'True'
20	M	titled	"Europe" → 'True'
21	M	all upper-case characters	"ISO" → 'True'
22	M	ends with an *s*	"textiles" → 'True'
23	M	Lancaster stem	"existing","shareholders" → 'ex', 'sharehold'
24	M	Porter stem	"definitions" → 'definit'
25	M	prefix (first three characters)	"shareholders" → 'sha'
26	M	suffix (last three characters)	"shareholders" → 'ers'
27	M	tense	"Clinton","discussed" → 'none', 'past'
28	M	token with no letters	"8am" → '8'
29	M	token with no letters and numbers	"8am" → 'False'
30	M	non-common word	"and","maiming" → 'False', 'True'
31	M	collapsed vocal pattern	'murder" → 'cvcvc'
32	M	first phoneme	"automobile" → 'AO1'
33	M	phonetic form	"automobile" → 'AO1-T-AH0-M-OW0-B-IY2-L'
34	M	last phoneme	"automobile" → 'L'
35	M	number of phonemes	"automobile" → '8'
36	M	polarity	"will","benefit" → 'neu', 'pos'
37	M*	ordinal number	"first", "second", "third", ...
38	M*	cardinal number + period	"2-year", "3-time", "5-month", ...
39	M*	contains only digits	"2012", "26", "0", ...
40	M*	festival expression	"christmas", "Easter", "thanksgiving", ...
41	M*	temporal future trigger	"next", "tomorrow", "coming", ...
42	M*	temporal fuzzy quantifier	"approximately", "few", "several", ...
43	M*	literal number	"zero", "three", "fourteen", ...
44	M*	temporal modifier	"end", "start", "beginning", ...
45	M*	month	"January", "sep", "february", ...
46	M*	ordinal number in digits	"15th", "100th", "1st", ...
47	M*	ordinal trigger	"st", "rd", "th", "nd"

Table 3.1: Features used in the experiments (first part). Type column indicates whether a feature belongs to the (M)orpho-lexical, (S)yntactic, (G)azetteer or (W)ordNet category. Regular expression-based features, denoted with an *, are presented with a list of matching expressions whereas for the rest of them the notation (tokens → values) has been used. Feature #15, #16 and #18 are computed using the Python 2.x built-in operators. #23 uses the Lancaster Stemmer [122], #24 uses the Porter Stemmer [128]. #37 and #38 are computed at token-level.

#	Type	Feature	Example
48	M*	temporal past trigger	“ago”, “earlier”, “previous”, ...
49	M*	temporal period	“centuries”, “week”, “hour”, ...
50	M*	part of the day	“morning”, “night”, “sunrise”, ...
51	M*	temporal present trigger	“tonight”, “current”, “nowadays”, ...
52	M*	season	“winter”, “Summer”, “springs”, ...
53	M*	temporal signal	“on”, “during”, “for”, ...
54	M*	temporal adjective	“soon”, “late”, “fiscal”, ...
55	M*	temporal adverb	“daily”, “early”, “lately”, ...
56	M*	temporal entity	“period”, “course”, “age”, ...
57	M*	temporal conjunction	“until”, “while”, “when”, ...
58	M*	temporal prepositions	“pre”, “mid”, “over”, ...
59	M*	time	“11:15am”, “12.23p.m.”, “8:00 pm.”, ...
60	M*	weekday	“Monday”, “tuesday”, “Thu”, ...
61	M*	year	“1996”, “2013”, “’50”, ...
62	G	gazetteer of country names	“from”,“United”,“Kingdom” → ‘O’, ‘B’, ‘I’
63	G	gazetteer of female names	“to”,“Marie”,“Claire” → ‘O’, ‘B’, ‘I’
64	G	gazetteer of world festivals	“Christmas” → ‘B’
65	G	gazetteer of country ISO names	“Italy” → ‘B’
66	G	gazetteer of male names	“Michele”,“and” → ‘B’, ‘O’
67	G	gazetteer of nationalities	“Britain” → ‘B’
68	G	gazetteer of U.S. cities	“Springfield”,“in” → ‘B’, ‘O’
69	S	lexical chunk	“an”,“offer”,“from” → ‘B-NP’, ‘I-NP’, ‘O’
70	S	prepositional noun phrase	“with”,“a”,“fork” → ‘B-PNP’, ‘I-PNP’, ‘I-PNP’
71	W	first sense	“chief” → ‘Synset(’head.n.04’)’
72	W	second sense	“chief” → ‘Synset(’foreman.n.01’)’
73	W	first antonym	“including” → ‘Lemma(’exclude.v.02.exclude’)’
74	W	second antonym	“including” → ‘Lemma(’exclude.v.03.exclude’)’
75	W	third antonym	“including” → ‘None’
76	W	fourth antonym	“including” → ‘None’
77	W	first entailment	“pay” → ‘Synset(’pay.v.01’)’
78	W	second entailment	“pay” → ‘Synset(’choose.v.01’)’
79	W	third entailment	“pay” → ‘None’
80	W	fourth entailment	“pay” → ‘None’
81	W	first hypernym	“six” → ‘Synset(’die.n.01’)’
82	W	second hypernym	“six” → ‘Synset(’digit.n.01’)’
83	W	third hypernym	“six” → ‘Synset(’domino.n.04’)’
84	W	fourth hypernym	“six” → ‘Synset(’spot.n.13’)’
85	W	first hyponym	“started” → ‘Synset(’attack.v.05’)’
86	W	second hyponym	“started” → ‘Synset(’recommence.v.01’)’
87	W	third hyponym	“started” → ‘Synset(’auspicate.v.02’)’
88	W	fourth hyponym	“started” → ‘Synset(’inaugurate.v.03’)’
89	W	first lemma	“ground” → ‘ground’
90	W	second lemma	“ground” → ‘dry_land’
91	W	third lemma	“ground” → ‘reason’
92	W	fourth lemma	“ground” → ‘land’
93	W	number of senses	“hold” → ‘45’
94	-	LABEL	“during”,“March” → ‘O-TIMEX3’,‘B-TIMEX3’

Table 3.2: Features used in the experiments (second part). Type column indicates whether a feature belongs to the (M)orpho-lexical,(S)yntactic, (G)azetteer or (W)ordNet category. Regular expression-based features, denoted with an *, are presented with a list of matching expressions whereas for the rest of them the notation $(tokens \rightarrow values)$ has been used. The WordNet-based features are computed from the TreeTagger lemma of each token. No word-sense disambiguation algorithm has been used.

We performed an extensive evaluation by repeating the experiments a number of times and assessing whether there is any statistical difference among the models. This allowed us to select the model that provides the highest $F_{\beta=1}$ score among the four proposed.

All the data provided by TempEval-3 (see Table 3.6), except for the TempEval-3 official benchmark test set, have been merged, shuffled at sentence level (seed = 490) and split into two sets: 80% as a training set and 20% as a test set. The training set has been shuffled 5 times, and for each of these, the 10-fold cross validation technique has been applied.

Table 3.3 shows the post-hoc ANOVA analysis and Figure 3.3 shows the box-plot comparison of the models ($F_\beta = 1$ measure). The analysis is statistically significant ($p = 0.0054$ with ANOVA test) and provides two important outcomes:

1. There is no statistically significant difference among the first three models (see Table 3.3), despite the presence of apparently important and computationally expensive information such as chunks, prepositional noun phrases and gazetteers.

2. the set of WordNet-based features negatively affects the overall classification performance, as already noticed in the literature [145]. This is mainly due to the sparseness of the labels: many tokens do not have any associated WordNet sense.

By virtue of this analysis, we opted for the smallest feature set, Model 1, which has two positive consequences: to help mitigate overfitting due to the smaller feature space, and reducing the computational cost of the system.

In order to get an educated estimation of the Precision/Recall performance of the selected model in the wild, we then trained it on the entire training set and

	Model 2	Model 3	Model 4
Model 1	0.994	0.151	$2.16E^{-9*}$
Model 2	-	0.267	$4.00E^{-10*}$
Model 3	-	-	$2.75E^{-10*}$

Table 3.3: Post-hoc ANOVA analysis of the models ($F_\beta = 1$ measure): p-values of two-tailed paired T-tests for each pair of models. Small p-values indicate statistical significance. Pairs of models denoted with * have a statistically significant difference. Model 4 is significantly worse than the rest of the models. At the same time, there is no statistically significant difference among the first three models.

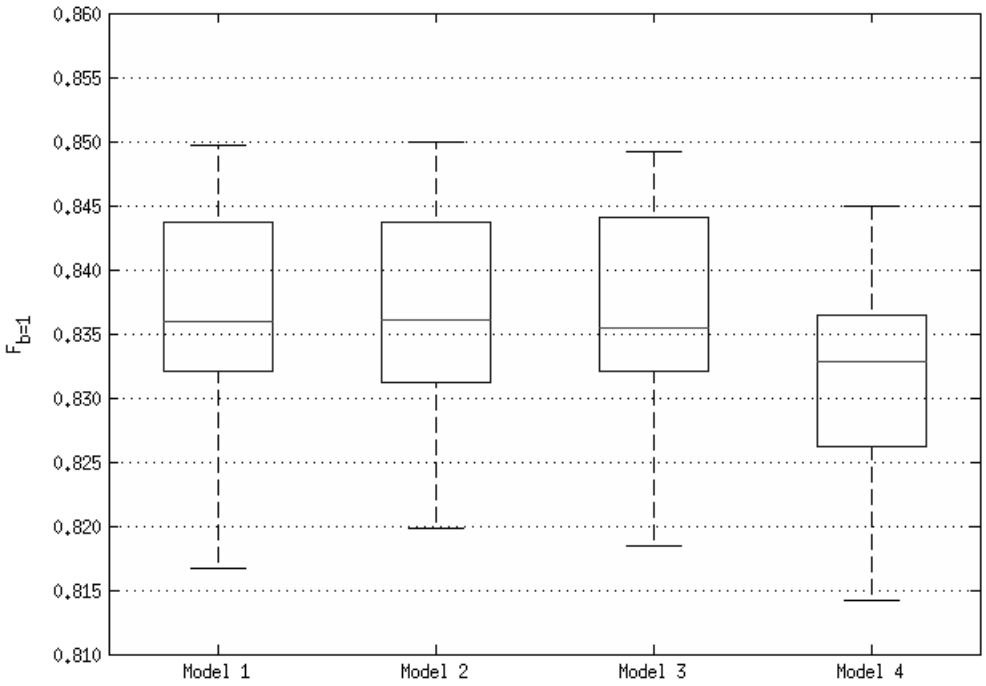


Figure 3.3: $F_\beta = 1$ measure across the four models. 5x10-fold cross validated. The box indicates the upper/lower quartiles, the horizontal line inside each of them shows the median value, while the dotted crossbars indicate the maximum/minimum values. There is no significant difference among the first three models, whereas the last one is statistically worse than the rest.

tested it against the test set. The results for all the models are shown in Table 3.4. Model 1 showed a slightly better $F_{\beta=1}$ score, which corroborated our choice.

The models used for the final evaluation of the TempEval-3 benchmark data have been trained using all the data, except for the ones in the benchmark data set.

3.4.1.3 A posteriori label adjustment pipeline

Although the CRF-based labeller already provided reasonable performance on the training data, equally balanced in terms of precision and recall, we focussed on boosting the baseline performance through a post-processing pipeline composed of three modules, which aimed to adjust the CRF-predicted labels.

Probabilistic correction module: We noticed that the CRF-based labeller tends to assign labels with high confidence even for ambiguous tokens. We therefore aimed to design a module that would make predictions less strict and in some cases have the effect of changing the most likely label (mainly expected to bring an improvement in terms of recall).

For each token, we thus average the conditional probabilities from the trained CRF model with the prior probabilities extracted from the gold data only (see Section 3.5.1 for details about data).

	Precision	Recall	$F_{\beta=1}$
Model 1	83.20	**85.22**	**84.50**
Model 2	**83.57**	85.12	84.33
Model 3	83.51	85.12	84.31
Model 4	83.15	84.44	83.79

Table 3.4: Estimation of the expected results for the benchmark. Precision, Recall and $F_{\beta=1}$ score have been computed using strict matching. Model 1 performed slightly better with respect to $F_{\beta=1}$.

For each token w in the gold data, we extracted the conditional probability $P(L|w)$, where $L = \{`B', `I', `O'\}$. The probabilities have been estimated using frequencies. The list of tokens taken into account has been restricted to those appearing within temporal expressions at least twice. This process allowed us to obtain the prior label probabilities. For example, $P(B|Monday) = 0.97$, $P(I|Monday) = 0.03$ and $P(O|the) = 0.95$.

From the CRF-based labeller we extracted, for each token, the internal conditional probability of each label. The two probabilities (from the gold data and the CRF) were then averaged for every label of each token.

An example is given in Table 3.5.

Threshold-based label switcher: Some tokens have a high a priori probability of being part of a temporal expression (e.g., *"Monday"* or *"today"*). However, some of these tokens might have been erroneously labelled as 'O' by the CRF labeller.

This module changes the predicted label to the most likely one based on the a priori probabilities from the gold data only. This is triggered only when the prior probability of a certain label in the gold data is greater than a given threshold. Therefore, the application of this module forces the prior

	$P(O)$	$P(I)$	$P(B)$
CRFs probabilities	**0.526**	0.004	0.470
Gold probabilities	0.000	0.063	**0.937**
Result	0.263	0.033	**0.704**

Table 3.5: Probabilities updated for the token *'Saturday'* in the sentence *"Northern Ireland's World Cup qualifier with Russia has been postponed until Saturday due to heavy snow"*. The predicted label changes from the 'O' (predicted by CRF) to 'B'.

probabilities extracted from the human-annotated data. Through repeated empirical experiments on a small sub-set of the training data, we found an optimal threshold value (0.87).

BIO fixer: Although CRFs are designed to handle sequences, they assign labels token-by-token. This leads to possibly inconsistent sequences of labels[8]. For the BIO labelling schemas, the only possible source of inconsistency is the sequence *O-I*, as there should be a 'B' between them. We found that, among the possible corrections (B-I or I-B), *B-I* applies to most cases (i.e. the first token has been most often incorrectly annotated). For example, *"Three/O days/I ago/I ./O"* should be converted into *"Three/B days/I ago/I ./O"*.

We also merged adjacent expressions such as *B-B* or *I-B*, because different temporal expressions are always divided at least by a symbol or a punctuation character (e.g. *"Wednesday/B morning/B"* becomes *"Wednesday/B morning/I"*, *"21st/B November/I 1990/B"* becomes *"21st/B November/I 1990/I"*).

We performed an extensive evaluation of the possible label adjustment pipeline configurations, which has been carried out with 5x10-fold cross validation (as described in Section 3.4.1.2). The results are presented in Figure 3.4. The first configuration corresponds to the CRFs only. All the differences among the settings are statistically significant (measured with ANOVA test). Using the pipeline always leads to an improvement in performance, with the BIO fixer component as the major contributor. The optimal pipeline configuration provides a 2.76% averaged statistically significant increment (with respect to the strict $F_{\beta=1}$ scores of the CRF model) and is composed of:

1. Probabilistic correction module

[8]This could have been avoided by using CRFs toolkits other than CRF++ [159]

2. BIO fixer
3. Threshold-based label switcher
4. BIO fixer

3.4.2 Normalisation

The normalisation phase aims to interpret and represent the temporal meaning of each pre-identified expression using the TimeML format [132]. Two attributes are particularly important in this respect: TYPE and VALUE. The first one can be either 'DATE', 'TIME', 'DURATION' or 'SET'. The second one expresses the ISO 8601 representation of each expression.

The proposed temporal expression normalisation approach is based on rules and it extends TRIOS [179]. TRIOS' input is the temporal expression and the utterance time (Document Creation Time) and its rules have the form of dictionary-driven regular expressions in a switch architecture: the activation of one of them excludes the activation of the remaining ones.

Our normalization system, called NorMA (depicted in Figure 3.5), is composed of three modules: pre-processing rules, extension rules and post-manipulation rules.

Pre-processing rules: This set of rules has been introduced to turn recognised temporal expressions into a more suitable form for normalisation. Some examples from this rule set are: determiners removal (e.g., *"the day after"* $\rightarrow$ 'day after'), misspelling correction (e.g., *"wendsday"* $\rightarrow$ 'Wednesday'), and lower-case and trimming transformation (e.g., *" every Friday morning ."* $\rightarrow$ 'every friday morning').

Extension rules: The extension rules are new rules that cover temporal expressions not handled by TRIOS. Such rules are matched before the TRIOS'

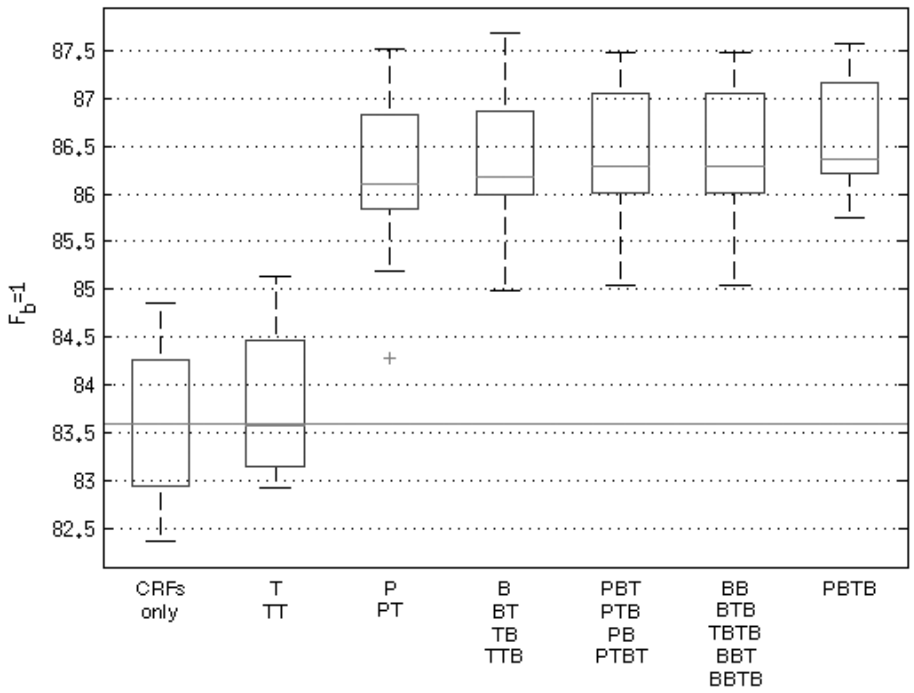


Figure 3.4: Analysis of different post-processing pipeline configurations (with respect to the $F_\beta = 1$ measure). 5x10-fold cross validated. *P* stands for *Probabilistic Correction Module*, *B* for *BIO-fixer* and *T* for *Threshold-based label switcher*. All the differences among the settings are statistically significant (measured with ANOVA test). The configurations have been collapsed when they provided the same result. The box indicates the upper/lower quartiles, the horizontal line inside each of them shows the median value, while the dotted crossbars indicate the maximum/minimum values. The horizontal line is the median of the configuration without pipeline.

ones. Examples of those are duration expressions (e.g. *"3-year"*, *"3-day"*), frequency expressions (e.g. *"every half an hour"*, *"every two days"*) or period expressions (e.g. *"'90s"*, *"eighties"*).

Post-manipulation rules: The post-manipulation rules are mainly used to validate the syntax of the predicted VALUE attribute and to normalise frozen expressions transformed by the previous groups of rules. For example, some of the rules are used to normalise expressions of festivity dates such as *"Queen's birthday"* or *"Saint Patrick's day"*.

Overall, NorMA extends TRIOS with 40 new rules: 16 pre-processing rules, 20 extension rules, and 4 post-manipulation rules (see Appendix A). The system has already been proven to provide statistically better performance with respect to TRIOS and consequently state-of-the-art performance against the TempEval-2 benchmark test set [54].

3.5 Experiments and Results

In this section we present the experiments performed. In particular, we describe the data, the evaluation metrics and the results. Also the findings of the error analysis are presented in order to investigate the system annotation errors.

3.5.1 Data

The human-annotated data come from two existing corpora: AQUAINT and TimeBank[9]. Both data sets have been revised by the TempEval-3 organizers in order to fix erroneous annotations. These two corpora have been used for training purposes

[9]Both corpora are available at `http://www.cs.york.ac.uk/semeval-2013/task1/index.php?id=data`

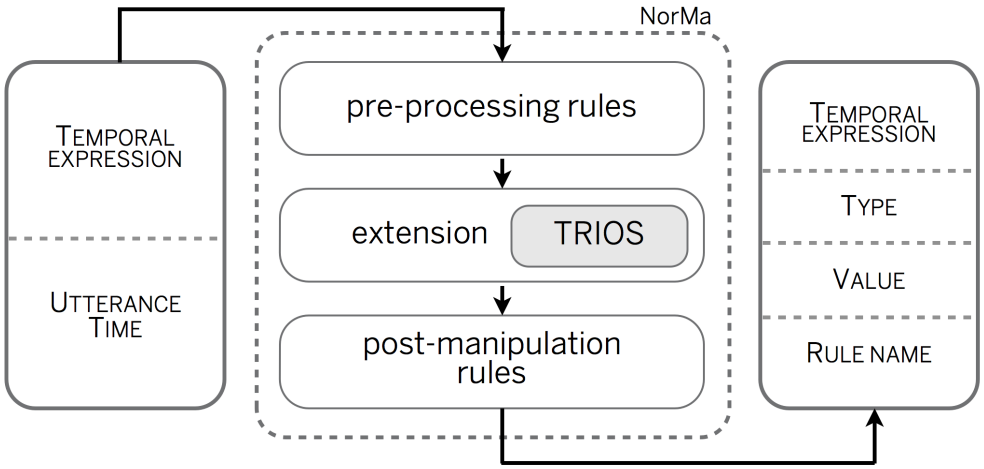


Figure 3.5: NorMA architecture diagram. Each pre-identified temporal expression, along with the document creation time, is pre-processed and then subjected to rules matching. Post-manipulation rules are activated to cope with exact matchings like season or festival names.

as opposed to a human-annotated corpus, *TempEval-3 benchmark*, which has been used as a test set.

In addition, for training we used the TempEval-3 silver corpus, which has been made by merging, through an ad-hoc algorithm [177], three state-of-the-art temporal extraction systems: TIPSem, TipSEM-B [102] and TRIOS [179]. This corpus is much larger than the gold ones, although its annotations are not as reliable. Table 3.6 summarises the main characteristics of each corpus.

Every document has been annotated using the TimeML standard and released with its DCT. Each annotated temporal expression carries its TYPE and VALUE attributes.

Corpus	# docs	# sentences	# words	# timexes	annotation	used for
AQUAINT	73	956	33973	652	gold	training
TimeBank	183	2624	61418	1426	gold	training
TempEval-3 silver	2452	12692	666309	12739	silver	training
TempEval-3 benchmark	20	219	6375	158	gold	test

Table 3.6: Corpora used in the experiments. The final column indicates how each corpus has been annotated: *gold* means annotated by human experts, whereas *silver* means generated by automatic systems.

3.5.2 Evaluation metrics

The identification phase (prediction of the temporal expression boundaries) has been evaluated using *Precision*, *Recall* and $F_{\beta=1}$ measure, according to the following formulae:

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

$$F_{\beta=1} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.5}$$

where *TP*, *FP* and *FN* stand for the number of true positive, false positive and false negative examples respectively.

Precision, Recall and $F_{\beta=1}$ measures are computed according to two different definitions of matching: strict and lenient, following TempEval-3 [181]. The strict matching considers a predicted boundary correct only if it strictly matches the gold boundary, whereas the lenient matching considers a predicted boundary correct as long as it overlaps with the gold one.

The performance of the normalisation task is measured on two temporal attributes: TYPE and VALUE (ISO 8601 representation). What is measured here is the prediction accuracy of the correctly identified temporal expressions only, according

to the following formulae:

$$Type_{accuracy} = \frac{\#\ correct\ types}{\#\ identified\ temporal\ expressions} \tag{3.6}$$

$$Value_{accuracy} = \frac{\#\ correct\ values}{\#\ identified\ temporal\ expressions} \tag{3.7}$$

The type of each temporal expression can be inferred from the VALUE attribute. Consequently, the overall score for temporal information extraction is computed using the following formula (also used at TempEval-3):

$$Score = \tilde{F}_{\beta=1} * value_{accuracy} \tag{3.8}$$

where $\tilde{F}_{\beta=1}$ denotes the lenient matching measure [181].

3.5.3 Results

Six different experimental settings have been evaluated as combinations of different training sets (gold, silver, gold&silver) with or without the application of label adjustment pipeline. The results are shown in Table 3.7 where the *overall score* is computed by Formula 3.8. We point out that the setting #4 was submitted as an official submission for the TempEval-3 challenge (Task A identification and normalization of temporal expressions) and has been ranked 5th out of 21 submitted runs, as the best performing machine learning-based system.

All the settings showed high precision (strict ranging from 0.76 to 0.82, lenient ranging from 0.87 to 0.92) and reasonable coverage (strict ranging from 0.63 to 0.70, lenient ranging from 0.79 to 0.85) in the identification stage. This indicates the fact that the system has partially generalised from the training data.

The training of the system by using the gold data only combined with the use of the label adjustment pipeline proved to be the best overall result, although not

#	Training data	IDENTIFICATION						NORMALIZATION		Overall
		strict matching			lenient matching			accuracy		
	(label adjustment)	Pre.	Rec.	$F_{\beta=1}$	Pre.	Rec.	$\tilde{F}_{\beta=1}$	TYPE	VALUE	score
1	Gold&Silver (✕)	0.79	0.64	0.70	0.97	0.79	0.87	0.89	0.77	0.672
2	Gold&Silver (✓)	0.80	0.66	0.72	0.97	0.80	0.88	0.87	0.76	0.667
3	Gold (✕)	0.76	0.64	0.70	0.95	0.80	0.87	0.87	0.77	0.675
4	Gold (✓)	0.79	**0.70**	**0.74**	0.95	**0.85**	**0.90**	0.86	0.77	**0.690**
5	Silver (✕)	0.78	0.63	0.70	0.97	0.80	0.87	0.89	0.77	0.672
6	Silver (✓)	**0.82**	0.66	0.73	**0.98**	0.79	0.88	**0.91**	**0.78**	0.683

Table 3.7: Performance on the TempEval-3 official benchmark test set. The use of the label adjustment pipeline (highlighted with the symbol ✓ as opposed to ✕ which means the pipeline has not been applied) always improves the $F_{\beta=1}$ score (both strict and lenient matching). The normalisation phase proves to be agnostic with respect to the identification precision.

leading to the highest normalisation accuracy. Somewhat surprisingly, the use of the silver data did not improve the performance, neither when used alone nor in addition to the gold data (regardless of the label adjustment usage).

The *a posteriori label adjustment pipeline* showed the highest precision when applied to the silver data only. In this case, the pipeline acted as a reinforcement of the human-annotated data, helping improving the boundaries. As expected, the post-processing pipeline boosted the performance of both precision and recall. Still, we note the best improvement with the human-annotated data.

We also investigated the contribution of each component in the label adjustment pipeline with respect to the test set. Figure 3.6 shows the results. The probabilistic correction module negatively affects the performance (making less strict predictions) although its output is then corrected by the use of the BIO fixer module. The threshold-based label switcher introduces an equal number of false and true positives. False positives are always 'I' labels, which are then propagated by the next component in the pipeline, BIO-fixer, by adding a 'B' label to the previous tokens. This explains the slight downward trend visible in the last step of Figure 3.6. The limited size of the TempEval-3 benchmark test set, on which this analysis is based, might not be enough to explain this behaviour. Therefore, the effect should be taken with caution.

The normalisation task proved to be challenging. Among the correctly typed temporal expressions, there was still about 10% for which an incorrect value is provided (VALUE ranges from 0.76 to 0.78).

3.5.4 Error analysis

The analysis of the predicted annotations against the gold ones allows us to pinpoint errors both in identification and normalisation phase. We analysed the errors in the experimental setting #4.

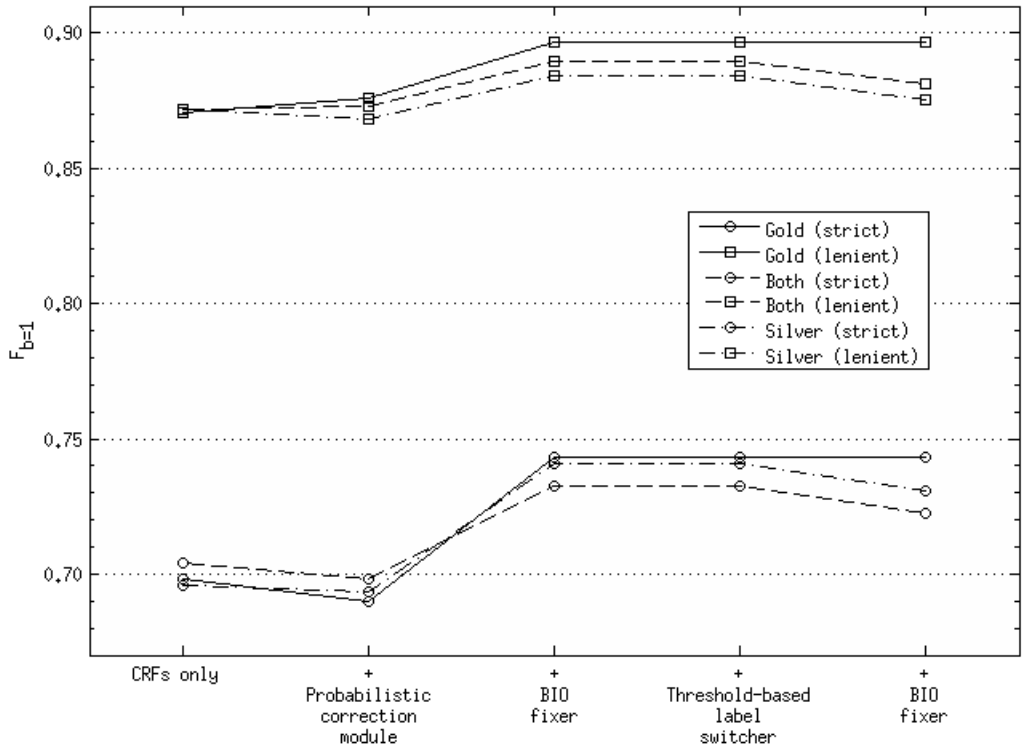


Figure 3.6: Analysis of the a posteriori label adjustment pipeline components. The upper group of curves refers to the lenient matching, whereas the bottom refers to the strict matching. Every component on the x-axis is applied on top of the previous ones.

3.5.4.1 Identification errors

The system correctly identified the majority of temporal expressions annotated in the test set, and incorrect annotations are mainly due to specific limitations of the system in addition to some issues in the gold standard data.

Examples of false positives (incorrectly recognised expressions) due to the CRF model are *"of flu"* and *"and"*. Those expressions have been wrongly classified and the post-processing pipeline has not been able to discard them from the predictions. This is due to a very high confidence from the CRF module.

We noticed a significant amount of partial errors mainly due to errors in the tokenization phase. For example, in *"early 2012."* and *"2007."* the full stop should have been removed, whereas *"2009-2010"* should have been split in three different tokens. It appears that wrong tokenisation is the major cause of the difference between strict and lenient performance. In few cases, the system excluded modifiers (e.g., *"late"* in *"late last July"*) or signals (e.g., *"every"* in *"every morning"*) at the beginning (or at the end) of the expressions, leading to false negatives. Those errors are due to the CRF model which discarded such words with very high confidence.

These results suggests that reducing the complexity of the CRFs factor graph (see Section 3.4.1.1) and using a better tokeniser may lead to better performance.

False negatives (missed temporal expressions) are also connected to the low frequency of some types of expression in the training data: *"15:00GMT Saturday"*, *"a mere 24 hours"*. We also noticed cases of false negatives due to rare surrounding morphological contexts in the training data.

In three cases (2%) out of a total of 138 temporal expressions, the errors are due to questionable human annotations in the test set: *"digital"* alone (in the expression "digital age"), *"tenure"* and *"second term"*. In five cases (4%), the system correctly annotates expressions missed by the human annotators (e.g., *"the next decade"* or *"every morning"*).

3.5.4.2 Normalisation errors

The normalisation error analysis has been carried out on the correctly identified temporal expressions and it consists of checking whether the content of the VALUE and TYPE attributes are equal to the ones provided by the human annotators.

A total of 33 temporal expressions have been correctly identified but wrongly normalised (*VALUE*). The major source of error (18/33 cases: 55%) remains the normalisation of partially extracted temporal expressions (e.g., *"100"* instead *"100 days"*, or *"a mere 24"* instead *"a mere 24 hours"*). In eight cases (24%), the normaliser failed to correctly distinguish between dates and durations (e.g., *"the 99th day"* was normalised as a duration of 99 days, instead of a precise day), whereas in five (15%) it failed to detect the right orientation in time (future or past), leading to the choice of a wrong year (e.g., *"early August"* normalised as *"2013-08"* instead of *"2012-08"*).

We found only one (3%) possibly wrongly annotated temporal expression in the benchmark test set, i.e. for the expression *"20th Century"*, a value *"19"* was provided instead of *"19XX"*. In another case, the expression *"a decade"* was normalised with *"P10Y"* instead of the more correct *"P1E"*. In both cases the normaliser provided the right value, although these were considered errors.

3.6 Conclusions

This paper has presented a novel architecture for temporal information extraction (identification and normalisation) of texts from general domain with an extensive feature type selection. We also described the results with respect to the TempEval-3 benchmark test set and the error analysis for both identification and normalisation phases.

3.6.1 Summary of contributions

In summary, the contributions of this paper are:

- We conducted an extensive evaluation of the feature space and training configurations, which, to the best of our knowledge, has never been done before in the context of temporal expression extraction. The results indicate the key importance of morpho-lexical features to the detriment of syntactic features, as well as gazetteer and WordNet-related ones. In particular, while syntactic and gazetteer-related features do not affect the performance, WordNet-related features appear not to have positive impact. This conclusion, although statistically significant, is necessarily limited by the fact that the features analysis strictly depends on the way previous work has used WordNet. It does not mean that there is not a different way of using WordNet which may positively contributing to the temporal expression identification. Also, the feature analysis is meant to be relevant only in the temporal information extraction context. We do not suggest that some of the features experimented with here will produce the same effects in a different NER task.

- We designed and built an automatic a posteriori label adjustment pipeline on top of the CRF module which we show to provide statistically significant positive impact on the results. We have also investigated the contribution of different possible configurations. Somewhat surprisingly, the use of the label adjustment pipeline, originally introduced mainly to be used with models trained on silver data, proved its efficacy with the gold data too. We provided an extensive statistical analysis on the a posteriori label adjustment pipeline which sheds light on the contribution of each pipeline component in isolation and in the context of others. The experiments also proved its use to be promising for both precision and recall enhancement.

- Furthermore, we found out that the use of silver data does not improve the performance, although we consider the benchmark test set arguably too small to made this conclusion generalisable.

3.6.2 Future work

The a posteriori label adjustment pipeline proved to be promising and it constitutes, *de facto*, a novel approach to temporal expression extraction. We believe it can be improved from many aspects, including:

- Using the N most likely predicted sequences from the CRFs-based labeller in order to discriminate the most ambiguous/difficult tokens.

- Using the rules from the normaliser in order to enhance the accuracy of the identification phase: discarding identified expressions not recognised by the normaliser (false positives reduction) and adding expressions recognised by the normaliser but ignored by in the identification phase (increment of true positives).

Our other future work will focus on the investigation of *local semantics* representation for temporal expressions [107]. This representation provides a way to separate the temporal expressions' semantics from the contextual information.

To aid replicability of this work, the source code of the entire system, the machine learning pre-trained models, the statistical validation details and an online demo are available at:
`http://www.cs.man.ac.uk/~filannim/mantime.html`

Acknowledgements

The authors wish to express their gratitude to Dr. G. Brown (The University of Manchester) who, by his advice about statistical validation and model selection, aided greatly in this work. We thank Marilena Di Bari (University of Leeds) for proofreading the manuscript. We also would like to thanks the reviewers for their efforts to suggest improvements to the paper. Finally, we thank the organizers of TempEval-3 for the data. This work was supported by a doctoral training grant from the UK Engineering and Physical Science Research Council.

Chapter 4

Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives

"Learn what is to be taken seriously and laugh at the rest."

– Herman Hesse, *Steppenwolf*

This chapter is directly adapted from the following journal paper:

- Aleksandar Kovačević, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, 20(5):859–866, 2013

The author of this thesis developed the clinical temporal expression normalization component, performed the error analysis and drafted the manuscript.

4.1 Abstract

Objective: Identification of clinical events (eg, problems, tests, treatments) and associated temporal expressions (eg, dates and times) are key tasks in extracting and managing data from electronic health records. As part of the i2b2 2012 Natural Language Processing for Clinical Data challenge, we developed and evaluated a system to automatically extract temporal expressions and events from clinical narratives. The extracted temporal expressions were additionally normalized by assigning type, value, and modifier.

Materials and methods: The system combines rule-based and machine learning approaches that rely on morphological, lexical, syntactic, semantic, and domain-specific features. Rule-based components were designed to handle the recognition and normalization of temporal expressions, while conditional random fields models were trained for event and temporal recognition.

Results: The system achieved micro F1 scores of 90% for the extraction of temporal expressions and 87% for clinical event extraction. The normalization component for temporal expressions achieved accuracies of 84.73% (expression's type), 70.44% (value), and 82.75% (modifier).

Discussion: Compared to the initial agreement between human annotators (87-89%), the system provided comparable performance for both event and temporal expression mining. While (lenient) identification of such mentions is achievable, finding the exact boundaries proved challenging.

Conclusions: The system provides a state-of-the-art method that can be used to support automated identification of mentions of clinical events and temporal expressions in narratives either to support the manual review process or as a part of a large-scale processing of electronic health databases.

4.2 Background

Recent advances in the availability of Electronic Health Records(EHRs) provide an opportunity to improve the quality of clinical care (eg, through large-scale data sharing and integration that can be used to build clinical decision support systems [189]) and to support medical research (e.g., identification of patients with specific conditions to support clinical trials [113]). While key issues remain in the adoption of EHRs and in managing data confidentiality [51], automated processing of available clinical data is also a major challenge: manual identification of such information is time consuming and often inconsistent and incomplete [121]. This is particularly the case with clinical narratives, which are often the primary, preferred, and richest source of patient information. Several efforts have been reported in the area of clinical text mining to bridge the gap between unstructured clinical notes and structured data representation [148, 154], including tools such as MetaMap [12, 11], and KnowledgeMap [43] that have been developed to automatically annotate medical concepts in free text, along with systems to identify patient disease status [174, 194], medication information [175, 176], etc.

The i2b2 Natural Language Processing for Clinical Data challenge series provides a framework for common evaluation of clinical text mining systems. The topic of the 2012 challenge was the identification and linking of mentions of Temporal Expressions(TEs) (eg, dates, times, durations, and frequencies) and clinically relevant events (eg, patient's problems, tests, treatments) in narratives [169].

Extraction of clinical events has recently attracted considerable attention, and was, for example, one of the tasks in the i2b2 2010 challenge [176]. A wide variety of approaches (semi-supervised [41], supervised [68, 126], hybrid models [77, 82]), features (orthographic, lexical, morphological, contextual, semantic), terminological resources (Unified Medical Language System (UMLS) [97, 197],

MedDRA [114], DrugBank [191]), and heuristic post-processing methods [77] were used. The best lenient F1-score for the extraction of clinical events ranged from 89.80% [82] to 92.40% [41].

On the other hand, previous research on TE extraction has been mainly focused on the general domain [184, 188]. The clinical domain has been considered only relatively recently and often as an extension of general systems. For example, Med-TTK [142] is built on top of a newswire system (TTK, TARSQI Toolkit [187]) by modifying and expanding rules developed on a set of 200 clinical narratives. The system identifies mentions of date, time, duration, and frequency TEs (with an overall F1 score of 85%), but does not provide their normalized values.

In this paper we describe, discuss, and evaluate a system that we have developed as part of our contribution to the i2b2 2012 challenge.

4.3 Objective

The aim of the 2012 challenge was to create clinical patient timelines from a set of clinical narratives. The TE extraction task focused on recognition and normalization of TE mentions. Normalization involved assigning three attributes:

- *value*, using the ISO 8601 representation (eg, "2012-10-31T09:00")
- *type* of the TE: Time (eg, "the morning of admission"), Date ("15 May 2007"), Duration ("4 minutes"), or Frequency ("PT48H")
- *modifier* that may be associated with the TE (eg, "approx").

The event extraction task included recognition of instances of PROBLEM (eg, "hematoma"), TEST (eg, "an echocardiogram"), and TREATMENT (eg, "heparin IV") events. In addition, events included mentions of a CLINICAL DEPARTMENT

(eg, "the ER"). Mentions that indicate an evidential source of some specified information (eg, "CT [shows]," "the patient [complained]") are considered EVIDENTIAL events (note that these can be verbs). Occurrences of all clinically relevant events that occur to the patient but do not belong to other event categories are considered OCCURRENCES (eg, "follow up," "transport," etc). In addition to the type of an event, each mention was assigned a modality (FACTUAL, CONDITIONAL, POSSIBLE, and PROPOSED) and a "polarity" (ie, negated or not).

In this paper we focus on the methodologies engineered for the extraction and normalization of TEs and identification of events from clinical narratives.

4.4 Materials and methods

We approached the tasks as Named Entity Recognition (NER) problems, with the aim to identify relevant text spans and assign required attributes. The system (see Figure 4.1) comprised two tracks: (a) TE identification and normalization, and (b) event recognition. Both tracks start with a common pre-processing step (in order to produce the features for subsequent steps). For identification of TEs we have developed two approaches (a rule-based and a machine learning), which are combined before the TE normalization module. Six separate Machine Learning (ML) modules were developed for events. Given that target annotations comprise spans of text, we approached the task as a sequence labeling problem and trained a separate Conditional Random Fields(CRFs) [94] model with a number of shared features. The results from the CRF modules are followed by a set of post-processing rules that are designed to improve the boundaries of the resulting text spans. In addition to CRF models, a dictionary-based module was developed for one of the event classes (*Clinical Department*).

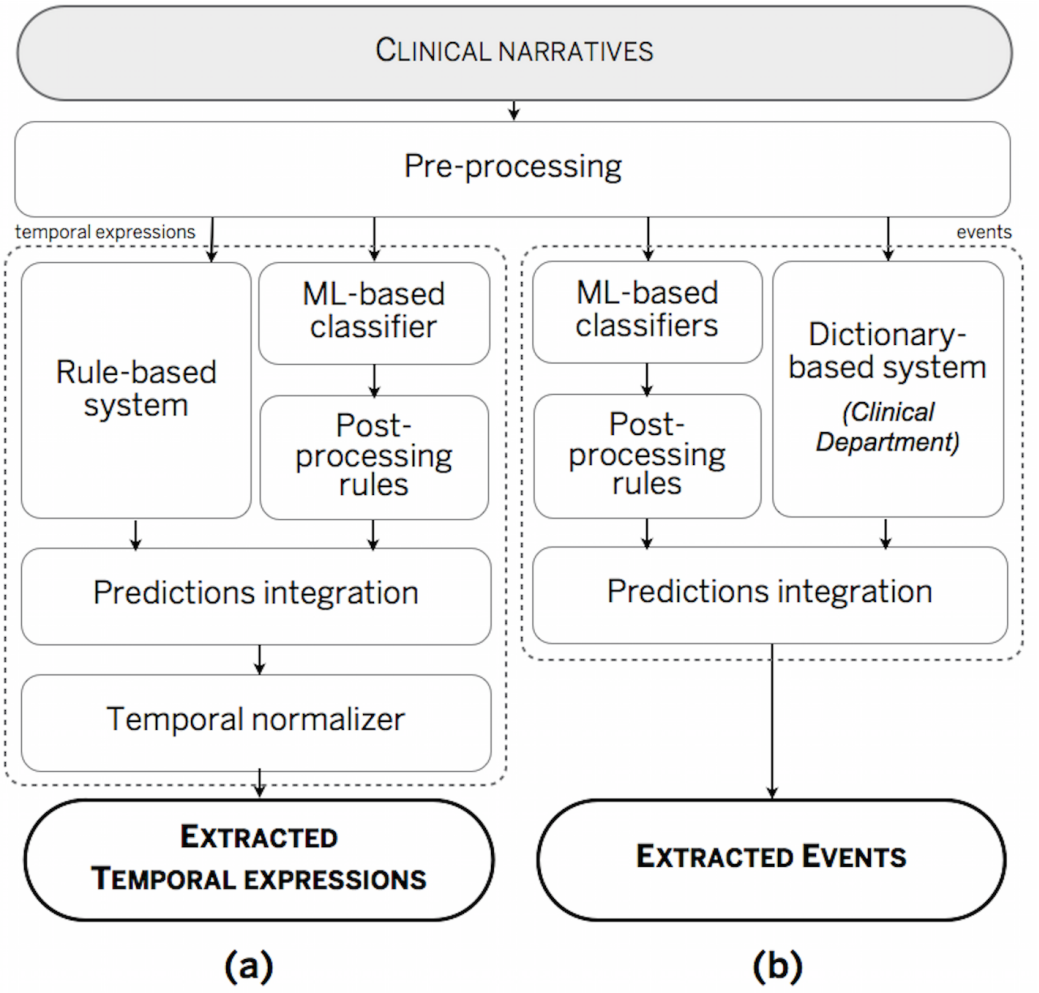


Figure 4.1: The overall system architecture. ML, machine learning.

4.4.1 Data

4.4.1.1 Dictionaries

We manually crafted a dictionary of temporal terms that included five common types of constituents of TEs: weekdays and months; times of day; spelled-out numbers; medical temporal abbreviations (eg, "OBD"); and common TE references (such as "previously," "today", etc). This dictionary was used in the feature extraction process for our ML models. In addition, a dictionary of clinical departments was semi-automatically collected using OpenNLP [10] NER to automatically extract candidate clinical department names from the i2b2 2010 and 2012 datasets. The candidates were then manually filtered to remove ambiguous terms. This dictionary was used for recognition and classification of mentions of *Clinical Departments* only.

4.4.1.2 Annotated corpora

For training, we used the training corpora provided by the i2b2 2012 challenge (190 mention-level annotated narratives, 2420 timexes, 16.534 events). Additionally, the i2b2 2010 challenge corpus comprising 426 narratives annotated with Problem, Test, and Treatment concepts [176] was used to support the event recognition track. The methods were tested on the i2b2 2012 test dataset (120 narratives, 1820 timexes, 13.594 events).

4.4.2 Pre-processing

The narratives were first pre-processed (tokenization, sentence splitting, Part-of-speech (POS) tagging, chunking) by GATE [40] which was used to develop our rule- and dictionary-based modules. Additionally, lexical features for our ML modules were generated by cTAKES, which provided tokens, POS tags, and chunks.

Mentions of *Problem*, *Test*, and *Treatment* were pre-generated by the Assertion module of cTAKES. Recognition of other medical entities was done by mapping all nominal chunks using MetaMap. The presence of negation was detected using NegEx [31], and sections (eg, "history of present illness", "hospital course") within narratives were detected using simple heuristics. We also extracted semantic roles using the Clear Parser semantic role labeller [35] module from cTAKES: each token was linked to an associated verb and assigned a role (eg, object, subject) in relation to that verb; verb tokens heading a sentence or sub-sentence were assigned a set of all participants and their roles linked to the verb. For example, the token "Thoracentesis" in the sentence "Thoracentesis was performed on 7-12-91". would be marked as an object of the verb "perform".

4.4.3 Extraction and normalization of temporal expressions

This module aims to identify and normalize TEs in pre-processed clinical narratives. The TE identification component accepts plain-text narratives and produces the spans of text recognized as TEs. We developed two identification modules: one based on rules and one ML-based. The results of both modules were integrated and passed to the normalization component, which provided the *type*, *value*, and *modifier* for the identified TEs.

4.4.3.1 The rule-based module

The rule-based module was developed using GATE. A total of 65 rules were engineered containing literal expressions derived from initial collocation extraction of TEs in the training data. The rule set is made up of (a) JAPE [40] *macros* which defined a set of recurring literals and symbols (e.g., temporal modifiers, weekdays, name of months, temporal medical abbreviations, etc.) and (b) JAPE *rules* which combine *macros* and JAPE *grammar* for rule formalism. The effectiveness of rules

(in terms of precision, recall, and F1 score) was analyzed on the training data to identify those that could have a positive effect on precision, recall, or F1 score.

4.4.3.2 The CRF-based module

The CRF-based module used token-level features that included the token's own properties and context features of the neighbouring tokens (the experiments on the training data showed that two tokens each side provided the best performance). We used the inside-outside (I-O) annotation. The following features were engineered for each token:

1. **Lexical features** included the token itself, its lemma, and POS tag, as well as lemmas and POS tags of the surrounding tokens. Each token was also assigned features from its associated chunk (phrase): the type of phrase (nominal, verbal, etc), tense and aspect (if the phrase was verbal), the location of the token within the chunk (beginning or inside), and the presence of negation as returned by NegEx.

2. **Domain features** capture mentions of specific clinical/healthcare concepts. All nominal chunks were fed to MetaMap and the returned UMLS semantic class was used as a feature for all tokens within that particular chunk. In the case of multiple semantic classes returned by MetaMap, we concatenated them alphabetically and used the resulting string as a hybrid semantic class. Additionally, mentions of Problem, Test, and Treatment (as generated by cTAKES) were assigned to the token.

3. **Semantic role features** model dependencies between the token and associated verb, following the approach of Llorens et al. [101] Each token is assigned the role, the verb, and their combination (eg, "object+perform") in order to capture particular verb-role preferences.

4. **Section type feature** represents the section type in which the token appeared.

5. **TE features** represent five features that indicated the presence of the five common types of constituents of TEs in a given token (see the temporal dictionary mentioned in Section 4.4.1).

The results of the CRF-based tagging were post-processed to adjust the boundary/scope of token-level tags (e.g., including determiners and pronouns where appropriate) and remove obvious false positives (e.g., single character predictions such as "/" or "a").

The results of both identification modules were integrated. In cases of overlap, the union at the token level is taken: for example, consider the segment "starting at 9am of the morning of admission;" if our rule-based method tagged the segment "9am of the morning" and our CRF model annotated "morning of admission," the final integrated result will be "9am of the morning of admission."

4.4.3.3 Temporal expression normalization module

The temporal expression normalization module has three components:

- A rule-based extractor of key reference dates within the clinical pathway (namely, time of *admission*, *discharge*, *operation*, *transfer*) uses a set of regular expressions to extract and associate these main clinical events to a date. The rules are based on the proximity of specific keywords (eg, "operative," "hospital," "discharge," "operation") and their direction (is the event mention before or/and after these keywords).

- A rule-based utterance time selector pairs each TE with a reference time by analyzing its component words. For example, the expression "the day after the admission" will be paired with the date of the admission, where the

expression "postoperative day 2" will be paired with the operation date. In the case of ambiguous expressions (such as "that time," "that period"), the module used the time assigned to the precedingTE. Otherwise, for all other TEs, the default reference time (*admission*) was used.

- Clinical NorMA, a rule-based clinical TE normalizer, provides the *value*, *type*, and (optional) *modifier* to identified TEs. It extends a pre-existing open-source general-domain normalizer [54]. To each TE and its associated reference time (from the utterance time selector), Clinical NorMA applies dictionary-driven regular expressions (83 general domain and 66 rules specifically designed for the clinical domain) to identify the *value* and *type* of the TE. The TE modifier is set only if a specific syntactic expression is triggered, for example, "in [number] or [number] days". If the modifier has not been assigned using such expressions, the *modifier* (MOD) component checks for the presence of trigger words (eg, "approximately," "several," "nearly"). These triggers have been mined from the training corpus by applying a feature selection algorithm based on mutual information. Finally, the post-processing component applies additional rules that correct systematic errors or provide default values (eg, the substitution of the undefined number of days in "PXD" with a default value, which was set as 3 for the i2b2 challenge).

Figure 4.2 summarizes the architecture for the identification and normalization of TEs.

4.4.4 Extraction of event mentions

This module aims at the extraction of event mentions. Apart from *Clinical Department*, all other event types were identified using CRFs only. The mentions

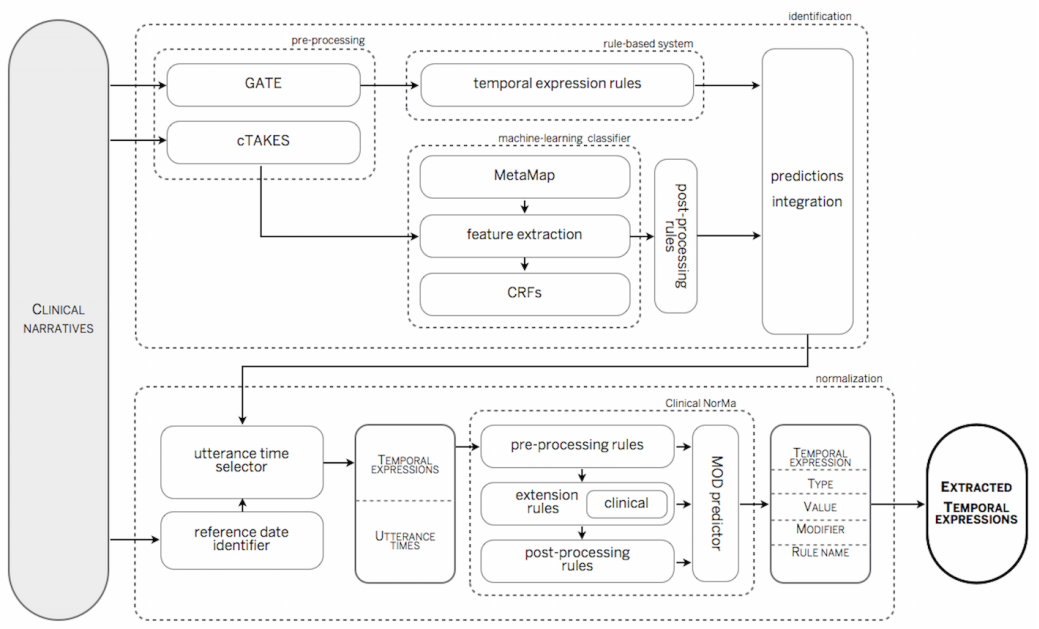


Figure 4.2: Temporal expression extraction and normalization architecture. CRF, conditional random field.

of *Clinical Departments* were identified using both a manually-curated dictionary (see section 4.4.1) and a CRF module, which were integrated at the token level (like the TEs above).

The event CRF models were trained on relevant (type-specific) subsets of the training data and they all shared a number of feature groups (see Table 4.1 for a summary). However, the *Evidential* and *Occurrence* types relied on additional feature groups as their scopes were not as focused as the scopes of other four event classes. We therefore added three additional feature groups for these event types:

- **Frequency** of the token annotated as *Occurrence* or *Evidential* in the training set, with the aim to help the model resolve confusion between them.

- **Co-occurring events**: An analysis of the training data revealed that mentions of these two event types correlate with the presence of other events in the same sentence. For example, the verb is "noted" often annotated as

Evidential if it is preceded with a *Problem* event (e.g., `<Problem> Oral cyanosis and shallow respirations </Problem>` were `<Evidential> noted </Evidential>`). We therefore decided to include predictions from *Problem*, *Test*, *Treatment*, and *Clinical Department* modules as features for the CRF models of the *Evidential* and *Occurrence* categories. The resulting tags of the *Evidential* model were also used as features in the *Occurrence* CRF model. We note that therefore the CRFs were run in a particular order: the models for *Treatments*, *Tests*, *Problems*, and *Clinical Departments* were run in parallel; the outputs of these models were then used as features for the *Evidential* CRF, whose predictions were used as a feature in the *Occurrence* model.

- **Expanded lexical features**: The initial experiments also revealed that the lexical variability of the *Occurrence* class was high, so an additional feature was added to indicate if a token is a typical *Occurrence* unigram. These unigrams were derived manually from a list of the 500 most frequent unigrams associated with this category (as obtained from the training data); after removing ambiguous terms, the list comprised 289 unigrams. The feature was also considered for the *Evidential* events, but associated words were heavily context dependent and thus not useful.

All CRF-based results were post-processed in the same way as TEs. Figure 4.3 summarizes the architecture developed for extraction of event mentions.

Finally, each of the recognized events was checked with NegEx to determine polarity. For modality, lexical clue-based rules were explored during the development phase, but produced no significant improvement as compared to setting this attribute to *Factual* for every recognized event (95% of all events in the training data were *Factual*).

Entity type	Lexical features	Domain features	Semantic role features	Section type feature	Temporal expression features	Frequency features	Co-occurring event features	Expanded lexical features
Problem	✓	✓	✓	✓	✓			✓
Test	✓	✓	✓	✓	✓			✓
Treatment	✓	✓	✓	✓	✓			✓
Clinical department	✓	✓	✓	✓	✓			✓
Evidential	✓	✓	✓	✓	✓	✓	✓	✓
Occurrence	✓	✓	✓	✓	✓	✓	✓	✓
Temporal expressions	✓	✓	✓	✓	✓			✓

Table 4.1: Groups of features used in the CRF models.

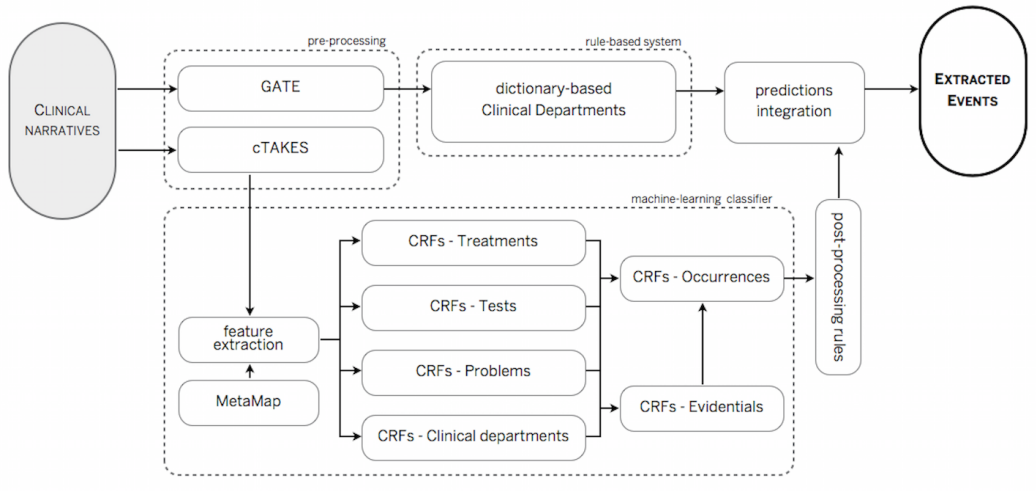


Figure 4.3: Event extraction architecture.

4.5 Results

4.5.1 Extraction and normalization of temporal expressions

The main evaluation metric for the TE recognition task was the product of the F1 score calculated with the lenient matching strategy (requiring that the system output overlaps the gold standard) and the accuracy obtained for the *value* attribute. Three different results were officially evaluated, based on the way temporal extractions were identified (the normalization module was always applied in full):

- run 1: only TEs identified by the rules optimized for F1 score;
- run 2: the union of recall-optimized rule-based predictions and tags generated by CRFs;
- run 3: only TEs identified by the rules optimized for precision.

The optimization has been performed on the training set with respect to the lenient matching strategy. Table 4.2 and 4.3 provides the results: run 2 provided the best F1 score (90.08%) along with the highest recall (91.54%). The strict evaluation scores (requiring an exact match between the system output and gold standard) were significantly lower (by 10-12% for the F1 score) indicating that both the ML and the rule-based approaches would benefit from a better method of boundary adjustment. The normalization scores were also highest in run 2 (*type*: 84.73%, *value*: 70.44%, *modifier*: 82.75%). We note that this is a state-of-the-art result as our run 2 was a top ranked outcome of the 2012 challenge (there were no significant differences between the top three runs, coming from three different teams). Compared to the results on the training data, there was some drop in the strict F1 score values (see Section B.1 in the supplementary data), in particular for the rule-based runs, but overall lenient F1 score and normalization results were

comparable with around a 1% difference between them (in some cases, the test results were even better).

4.5.2 Extraction of event mentions

The event extraction task was evaluated using the F1 score calculated with the lenient matching strategy, averaged across all the annotations in the test corpus. Accuracy was used to evaluate polarity and modality attributes. Two different results were officially evaluated, different only in how mentions of *Clinical Departments* were identified:

- run 1: targeted precision by choosing *Clinical Department* predictions based on dictionary matches only;
- run 2: targeted recall, so *Clinical Department* predictions included the union of CRF- and dictionary-based tags.

The mentions of other event types were identical in both runs. Table 4.4 and 4.5 provides the results: overall, run 2 gave better results, with the better F1 score (87.29%), recall (85.32%), and accuracies: polarity (79.45%) and modality (81.53%). Nonetheless, as expected, better precision (89.64%) was achieved in run 1. The strict F1 scores were 8% lower in both submissions, indicating again that

	Identification					
	Strict matching			Lenient matching		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Run 1	78.03	78.41	78.22	89.23	89.62	89.42
Run 2	77.03	79.62	78.30	88.68	91.54	90.08
Run 3	79.85	77.09	78.45	90.38	87.25	88.79

Table 4.2: Temporal expression identification: micro-averaged results on the test data (120 narratives, 1820 temporal expressions)

	Normalization		
	Type (%)	Value (%)	Modifier (%)
Run 1	83.30	69.73	81.98
Run 2	84.73	70.44	82.75
Run 3	80.88	67.91	79.67

Table 4.3: Temporal expression normalization: micro-averaged results on the test data (120 narratives, 1820 temporal expressions)

determining the right boundaries for token-level recognized events was challenging. When compared to the training data (see Section B.1 in the supplementary data), the system seems to have generalized well as there was even a slight increase in F1 score (around 1%) as compared to the training data.

4.6 Discussion

4.6.1 Temporal expression recognition

The results indicate that textual spans that represent TEs can be identified with an F1 score of 90%, with no significant differences between rule-based and integrated models (as expected, the rule-based runs showed slightly better precision; the CRF model on its own showed lower performance, with an F1 score of 86.70%). Common errors include mentions of clinical findings that follow date patterns or

	Strict Matching			Lenient matching		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Run 1	82.05	77.05	79.71	89.64	84.66	87.08
Run 2	81.74	78.05	79.85	89.35	85.32	87.29

Table 4.4: Event identification: micro-averaged results on the test data (120 narratives, 13 593 events)

	Attributes	
	Polarity (%)	Modality (%)
Run 1	78.81	80.08
Run 2	79.45	81.53

Table 4.5: Event attributes extraction: micro-averaged results on the test data (120 narratives, 13 593 events)

ambiguous mentions (such as "now"). The supplementary data in Section B.2 provide a detailed error analysis.

4.6.2 Temporal expression normalization

The normalization results vary for different attributes: while *type* and *modifier* have reasonable accuracies (84.73% and 82.75%, respectively), the *value* attribute proved challenging (70.44%). This is expected given that the value prediction asked for complete identification of a TE, whereas the other two attributes provide categorization-like values. We note that in some cases the normalizer failed to correctly distinguish between *Date* and *Duration* (and less frequently between *Time* and *Duration*). This is mostly due to wrong boundaries inherited from the TE recognition (e.g., omission of an important preposition like "three days' (date or duration) versus "every three days' (frequency)). A detailed error analysis is provided at Section B.2 in the supplementary data.

4.6.3 Event recognition

The type-specific lenient evaluation results are given in Table 4.6. The best F scores were achieved for frequent, well-defined event types, such as *Problem* (91.38%), *Test* (91.11%), and *Treatment* (89.26%). This result showed that CRF models generalized well with an abundance of training data (the additionally used

2010 dataset) and benefited from the use of terminological processing (cTAKES, MetaMap). For example, when the 2010 dataset is removed, the F1 scores drop notably for *Problem* (by 21%), *Test* (19%), and *Treatment* (20%) (data not shown). We note that these results are in line with the top performing systems in the i2b2 2010 challenge (the lenient F1 score ranges from 89.80% [82] to 92.40% [41]).

4.7 Conclusion

This paper presents and evaluates various approaches to the extraction of clinically relevant events and TEs from clinical narratives, as part of our participation in the i2b2 2012 challenge. The methodology relies on combining rule-based approaches with feature-rich ML, which includes morphological, lexical, syntactic, semantic, and domain-specific features. The rule-based components were designed to handle the recognition and normalization of TEs and *Clinical Departments*, while CRF models were trained for all event and temporal recognition tasks.

The hybrid temporal recognition and normalization system provides state-of-the-art results with a micro F1 score of over 90% for lenient matching, and accuracies of 84.73% (*type*), 70.44% (*value*), and 82.75% (*modifier*) for the TE normalization. Clinical event extraction showed good performance with a micro F1

Event type	Frequency	P (%)	R (%)	F1 (%)
Problem	4309	95.24	87.82	91.38
Treatment	3285	95.68	83.65	89.26
Occurrence	2499	63.43	66.91	65.12
Test	2173	95.05	87.48	91.11
Clinical department	732	76.02	83.61	79.64
Evidential	595	64.99	75.80	69.98

Table 4.6: Event recognition: per category performance on the test data (run 2, lenient matching).

score of 87.29% (lenient). The well-scoped classes (such as *Problem*, *Treatment*, and *Test*) showed very good performance (F1 score of 90%), whereas unfocused and context-dependent categories (e.g., *Clinical Department*, *Occurrence*, *Evidential*) proved to be challenging. Our study also revealed that the use of additional annotated corpora can indeed benefit the models, relaxing the need for specific terminological information.

While performance based on lenient matching was good, the most challenging part remains deciding the right boundaries of mentions (strict F1 scores were 78% and 80% for TE and event mentions, respectively). In addition, future work needs to explore new methods and features to capture context-dependent mentions and model unfocused categories.

A comparison to the agreement between the human annotators (89% for TEs and 87% for event recognition) indicates that the quality of the system's performance is comparable to what can be expected from manual efforts, and thus can be used either as a pre-processing step for a manual review process or as a part of a large-scale processing of electronic health databases.

The methods described here are packed in the TERN and CliNER tools, which are freely available at `http://gnode1.mib.man.ac.uk/hecta.html`.

Contributors

AK developed the machine learning models and the overall system architecture, and drafted the manuscript; AD developed the rule-based temporal recognition models, collected the Clinical Department dictionary, and drafted the manuscript; MF developed the clinical temporal expression normalization component, performed the error analysis and drafted the manuscript; JAK performed critical revision of the manuscript and gave final approval; GN was the team leader in the i2b2

challenge, supervised the development of the whole system, and performed critical revision of the manuscript.

Funding

This work was partially supported by PhD scholarships from the University of Manchester and UK Engineering and Physical Sciences Research Council (EPSRC) (to AD, MF) and the following projects: “Development of new information and communication technologies, based on advanced mathematical methods, with applications in medicine, telecommunications, power systems, protection of national heritage and education” (III44006) and “Infrastructure for Technology Enhanced Learning in Serbia” (III47003) (the Serbian Ministry of Education and Science, to AK, GN); “Linked2Safety: A next-generation, secure linked data medical information space for semantically-interconnecting electronic health records and clinical trials systems advancing patients safety in clinical research” project (EU FP7 ICT, contract 288328, to JAK, GN); “A study using techniques from clinical text mining to compare the narrative experiences of patients with medulloblastoma with factors identified from their hospital records” (The Christie Paediatric Oncology Charitable Fund, to JAK, GN, AD). Health eResearch Centre (HeRC) is funded by a consortium of ten UK government and charity funders, led by the Medical Research Council (MRC). The i2b2 challenge is supported by Informatics for Integrating Biology & the Bedside (i2b2) award number 2U54LM008748 from the National Institutes of Health (NIH)/National Library of Medicine (NLM), by the National Heart, Lung, and Blood Institute (NHLBI), and by award number 1R13LM01141101 from the NIH NLM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM, NHLBI, or the NIH.

Competing interests

None

Provenance and peer review

Not commissioned; externally peer reviewed.

Acknowledgements

AK would like to acknowledge the Serbian Ministry of Education and Science for funding his stay at the University of Manchester (June-August 2012).

Chapter 5

Using machine learning to predict temporal orientation of search engines' queries in the Temporalia challenge

"Inherited Will, The Destiny of the Age, and The Dreams of the People. As long as people continue to pursue the meaning of Freedom, these things will never cease to be!" - Gol D. Roger

– Eiichiro Oda, *One Piece*

This chapter is directly adapted from the following paper:

- Michele Filannino and Goran Nenadic. Using machine learning to predict temporal orientation of search engines' queries in the temporalia challenge. In *NTCIR-11, EVIA 2014 (NII Testbeds and Community for Information Access Research)*, 2014

5.1 Abstract

We present our approach to the NTCIR-11 Temporalia challenge (introduced in Chapter 2), Temporal Query Intent Classification: predicting the temporal orientation (present, past, future, atemporal) of search engine user queries. We tackled the task as a machine learning classification problem. Due to the relatively small size of the training set provided, we used temporal-oriented attributes specifically designed to minimise the features' sparsity. The best submitted run achieved 66.33% of accuracy, by correctly predicting the temporal orientation of 199 test instances out of 300. We discuss the results of the manual error analysis performed on the predicted classes, which sheds light on the main sources of error. Finally, we present some a-posteriori improvements to the best submitted run, which lead to a 6% improvement in terms of accuracy (72.33%).

5.2 Introduction

Temporal information extraction [184, 188, 181] is pivotal for many Natural Language Processing (NLP) applications such as question answering, text summarisation and machine translation. The use of such information also plays a crucial role in the field of Information Retrieval (IR).

Research in this context has lead to IR systems which consider temporal information of indexed documents and users' queries to improve their accuracy by temporally filtering results in order to better capture user's intent. Being able to predict the temporal orientation of a query like `'weather in manchester'`, makes search engines able to show updated real-time meteorological information, whereas in the case of `'Weather forecast manchester'` they are more likely to show results about the immediate future. Some queries (e.g. `'sunday times'`, `'galileo Galilei'`), on the other hand, do not have a specific temporal orientation.

To address this issue, a shared task (called Temporalia [79]) was organized by the Japanese National Institute of Informatics (NII) in which systems are asked to automatically predict the temporal orientation of a given user query in one of the following categories: past, present, future and atemporal.

Search queries are atemporal when they do not have a temporal intent. Therefore the corresponding search results are in principle not expected to change due to the passing of time. On the other hand, search results for past, recency and future queries are related to time. Recency queries refer to present events, future queries refer to predictions or scheduled events, and past queries are related to events already happened.

This paper describes how we tackled this problem. Section 5.3 introduces the characteristics of the data provided by the challenge organisers. Section 5.4 illustrates the proposed machine learning-based methodology along with the attributes explicitly designed to minimise features' sparsity. The Results section (5.5) presents the accuracy of the different submitted runs, investigates the main sources of error, and presents some further a-posteriori improvements to our best performing model. We conclude the paper with a Discussion section (5.6) and Conclusions (5.7).

5.3 Data

The organisers of the Temporalia challenge released a data set of 80 search engine queries where each one consisted of the textual representation (query), the submission time and the gold temporal orientation class (`atemporal`, `past`, `recency` or `future`).

They also provided a set of 20 queries to be used as a preliminary test set, therefore without temporal orientation (unlabelled). We manually annotated them

and once the organizers confirmed the quality of the annotation (95% of accuracy, 19/20 correctly classified) we included them in the training set.

The official benchmark test set for the challenge consisted of 300 unlabelled queries. The Table 5.1 shows an excerpt of the training data.

Query	**Submission**	**Class**
Movies 2012	Feb 28, 2013	Past
Upcoming Movies in 2013	Jan 1, 2013	future
2013 MLB Playoff Schedule	Jan 1, 2013	future
current price of gold	Feb 28, 2013	recency
Amazon Deal of the Day	Feb 28, 2013	recency
Number of Neck Muscles	Feb 28, 2013	atemporal
...	...	...

Table 5.1: Example of the training instances. All the queries have been submitted at GMT+0.

5.4 Methodology

The task can naturally be seen as a 4-class classification problem since each query is associated with one and only one class. We therefore tackled it using a supervised machine learning-based approach. We mostly focussed our work on designing and testing a set of temporal-related attributes with a small set of possible values. As a consequence, this allowed us to minimise the total number of features required to model the classification problem.

While *attribute*, *feature* and *value* are often used synonymously, in this paper we use them with a definition mutuated from the machine learning community [1]. In particular, a feature F is a true predicate expressing the pairing of a particular attribute h and its value v. For example, `lower=upcoming` is a feature, where `lower` is the attribute and `upcoming` is its value.

ID	Object	Attribute	$\|V\|$	Example: query/submission → attribute value
1	Q	Is it a Wikipedia page title?	2	"New York Times" → 'YES'
2	Q	Does it contain a temporal expression?	2	"june 2013 movies" → 'YES'
3	S	Submission's term	3	"Feb 28, 2013 GMT+0" → 'B'
4	S	Submission's trimester	4	"Aug 26, 2013 GMT+0" → 'M2'
5	B	Timing	4	"Movies 2012", "Feb 28, 2013 GMT+0" → 'past'
6	Q	Most frequent trigger class	5	"peso dollar exchange rate" → 'present'
7	Q	Wh type	5	"how did hitler die" → 'how'
8	Q	Most frequent TempoWordNet class	5	"current stock prices" → 'present'
9	Q	Most frequent POS tag tense	7	"what is stop kony 2012" → 'VBZ'
10	Q	Most frequent coarse-grained POS tag	8	"kony 2012 fake" → 'N'
11	Q	Trigger classes footprint	11	"what was I thinking lyrics" → 'past-atemporal'
12	B	Temporal Δ between submission and query	16	"father's day 2010", "Feb 28, 2013 GMT+0" → 36.0
13	Q	Tenses footprint	18	"when does fall start" → 'VBZ-VB'
14	Q	Ordered TempoWordNet classes	18	"the last song" → 'past-future-present-atemporal'
15	Q	Most frequent fine-grained POS tag	21	"kony 2012 fake" → 'NN'
16	Q	Coarse-grained POS tag ordered footprint	119	"when is labour day" → 'N-W-V'
17	Q	Fine-grained POS tag ordered footprint	202	"when is labour day" → 'NN-WRB-VBZ'
18	Q	Coarse-grained POS tag footprint	204	"when is labour day" → 'W-V-N-N'
19	Q	Fine-grained POS tag footprint	265	"when is labour day" → 'WRB-VBZ-NN-NN'

Table 5.2: List of attributes used, ordered by number of possible values. Object column indicates whether the attribute is computed from the (Q)uery, the (S)ubmission date or (B)oth. The $|V|$ column contains the cardinality of the value set per attribute (measured on the entire data set). Coarse-grained POS tags have been computed by considering just the first letter of the *Penn Treebank Tag Set*. POS tags are computed using the MaxEnt Treebank POS tagger from the Python NLTK library. The TempoWordNet-related attributes (#8 and #14) use the WordNet-based lemmatiser.

5.4.1 Pre-processing

All the user queries from the training and test data have been firstly pre-processed: for each user query we computed its lower-case version, its tokenisation and POS tags. The submission times have been pre-processed too: for each of them we firstly normalised[1] it via NorMA [54] (a temporal expression normaliser described in Chapter 3 and Appendix A), and from this we separately extracted the numerical representation of year, month and day. The time of the query submission has not been taken into account.

5.4.2 Attributes

The limited size of the training set made the task challenging for machine learning since the use of the attributes commonly used in NLP would have easily lead to sparse feature space, potentially leading to high-variance models (overfitting) in a real search engine's use scenario. By using just 100 samples, bag-of-words and n-grams representations would not have provided any support due to the huge number of possible different values to be learned.

We proposed 19 different attributes each one with a different number of possible values. An overview of them, along with explanatory examples is presented in Table 5.2.

Sometimes search engines are used as a faster alternative to typing the precise URL of our preferred web sites. This is the case, for example, of queries such as "the sunday times" or "wikipedia". We introduced the attribute #1 (see Table 5.2) to capture such cases. The titles of all the Wikipedia English pages have been collected via DBPedia [14]. The attribute value is positive only if a Wikipedia title and the query (as it is) are case-insensitively equal.

[1] A temporal normaliser provides a standard ISO 8601 representation of any temporal expression: dates, durations, times and sets.

The information about the presence of temporal expressions in the query text (attribute #2) is important to separate the atemporal queries from the rest of them. We used ManTIME [55], a temporal expression extraction system, to extract the temporal expressions from the queries' text. We also used a backup regular expression-based system to spot date mentions (e.g. *"2012"*, *"1900"*), only in the case ManTIME does not find any temporal expression. The attribute has a positive value only if at least one temporal expression, or date mention, has been extracted.

Via a preliminary analysis of the training data we noticed that the part of the year in which the query has been submitted could play a crucial role in the classification task. Consequently, we designed two attributes (#3 and #4). The first one assigns 'B', 'M' or 'E' if respectively the query has been submitted in the first, second or third term (four-month period) of the year. The second one uses trimesters instead, leading to 4 possible values: 'B', 'M1', 'M2' or 'E'. Table 5.2 provides some examples. Using the normalised submission time and the extracted temporal expressions from the query text, we also compute two supplementary temporal attributes: #5 and #12. The latter is a numerical attribute corresponding to the difference, in terms of months, between the temporal expressions in the query and the submission date. The attribute #5 represents just its "sign" in the following categories: `present`, `past`, `future`.

From the training data we extracted the word and bigram vocabulary of the queries and filtered them as attributes by using RELIEF [88], a feature selection algorithm. We have been able to obtain a ranked list of the most (and least) influential unigrams and bigrams with respect to the classification task. Through a manual analysis, we grouped them in temporal trigger gazetteers, one per temporal class, according to their pertinence. For example, the future triggers include words such as *"forecast"*, *"upcoming"*, whereas the past triggers include words such as *"last"* and *"previous"*. The attribute #6 represents the most frequent temporal

trigger type in the query, whereas the attribute #11 represents the entire sequence of triggers in the order they appear in the query ("footprint").

We integrated TempoWordNet [46], a lexical knowledge-base for temporal analysis which provides a probabilistic measure of temporal orientations for the WordNet's synsets. Since WordNet's synsets are sets of lemmas, we lemmatised the search query and represented the most likely temporal orientation class according to TempoWordNet (attribute #8) and the sorted list of them (attribute #14). For each lemma, the most likely corresponding WordNet sense has been used.

We also checked if a query is a wh-question. The attribute #7 represents which type of question the query is among the following possibilities: "what", "when", "where", "who", "why", and "how". The attribute just checks the query's first word.

Since queries are usually small multi-word expressions, we investigated the use of POS tags in different ways. The hypothesis was that specific sequences of tags could be correlated with some classes. The attributes #9 and #13, in particular, are focussed on verbs only. They represent the POS tag of the most frequent tense and the POS tag sequence, respectively. Attributes #10 and #15 are the most frequent coarse and fine-grained POS tag, respectively. Finally, the last four attributes (#16-19) are POS tag sequence ordered by the frequency or by order of appearance, using coarse and fine-grained tags.

For each of the attributes presented we also counted the cardinality of their value sets ($|V|$ column in Table 5.2): the number of different values each attribute can take. The counts have been computed using the entire data set (training and test) and it provides a rough, but useful, estimation of their sparsity.

5.4.3 Submitted Runs

We experimented with different machine learning models: SVM with linear, polynomial and RBF kernel, Naïve Bayes, C4.5 decision tree and Random Forests.

ID	Attribute	Run 1	Run 2	Run 3
1	Is it a Wikipedia page title?		✓	✓
2	Does it contain a temporal expression?	✓	✓	✓
3	Submission's term			✓
4	Submission's trimester			✓
5	Timing	✓	✓	✓
6	Most frequent trigger class	✓		✓
7	Wh type		✓	✓
8	Most frequent TempoWordNet class			✓
9	Most frequent POS tag tense	✓	✓	✓
10	Most frequent coarse-grained POS tag		✓	✓
11	Trigger classes footprint	✓	✓	✓
12	Temporal Δ between submission and query		✓	✓
13	Tenses footprint		✓	✓
14	Ordered TempoWordNet classes			✓
15	Most frequent fine-grained POS tag		✓	✓
16	Coarse-grained POS tag ordered footprint			✓
17	Fine-grained POS tag ordered footprint			✓
18	Coarse-grained POS tag footprint			✓
19	Fine-grained POS tag footprint			✓

Table 5.3: List of attributes used in the submitted runs with reference to Table 5.2.

The parameters for SVMs have been preliminary optimised on a sub set of the training data (20 samples) and 10 cross-fold validation has been used for all the experiments. We noticed the SVM (with polynomial kernel) and Random Forest algorithm systematically outperforming the rest. We used the former in Run 1 and 2, and the Random Forest algorithm for the Run 3. The attributes used for each run are illustrated in Table 5.3.

For the Run 1, called `minimal`, we selected the first 11 attributes and discarded the ones that did not positively contribute to the model (measured with RELIEF). In particular, we registered no improvements in the use of TempoWordNet-based attributes (#8 and #14), as well as the ones related to the submission part of the year (#3 and #4). The second run, called `intermediate`, is built on top of the first one, except for the absence of the most frequent trigger classes (#6). We added all the features with a cardinality less than 100, except for the TempoWordNet-related

ones. The third run, called `full`, uses all the attributes presented in Section 5.4.2.

5.5 Results

Run 1 obtained the highest accuracy by correctly predicting the temporal orientation of 199 queries (66.33%) out of 300. The intermediate and full models achieved, as predicted, lower accuracy.

Name	Accuracy	#
TUTA1	74.00%	222
And7	72.00%	216
HULTECH	68.00%	204
HITSZ	67.67%	203
UniMan Run 1 (minimal)	66.33%	199
mpii	64.00%	192
UniMan Run 2 (intermediate)	61.33%	184
UniMan Run 3 (full)	55.00%	165

Table 5.4: Results of the three submitted runs with respect to the other participants' best runs. Attribute set names, accuracies and number of correctly predicted instances are shown.

In the challenge, the Run 1 ranked 5th among the best runs, and 11th out of the 17 submitted runs. Further analysis on the submitted models showed that there is no statistically significant difference between the minimal and intermediate model. On the contrary, there is a statistically significant difference between minimal and full, and intermediate and full.

5.5.1 Error analysis

An analysis of the confusion matrix for the minimal run (see Table 5.5, below) highlights interesting issues.

	Classified as:			
	Recent	*Past*	*Future*	*Atemporal*
Recent	**43**	0	*21*	11
Past	3	**60**	6	6
Future	*38*	0	**35**	2
Atemporal	6	5	3	**61**

Table 5.5: Confusion matrix of the minimal run predictions for the official benchmark test set. True positive diagonal is in bold. Problematic cases are italicized.

We are able to identify three different major sources of classification mistakes, presented by their frequency:

Future as recent 38 `future` instances have been misclassified as `recent`. Some example of misclassified queries are: *"college rankings in 2013"*, *"2013 wimbledon"* and *"voice 2013 winner"*, which have all been submitted on the 1st of May 2013. The events described in the queries did not happen yet at the time of search and therefore the temporal orientation should have been future.

Recent as future. In 21 cases, `recent` instances have been misclassified as `future`. Some examples of misclassified queries are: *"bruins game tonight time"*, *"weather for nyc"*, *"when does spring start 2014"* submitted on 1st of May 2013. The first two examples are clear cases of recent temporal orientation, since the user is searching for information related to the day of the search. The last example, on the contrary, is questionable: the query could have been annotated as atemporal since the information searched for does not dependent on time.

Recent as atemporal. Finally, 11 `recent` instances have been erroneously classified as `atemporal`. Some examples of misclassified queries are: *"value of silver dollars"*, *"time in hawaii"*, *"24 hour clock"*, and *"disney prices going up"*. In all these cases the users expect search results which are strictly

related with the current time. Prices, currency values and updated times are all examples of such category.

By manually investigating the attribute representation of these errors, we found that the major part of them are due to the absence of some trigger words in the gazetteers. In some cases, the misclassification is due to a wrong grouping of triggers. For example, the trigger *"tonight"* has been assigned to the future gazetteer instead of the recency one. Only a small part of them is due to the classifier limitations.

More generally, we also find some limitations in the representation of attributes, which if solved could have lead to better classification performance. Multi-valued attributes (#11 and #14) could have been substituted with groups of binary features. Some attributes (#10, #13 and #15-17) were affected by ordering problems, which lead to different string representation though conveying the same information. Due to the choice of attributes selected for the best run (minimal) only the wrong trigger classification (#11) affects the best performance.

5.5.2 A-posteriori improvements

By fixing the limitations mentioned above, the minimal model correctly classifies 217 instances (18 instances more) of the official benchmark test set, achieving an accuracy of 72.33% (+6%).

By using the fixed attribute set, we also determined which model would have provided the best performance. An exhaustive search among all the possible combinations of attribute sub-sets found the best of them providing 76% of accuracy (228 instances correctly classified). This level represents the upper bound for the accuracy of our attribute sets on the official benchmark test instances.

5.6 Discussion

We found that the task proved to be challenging due to some specific characteristics. The most important one is the dimension of the training set. We believe that 100 instances are surely not enough to train a robust machine learning classifier, due to the fact that many of the classic NLP attributes in the literature have a too sparse representation to be learned from such a small training set. At the same time, we perceive this limitation as a deliberately conceived characteristic of the data intended to avoid overfitting attributes/rules, which would have ultimately resulted in no future use for the community.

During our manual error analysis, we also found that some of the queries were particularly hard to classify even for people. An example is *"Ventura Stern 2016"* which refers to the nominee of a comedian duo to the 2016 USA elections. Some other queries were just partial (e.g. *""earth after 1"*). In some other cases, we faced the need for surfing the Internet to seek some temporal information about entities mentioned in the queries. This has been the case for *"season 2 dexter"* or *"season 3 game of thrones"*, which both refer to particular seasons of famous TV shows. These findings suggest a potential benefit from the use of a named-entity recogniser component along with some temporal contextualisation of the recognised concepts [56].

Finally, we found the contribution of TempoWordNet (as used by our attributes) to be negligible. The reason is that the temporal orientation of a word is related to its WordNet sense rather than its word-form which was essential in our task. Temporal orientation of all the verbs, for example, are inevitably missed since verbs in WordNet are represented through their infinitive form only. This also leads to a distribution of temporal orientation among senses which is skewed towards the atemporal class. 81.97% of senses have high probability of being atemporal, 13.72% of being present, 2.84% of being future, and just 1.48% of being past. If

the atemporal label, and to some extent the present label too, can be seen as a neutral choices, lots of examples from future and past categories seem not to have any relation at all with the temporal orientation of the sense.

5.7 Conclusions

In this paper we presented our approach to the Temporal Query Intent Classification subtask of Temporalia in the NTCIR-11 challenge. We tackled the task as a machine learning classification problem, by designing and proposing a set of temporal-oriented attributes which minimised the features' sparsity. An extensive overview of the attributes used, along with examples, has been illustrated in Section 5.4.2.

This piece of research contributes by presenting a ML-based strategy to classify queries with respect to their temporal orientations. The strategy adds Temporal Information Extraction (TIE)-based features to the ones commonly used in Natural Language Processing (NLP). The feature selection phase highlights the importance of the former type with respect to the latter.

Three different runs have been submitted, corresponding to three different attribute sets (minimal, intermediate and full) and two different machine learning classification algorithms (SVM with polynomial kernel and Random Forest). The minimal attribute set, which minimised the sparsity of the representation, achieved the best performance (66.33%) among our submitted runs. The model has been further improved, leading to a final accuracy of 72.33%.

A manual error analysis has been performed in order to highlight the main sources of classification error. We found that the major part of them are due to limitations related with the attribute representation.

To aid replicability of this work, the source code, the machine learning pre-trained models and the feature tables are available at `http://www.cs.man.`

ac.uk/~filannim/temporalia.html. All the data are available for the submitted runs and the fixed one.

5.8 Acknowledgements

MF would like to acknowledge the support of the UK Engineering and Physical Science Research Council in the form of doctoral training grant. We want to thank Gaël Dias and Mohammed Hasanuzzaman from the Normandie University for the availability in sharing with us TempoWordNet before its official public release.

Chapter 6

Mining temporal footprints from Wikipedia

"You'll stumble many times in the future, but when you do, each time you'll have more strength to bounce back."

– Nobi Nobita, *Doraemon*

This chapter is directly adapted from the following paper:

- Michele Filannino and Goran Nenadic. Mining temporal footprints from Wikipedia. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 7–13, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University

6.1 Abstract

Discovery of temporal information is key for organising knowledge and therefore the task of extracting and representing temporal information from texts has received an increasing interest. In this paper we focus on the discovery of temporal footprints

from encyclopaedic descriptions. Temporal footprints are time-line periods that are associated to the existence of specific concepts. Our approach relies on the extraction of date mentions and prediction of lower and upper boundaries that define temporal footprints. We report on several experiments on persons' pages from Wikipedia in order to illustrate the feasibility of the proposed methods.

6.2 Introduction

Temporal information, like dates, durations, time stamps etc., is crucial for organising both structured and unstructured data. Recent developments in the natural language community show an increased interest in systems that can extract temporal information from text and associate it to other concepts and events. The main aim is to detect and represent the temporal flow of events narrated in a text. For example, the TempEval challenge series [184, 188, 181] provided a number of tasks that have resulted in several temporal information extraction systems that can reliably extract complex temporal expressions from various document types [179, 101, 17, 55].

In this paper we investigate the extraction of *temporal footprints* [84]: continuous periods on the time-line that temporally define a concept's existence. For example, the temporal footprint of people lies between their birth and death, whereas temporal footprint of a business company is a period between its constitution and closing or acquisition (see Figure 6.1 for examples). Such information would be useful in supporting several knowledge extraction and discovery tasks. A question answering system, for example, could spot temporally implausible questions (e.g. *What computer did Galileo Galilei use for his calculations?* or *Where did Blaise Pascal meet Leonardo Da Vinci?*), or re-rank candidate answers with respect to their temporal plausibility (e.g. *British politicians during the Age of Enlighten-*

ment). Similarly, temporal footprints can be used to identify inconsistencies in knowledge bases.

Temporal footprints are in some cases easily accessible by querying Linked Data resources (e.g. DBPedia, YAGO or Freebase) [147], large collections of data [171] or by directly analysing Wikipedia info-boxes [119, 49, 192, 76, 93]. However, the research question we want to address in this paper is whether it is possible to automatically approximate the temporal footprint of a concept only by analysing its encyclopaedic description rather than using such conveniently structured information.

This paper is organised as follows: Section 6.3 describes our approach and four different strategies to predict temporal footprints. Section 6.4 provides information about how we collected the data for the experiments, and Section 6.5 presents and illustrates the results.

6.3 Methodology

In order to identify a temporal footprint for a given entity, we propose to predict its lower and upper bound using temporal expressions appearing in the associated text. The approach has three steps: (1) extracting mentions of temporal expressions, (2) filtering outliers from the obtained probability mass function of these mentions,

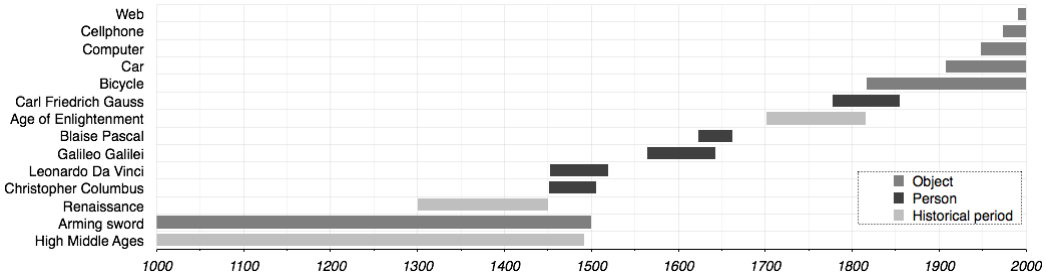


Figure 6.1: Examples of temporal footprints of objects, people and historical periods.

and (3) fitting a normal distribution to this function. This process is controlled by three parameters we introduce and describe below. We restrict temporal footprints to the granularity of years.

6.3.1 Temporal expression extraction (TEE)

For each concept we extract all the dates from its associated textual content (e.g. a Wikipedia page). There are numerous ways to extract mentions of dates, but we use (a) regular expressions that search for mentions of full years (e.g. sequence of four digits that start with '1' or '2' (e.g. *1990*, *1067* or *2014*) — we refer to this as TEE RegEx; (b) a more sophisticated temporal expression extraction system, which can also extract implicit date references, such as *"a year after"* or *"in the same period"*, along with the explicit ones and, for this reason, would presumably be able to extract more dates. As temporal expression extraction system we used HeidelTime [168], the top-ranked in TempEval-3 challenge [181]. We refer to this approach as TEE Heidel.

6.3.2 Filtering (Flt)

We assume that the list of all extracted years gives a probability mass function. We first filter outliers out from it using the Median Absolute Deviation [73, 95] with a parameter (γ) that controls the size of the acceptance region for the outlier filter. This parameter is particularly important to filter out present and future references, invariably present in encyclopaedic descriptions. For example, in the sentence "Volta also studied what we now call electrical capacitance", the word *now* would be resolved to '2014' by temporal expression extraction systems, but it should be discarded as an outlier when discovering of Volta's temporal footprint.

6.3.3 Fitting normal distribution (FND)

A normal distribution is then fitted on the filtered probability mass function. Lower and upper bounds for a temporal footprint are predicted according to two supplementary parameters, α and β. More specifically, the α parameter controls the width of the normal distribution by resizing the width of the Gaussian bell. The β parameter controls the displacement (shift) of the normal distribution. For example, in the case of Wikipedia pages about people, typically this parameter has a negative value (e.g. -5 or -10 years) since the early years of life are rarely mentioned in an encyclopaedic description. We compute the upper and lower bounds of a temporal footprint using the formula $(\mu + \beta) \pm \alpha\sigma$.

We experimented with the following settings:

a) The *TEE RegEx* strategy consists of extracting all possible dates by using the regular expression previously mentioned and by assigning to the lower and upper bound the earliest and the latest extracted year respectively.

b) In the *TEE RegEx + Flt* approach, we first discard outliers from the extracted dates and then the earliest and latest dates are used for lower and upper bounds.

c) For the *TEE RegEx + Flt + FND* strategy, we use the regular expression-based extraction method and then apply filtering and Gaussian fitting.

d) Finally, for the *TEE Heidel + Flt + FND* setting, we use HeidelTime to extract dates from the associated articles. We than apply filtering and Gaussian fitting.

The parameters α, β and γ are optimised according to a Mean Distance Error (MDE) specifically tailored for temporal intervals (see Appendix 6.A: Error measure), which intuitively represents the percentage of overlap between the predicted

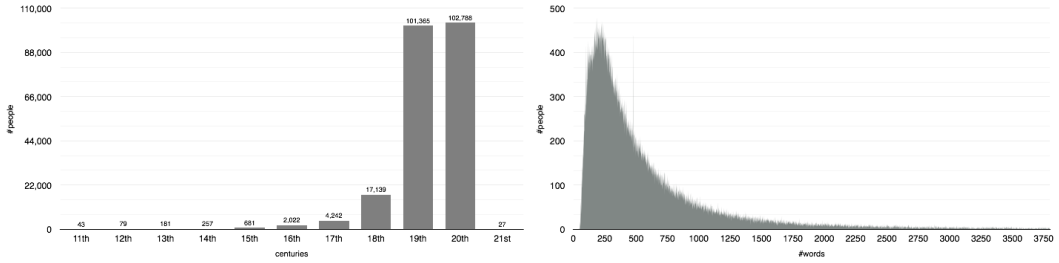
intervals and the gold ones. For each approach we optimised the parameters α, β and γ by using an exhaustive GRID search on a randomly selected subset of 220 people.

6.4 Data

We applied the methodology on people's Wikipedia pages with the aim of measuring the performance of the proposed approaches. We define a person's temporal footprint as the time between their birth and death. This data has been selected in virtue of the availability of a vast amount of samples along with their curated lower and upper bounds, which are available through DBpedia [14]. DBpedia was used to obtain a list of Wikipedia web pages about people born since 1000 AD along with their birth and death dates[1]. We checked the consistency of dates using some simple heuristics (the death date does not precede the birth date, a person age cannot be greater than 120 years) and discarded the incongruous entries. We collected 228,824 people who lived from 1000 to 2014. The Figure 6.2a shows the distribution of people according to the centuries, by considering people belonging to a particular century if they were born in it.

As input to our method, we used associated web pages with some sections discarded, typically containing temporal references invariably pointing to the present, such as *External links*, *See also*, *Citations*, *Footnotes*, *Notes*, *References*, *Further reading*, *Sources*, *Contents* and *Bibliography*. The majority of pages contain from 100 to 500 words (see Figure 6.2b).

[1]We used the data set `Persondata` and `Links-To-Wikipedia-Article` from DBpedia 3.9 (`http://wiki.dbpedia.org/Downloads39`)



(a) Distribution of Wikipedia pages per century.

(b) Distribution of Wikipedia pages per length (in words).

Figure 6.2: Exploratory statistics about the test set extracted from DBpedia.

6.5 Results

Figure 6.3 depicts the application of the proposed method to the Galileo Galilei's Wikipedia article. The aggregated results with respect to the MDE are showed in Table 6.1. The TEE Reg + Flt setting outperforms the other approaches. Still, the approaches that use the Gaussian fitting have lower standard deviation.

These results in Table 6.1 do not take into account the unbalance in the data due to the length of pages (the aggregate numbers are heavily unbalanced towards short pages i.e. those with less than 500 words, as depicted in Figure 6.2b). We therefore analysed the results with respect to the page length (see Figure 6.4). TEE RegEx method's performance is negatively affected by the length of the articles. The longer a Wikipedia page is, the worse the prediction is. This is expected as longer articles are more likely to contain references to the past or future history,

Strategy	**Mean Distance Error**	**Standard Deviation**
TEE RegEx	0.2636	0.3409
TEE RegEx + Flt	**0.2596**	0.3090
TEE RegEx + Flt + FND	0.3503	0.2430
TEE Heidel + Flt + FND	0.5980	0.2470

Table 6.1: Results of the four proposed approaches.

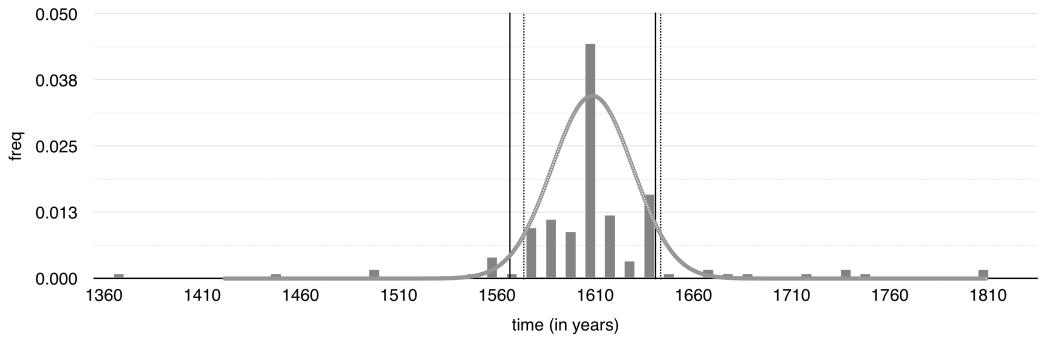


Figure 6.3: Graphical representation of the output on Galileo Galilei's Wikipedia page. Vertical continuous lines represent the prediction of temporal footprint boundaries, whereas dotted lines represent the real date of birth and death of the Italian scientist. The histogram shows the frequency of mentions of particular years in Galilei's Wikipedia page. The Gaussian bell is plotted in light grey.

whereas in a short article the dates explicitly mentioned are often birth and death only. The use of the filter (*TEE RegEx+Flt*) generally improves the performance. The approaches that use the Gaussian fitting provide better results in case of longer texts. Still, in spite of its simplicity, the particular regular expression used in this experiment proved to be effective on Wikipedia pages and consequently an exceptionally difficult baseline to beat. Although counter-intuitive, *TEE RegEx + Flt + FND* performs slightly better than the HeidelTime-based method, suggesting that complex temporal information extraction systems do not bring much of useful mentions. This is in part due to the English Wikipedia's Manual of Style [190] which explicitly discourages authors from using implicit temporal expressions (e.g. *now*, *soon*, *currently*, *three years later*) or abbreviations (e.g. *'90*, *eighties* or *17th century*). Due to this bias, we expect a more positive contribution from using a temporal expression extraction system, when the methodology is applied on texts written without style constraints.

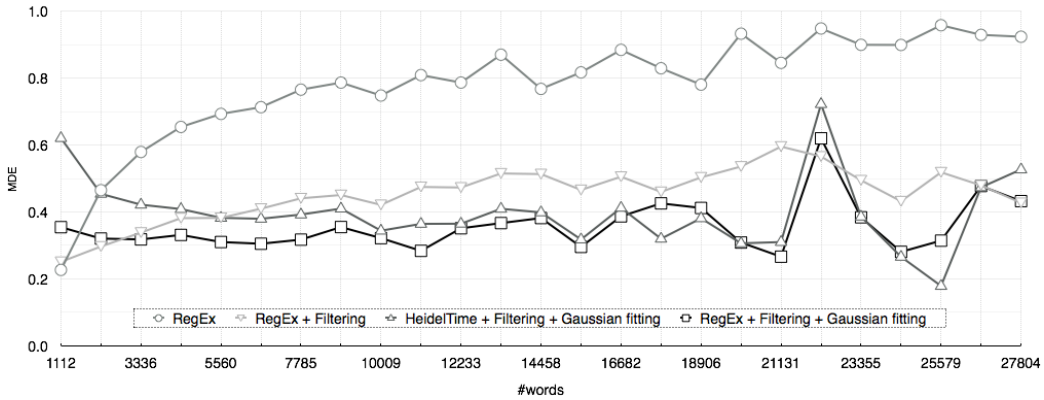


Figure 6.4: Observed error of the four proposed approaches with respect to the length of Wikipedia pages (the lower the better). Each data point represents the average of each bin. The *TEE RegEx* setting generally provide a very high error which is correlated with the page's length. The use of the outlier filter sensibly improves the performance (*TEE RegEx + Flt*). The approach *TEE RegEx + Flt + FND* is better than *TEE Heidel + Flt + FND* especially with short and medium size pages. The spike near 22000 words is due to a particular small sample.

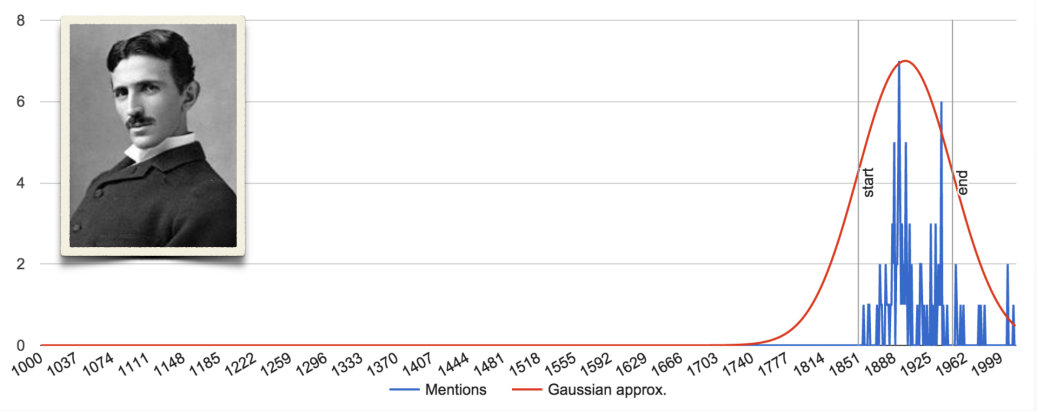
6.6 Discussion

The filtering component helps to ignore the noisy dates from the one belonging to the person's life span. The Nikola Tesla's prediction is explanatory (see Figure 6.5). The component filters out some temporal references which are posterior to the the inventor's life span, leading to an accurate prediction. When the filtering step is not performed, all the date mentions are taken into account to compute the prediction and this leads to a less accurate prediction.

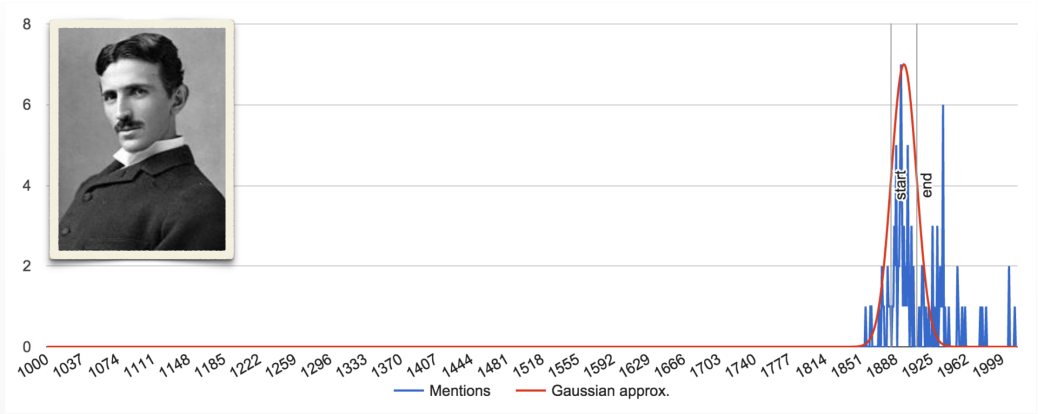
We now examine some typical cases of erroneous predictions to pinpoint the weaknesses of the methodology.

Figure 6.6a shows the prediction for Christopher Columbus. The predicted width of the Gaussian bell is too large, since its Wikipedia page contains many temporal expressions referring to facts happened after his life span[2]. The presence

[2]The historical role played by Christopher Columbus in the discovery of the American continent has been at the centre of a legal dispute at the beginning of the 20th century.

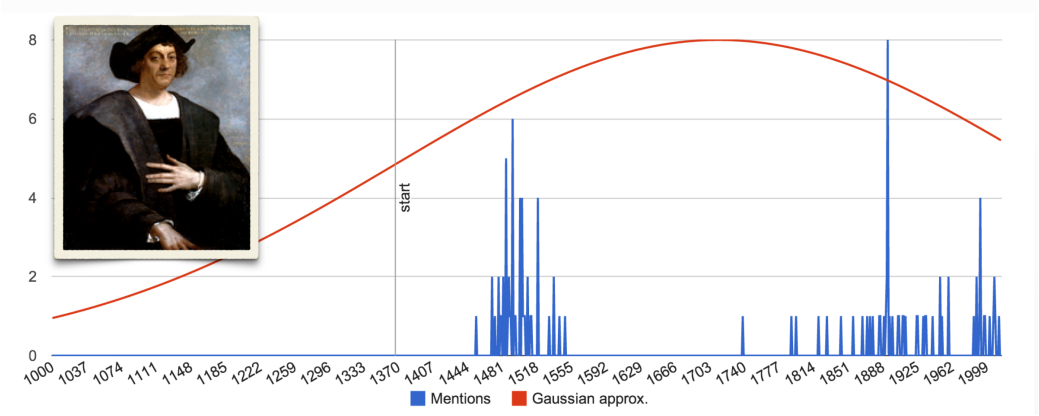


(a) With filtering. Prediction: *[1850-1948]*, MDE: 0.1111

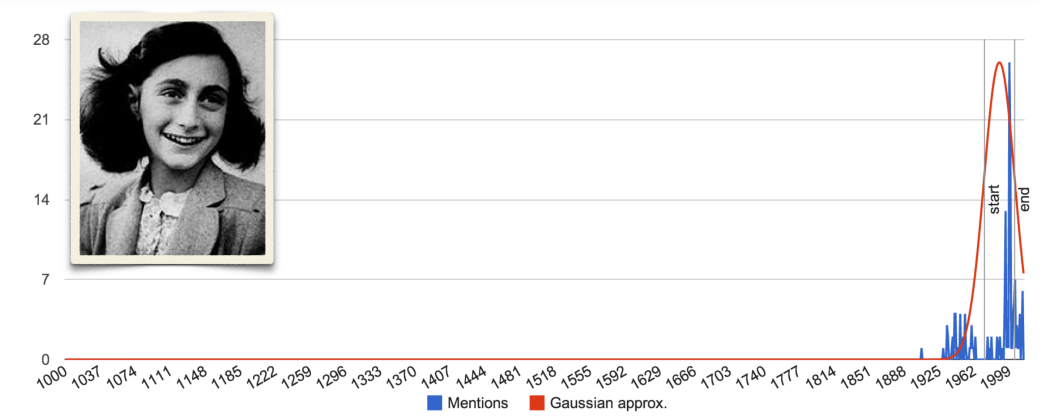


(b) Without filtering. Prediction: *[1882-1909]*, MDE: 0.6818

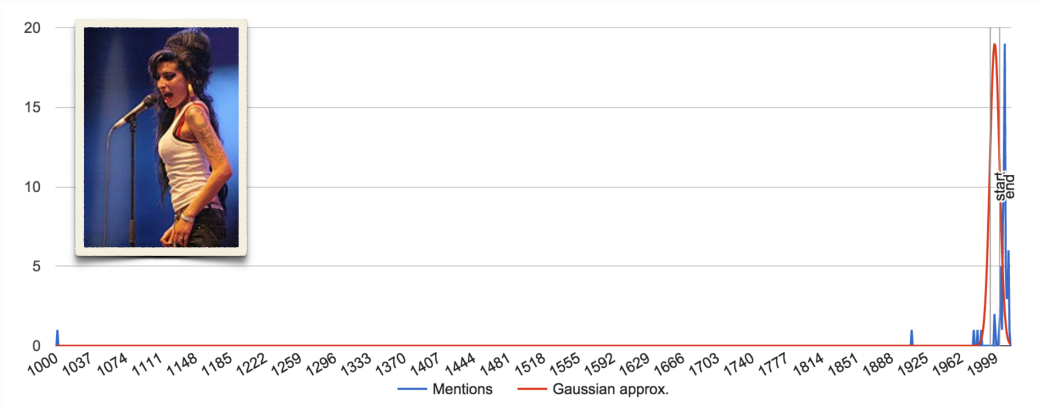
Figure 6.5: Impact of the filtering on the Wikipedia page of Nikola Tesla (1856-1943).



(a) Christopher Columbus (1451-1506) predicted as *[1362-2052]*, MDE: 0.9186



(b) Anna Frank (1929-1945) predicted as *[1962-2010]*, MDE: 1.0000



(c) Amy Winehouse (1983-2011) predicted as *[1992-2002]*, MDE: 0.6207

Figure 6.6: Example of erroneously predicted temporal footprints.

of those historical facts, not belonging to his life span, made the Gaussian fitting method unable to generate a good prediction. According to the system, Christopher Columbus would have been born during the year 1366 and will die in 2057.

Figure 6.6b shows the case of Anna Frank's prediction, which is similar to many others in the dataset. Here the life span of the person has a marginal coverage in the text with respect to an event which is posterior to her death[3]. As a result of this, the Gaussian bell shifts on the right and determines a larger error.

The last case is the one of Amy Winehouse (see Figure 6.6c), where the prediction is anterior to the correct life span. Unlike in the previous two examples, where multiple spikes are visible in the data, here a single spike is presented. The Gaussian bell fits the data correctly but the presence of some wrongly annotated temporal expressions near the years 1000 and 1925 moves the prediction on the left. Also, since her Wikipedia page contains several references to the period immediately before and after her death the predicted life span is shorten than the correct one.

To illustrate potential we show the predicted temporal footprints for four persons (see Figure 6.7).

The methodology here presented has some limitations, which are mainly related to the following points:

- The proposed methodology assumes that the distribution of persons' dates is Gaussian. According to the data presented here, skewed distributions may be better suited to approximate persons' life spans, since the first years of life are generally not covered as the later years are.

- The number of samples on which the parameters are estimated affects the Mean Distance Error. The parameters used for the experiments here pre-

[3]The publication and resonance of her wartime diary *The Diary of a Young Girl* which has been the basis for several plays and films.

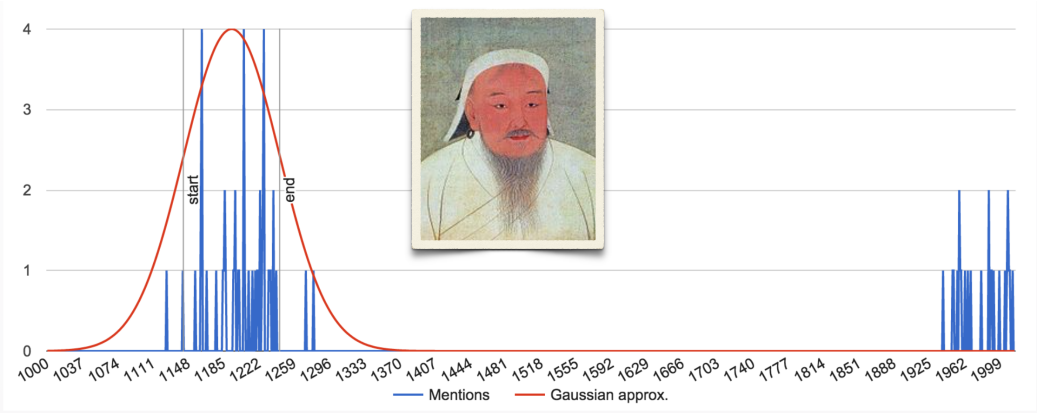
sented have been estimated on 220 randomly selected persons. Such data set represents less than a 100th of the total number of persons considered in this study.

- The assumption that the same parameters would work for all the persons, regardless for the century they belong and the length of their encyclopedic text, may be false. The study showed that the length of the page and the century are two important factors. In particular, the latter is a reflex of the former, since Wikipedia pages of contemporary people tend to be longer than non-contemporary ones.

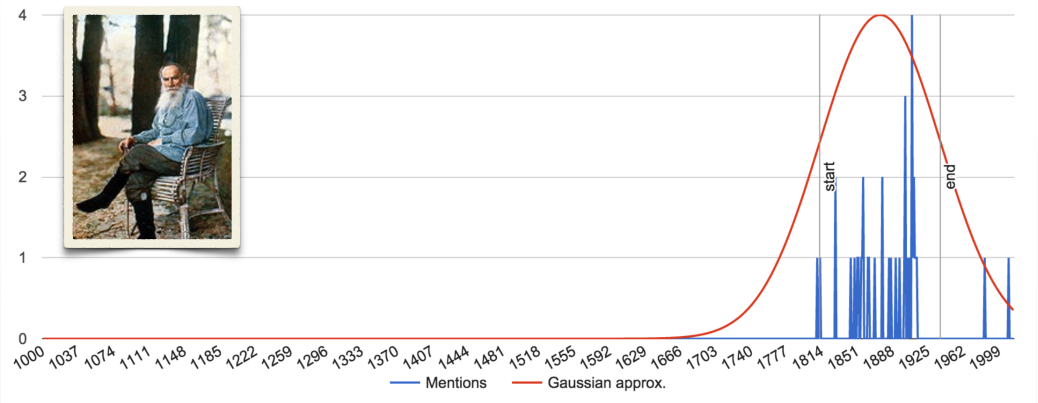
- The temporal expression extraction component produces a number of irrelevant dates. Some of them are irrelevant because related to historical periods which lay outside the boundary of the person life span. Some others are TEE false positives: expressions erroneously identified and normalised as temporal.

6.7 Conclusions

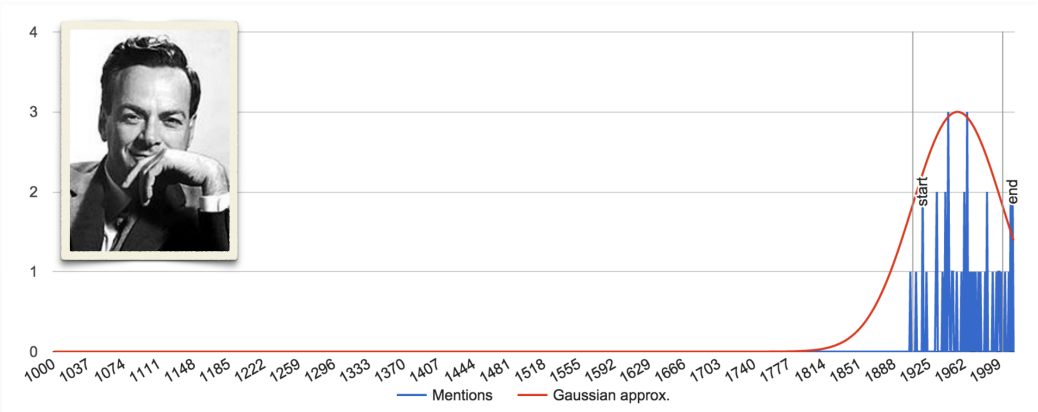
In this paper we introduced a method to extract temporal footprints of concepts based on mining their textual encyclopaedic description. The proposed methodology uses temporal expression extraction techniques, outlier filtering and Gaussian fitting. Our evaluation on people in Wikipedia showed encouraging results. We found that the use of a sophisticated temporal expression extraction system shows its strength only for long textual descriptions, whereas a simple regular expression-based approach performs better with short texts (the vast majority in Wikipedia pages).



(a) Genghis Khan (1162-1227) predicted as *[1142-1243]*, MDE: 0.3529



(b) Leo Tolstoy (1828-1910) predicted as *[1811-1937]*, MDE: 0.3465



(c) Richard Feynman (1918-1988) predicted as *[1907-2002]*, MDE: 0.2604

Figure 6.7: Tests on three different concept using the proposed approach. Concept name is followed by the gold birth and death date, predictions and error.

The notion of temporal footprint has not to be interpreted strictly. A more factual interpretation of temporal footprint could be explored, such as temporal projection of a person's impact in history. This would allow to distinguish between people that made important contribution for the future history from those who did not. The predicted interval of Anna Frank's Wikipedia page depicted in Figure 6.6b is an example of that.

The online demo of this project is available at: `http://www.cs.man.ac.uk/~filannim/projects/temporal_footprints/`. At the same address, we also provide the data, source code, optimisation details and supplementary graphs to aid the replicability of this work.

Focussing on the person's impact on the history rather than its physical existence could better support question answering systems to solve perhaps even more natural queries. For example, in the case of the query "American politicians during the Margaret Thatcher's era", evidently we are not interested in the footprint of the physical existence of the former Prime Minister of the United Kingdom, rather we are interested in the time period in which she was in charge.

Acknowledgements

The authors would like to thank the reviewers for their comments. This paper has greatly benefited from their suggestions and insights. MF would like to acknowledge the support of the UK Engineering and Physical Science Research Council (EPSRC) in the form of doctoral training grant.

Appendix 6.A: Error measure

In interval algebra, the difference between two intervals, $[A]$ and $[B]$, is defined as $[A] - [B] = [A_L - B_U, A_U - B_L]$ (where the subscripts $_L$ and $_U$ indicate lower and upper bound respectively). Unfortunately, this operation is not appropriate to define error measures, because it does not faithfully represent the concept of deviation [124].

We therefore rely on distances for intervals, which objectively measure the dissimilarity between an observed interval and its forecast [13]. In particular, we used De Carvalho's distance [42]:

$$d_{DC}([A],[B]) = \frac{d_{IY}^{\lambda}([A],[B])}{w([A] \cup [B])},$$

where $w([A] \cup [B])$ denotes the width of the union interval, and $d_{IY}^{\lambda}([A],[B])$ denotes the Ichino-Yaguchi's distance defined as follows:

$$d_{IY}^{\lambda}([A],[B]) = w([A] \cup [B]) - w([A] \cap [B]) + \lambda(2w([A] \cap [B]) - w([A]) - w([B])).$$

The Mean Distance Error (MDE) based on De Carvalho's distance is defined by:

$$MDE = \frac{1}{n}\sum_{t=1}^{n} \frac{d_{IY}^{\lambda=0}([A_t],[B_t])}{w([A_t] \cup [B_t])} = \frac{1}{n}\sum_{t=1}^{n} \frac{w([A_t] \cup [B_t]) - w([A_t] \cap [B_t])}{w([A_t] \cup [B_t])},$$

where n is the number of total samples. We set $\lambda = 0$ because we do not want to control the effects of the inner-side nearness and the outer-side nearness between the intervals.

The absence of any intersection between the intervals leads to the maximum

error, regardless to the distance between the two intervals. A predicted interval far from the gold one has the same error of a predicted interval very close to the gold one, if they both not even minimally overlap with it.

Chapter 7

Discussion

"History as well as life itself is complicated – neither life nor history is an enterprise for those who seek simplicity and consistency."
– Jared Diamond, *Collapse: How Societies Choose to Fail or Succeed*

This thesis has presented contributions to the field of TIE from both a methodological and applicative perspective. This chapter will firstly answer to the research questions stated in Section 1.2, and then discuss the proposed methodology.

7.1 Answers to the main research questions

7.1.1 Data-driven temporal information extraction

Data-driven approaches have been extensively tested in Information Extraction (IE) tasks and, for some of them, they currently constitute the state-of-the-art approach. Some of these techniques have been successfully used to identify temporal expressions and events, whereas the normalisation phase is almost exclusively approached with rule-based systems. The results from several TIE challenges [184, 185, 181] seem to suggest that rule-based approaches perform better than data-driven ones.

This thesis has presented a successful data-driven TIE methodology (see Chapter 3) in which the identification phase is based on ML classifiers and the temporal expression normalisation is carried out using a rule-based normaliser (NorMA). The identification phase is improved by the use of an *a posteriori* label-adjustment component which improves the ML-based predictions according to a probabilistic approach.

The method presented here has been officially benchmarked at the main TIE challenge [181]. The best submitted run out-performed all the other data-driven approaches and some of rule-based ones. In particular, the results show that there are no performance differences in the identification phase between rule-based systems and data-driven approaches. This was not the case at the previous editions of the same challenge [184, 188]. However, tackling the normalisation problem with data-driven approaches is still challenging (a complete discussion on the subject will be presented in Section 8.2.1).

7.1.2 Extensive attribute selection

This thesis has also addressed the challenge of investigating which type of linguistic features are beneficial for the Temporal Expression Extraction (TEE) task.

The literature in TIE reports several types of linguistic features (see Chapter 2), although often no statistical analysis of their effectiveness is provided. This problem has been specifically addressed in Chapter 3. Through an extensive literature review, the features typically used in TIE have been harvested and grouped according to their linguistic type: morphological, syntactic, gazetteer-based and knowledge-based (by using WordNet). The groups have been combined in four models and a rigorous model selection has been performed with the aim of measuring, within a statistical framework, the effectiveness of the models. The results showed that the use of WordNet-based features, as used in the literature,

negatively affects the performance in TEE. They also show that adding syntactic features and/or gazetteers on top of morphological features does not provide any statistical significant difference (negative or positive). This last finding indicates that, such features are computationally expensive and can be discarded without affecting the extraction performance.

The impact of this study is twofold. On one side, it gathers and summarises the most common features used in the literature. On the other side, it provides a statistically reliable indication of what types of features positively contribute to the TIE.

7.1.3 The role of silver data

With the advent of fairly accurate NER system, creating silver data is becoming more and more common [141]. As a consequence, the research question of whether such bigger corpora lead to better systems is arising interest [72, 181]. The work presented in this thesis has shed light on such research question in the context of temporal information.

The result of this study show that the use of automatically annotated data does not lead to systems which are better than the ones trained on gold data, even if the latter are notably smaller. Such result does not show that there are not other ways of using silver data which lead to better models [139, 125].

A similar result has already been found in different NLP tasks [83]. At the same time, experiments performed in the event extraction seem to suggest the opposite [181], leading to the conclusion that the utility of silver data needs to be studied per-task and is strictly related to the particular strategy used to learn from them. This further justifies our investigation on the TEE task.

7.1.4 Further improve data-driven predictions

The work presented in this thesis has pushed forward the capabilities of data-driven systems by introducing an *a posteriori* label adjustment module (see Chapter 3) on top of the CRF-based system. The use of such module improved the extraction performance, making it comparable to rule-based systems.

The main idea behind this module is to fix some invalid predicted sequences and correct them by using unigrams frequencies. This idea proved to be effective and boosted the identification performance of about 3%[1] in terms of $F_{\beta=1}$ measure. The results from TempEval-3 [181] prove the efficacy of this module. In fact, the runs submitted without *a posteriori* label-adjuster perform analogously to the rest of the systems.

7.1.5 TIE domain adaptation

The methodology tested in the general domain (newswire data) has been successively ported to the clinical domain (see Chapter 4). The data has presented several challenges due to the presence of typos, ungrammatical sentences and ambiguous punctuation usage. These characteristics made the TIE task harder since the currently available pre-processing tools (see Section 2.1) are not tailored for such irregularities. Moreover, the clinical sub-language involves a broader definition of clinical event (diseases, treatments, symptoms, department names, etc.) [169]. For these reasons, a set of dedicated CRF models has been trained, each for a different event type, and a further post-processing layer has been designed for the adjudication process. The TEE task proved to be challenging due to the lexical variability of expressions. For example, the clinical expression *postoperative day #5* has 12 different lexical variations (*pod 5*, *postoperative 5*, *postoperative d. 5*,

[1]The improvement is measured over several different tests and is statistically significant.

etc.) which convey exactly the same meaning[2]. The normalisation phase required an extension of NorMA [54] to include clinical temporal expressions: e.g. Latin abbreviations used for medication administration (i.e. *bid*: twice a day, *qd*: every day, *qds*: four times a day, etc.), and relative temporal anchoring with respect to clinical events (i.e. *postoperative day 5*, *hospital day #4*, *transfer day*).

Figure 7.1 shows the increment of accuracy for TYPE, VALUE and MODIFIER attributes as the number of update iterations increases. The initial level of accuracy refers to NorMA, the general-domain normaliser.

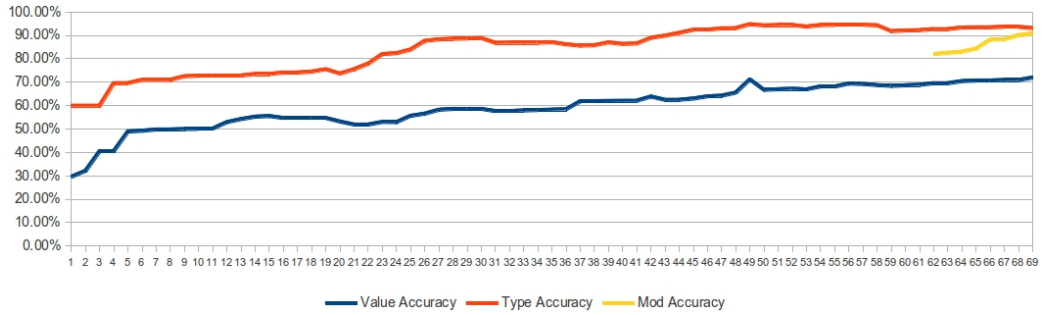


Figure 7.1: Number of normalisation rules vs. accuracy in the clinical domain

7.1.6 Novel TIE applications

Applications of TIE have played an important role since the beginning of the field [6]. Following this tradition, this thesis also explores some new applications of TIE to support the prediction of (I) temporal orientations of search engines' queries (see Chapter 5), and (II) temporal footprints of persons from their Wikipedia pages (see Chapter 6).

The temporal footprint prediction is based on the temporal expressions extracted in the Wikipedia pages according to four different strategies, one of which uses a

[2]It has been noticed that some clinical annotators would count the days from the operation day, whereas others starts from the day after. This leads to some inevitable errors in the temporal expression normalisation phase

TIE technique. The results show that a simpler regular expression-based approach is more effective with short Wikipedia pages.

The TIE methodology has also been tested to support the prediction of the temporal orientation class (past, present, future or atemporal) of search engines' queries. Data-driven approaches have been used (Support Vector Machines(SVMs) and Random Forests(RFs)) with different features, showing that the ones based on TIE techniques outperform the other approaches.

7.1.7 Large-scale experimental setting

As described in Chapter 3, CRF is a complex ML classifier which predicts sequences of labels for sequences of samples. Since the prediction of a single label is made on the premises of its neighbours (according to a factor graph), it results in an exponential expansion of the number of features and more time required for a training-test cycle. For example, the initial 93 attributes described in Table 3.1 and 3.2 generated more than 18 millions of features in the experiments performed in Chapter 3 (6 millions in those performed in Chapter 4) leading to an 8-hour long training-test cycle. In order to control overfitting and obtain a reliable estimation of the models' performance, the experiments have been 10-fold cross-validated and repeated 5 times [58].

Such setting posed time-performance challenges mainly at the level of parallelisation. By using three servers with a total of 48 dedicated cores, the entire computation have been performed in less than 3 weeks. To the best of my knowledge, there are no studies in TIE which use such robust experimental setting.

7.2 Limitations

Although the TIE methodology presented in this thesis provides a process to reliably extract temporal expressions and events from text, there are some limitations. Those are mainly related to the following points:

- the need of large annotated data.
- the coupling with the particular annotation schema.
- the need of small manual adaptation for cross-language porting.

Each of these limitations will be discussed in the following sub-sections.

7.2.1 No data, no party

The TIE methodology presented in this thesis relies on supervised ML approaches. For this reason, it requires annotated data in order to perform the learning phase. According to the complexity of the task, manual annotations can be very expensive in terms of time and money, and in some cases the annotation strategy itself affects the capacity of ML techniques to learn [135]. This is the case when the annotations are characterized by low Inter Annotator Agreement (IAA). Finally, ML algorithms can generalise from the data only when those are enough [69, 70, 7].

Although the methodology is designed to be re-trained on new languages or sub-languages, the assumption about the availability of annotated data in the destination language holds. In situations in which there are no annotation schemas already available it may be difficult to meet this requirement, since one should not only annotate the data, but also carefully design a convenient temporal annotation schema.

7.2.2 Different data implies different annotation schemas

The definition of what temporal information is depends on the annotation guidelines, the annotation schema used to manually annotate the data and more in general, by the task.

Since the methodology presented in this thesis relies on ML supervised approaches, the learned models are the result of an optimisation procedure which maximises the fit of the predictions with respect to the data.

This characteristic prevents the methodology to be applied to sources differently annotated because of their different annotation schemas. The three major temporally annotated corpora currently available follow three different annotation schema and cannot be merged together. WikiWars [108], for example, is annotated using the TIMEX2 tag which carries a stricter definition of temporal expression than the TIMEX3 tag used in TimeBank [179]. On the other hand, the i2b2/2012 corpus [169] is a collection of clinical notes annotated according to the clinical definition of temporal expression and event.

For these reasons, once the methodology here proposed is applied to a particular data set, it cannot be expected to reliably work on a different schema. The opposite it also true: the methodology will not be able to learn from multiple schemas even if it is trained on a merge of all the training data available.

7.2.3 Cross-language porting

The strategy for TIE presented in Chapter 3 and 4 is based on data-driven approaches, but it still partially relies on the use of gazetteers and a rule-based normalisation strategy. When such methodology needs to be ported to other languages or sub-languages (domains), these components have to be manually translated or updated accordingly.

The gazetteers used in ManTIME are related to:

- festivity names (*Christmas*, *Easter*, *Thanksgiving*, *Halloween*, *Epiphany*, etc.).
- temporal signals (*after*, *about*, *since*, *until*, etc.).
- temporal adverbs (*next*, *regularly*, *lately*, *continually*, *daily*, etc.).
- temporal prepositions (*by*, *than*, *to*, *during*, *for*, etc.).
- temporal adjectives (*late*, *soon*, *early*, etc.).

Except for the festivity names, the rest of them includes words which have been identified by linguists for the temporal role they have in English. These expressions need to be ported only in case of different language.

In addition, some of them can be extended according to the domain. For example, in clinical documents the adjective *postoperative* conveys a temporal meaning (a period or time point following a surgical operation). The inclusion of such term and its abbreviated forms (*pod* and *p.o.d.*) would positively contribute to the TIE.

The normalisation component needs to be adapted to the new language too. This process involves the adaptations of the rules, formalised in this research by using regular expressions. This is the most time-consuming activity since the deletion and creation of some new rules may be necessary, along with the update of most of them. Recent developments in this area suggest that the use of machine translation algorithms can provide a baseline normalisation system in all languages [167].

This chapter has illustrated how the work carried out during the PhD has answered its research questions. The limitations of the proposed general methodology have also been presented.

The following chapter will summarise the thesis contributions and present some of the possible new researches worth to be investigated for the advancement of the TIE field. Before the final conclusions, I will present a real case of temporal information analysis which has the purpose of showing how complex such task really is.

Chapter 8

Conclusions and future work

> *"Where does it all lead? What will become of us? These were our young questions, and young answers were revealed. It leads to each other. We become ourselves."*
>
> – Patti Smith, *Just kids*

The automatic extraction of temporal information from written texts is pivotal for many Natural Language Processing (NLP) applications such as question answering, text summarisation and information retrieval. It allows filtering information and inferring temporal flows of events. This thesis focussed on Information Extraction (IE) and presents a novel data-driven Temporal Information Extraction (TIE) strategy which has been tailored to be domain independent. Such methodology has been tested in two linguistic domains: general (see Chapter 3) and clinical (see Chapter 4).

Two novel applications of TIE have also been explored. The first one shows that temporal analysis can be used to improve the accuracy of search engines by filtering out temporally irrelevant results with respect to the users' queries (see Chapter 5). The second application shows that it is possible to predict the approximate date of birth and death of persons by temporally analysing the text in their Wikipedia pages (see Chapter 6).

8.1 Thesis contributions

This thesis successfully provides a method for the extraction of temporal expressions and events from texts written in English. The methodology can be applied to any domain, provided the availability of annotated data. The method consists in training Machine Learning (ML) models to recognise temporal expressions and events, whereas the normalisation phase is entirely accomplished through an ad-hoc rule-based system. Eventually, the predictions are automatically improved via a data-driven component. The method has been officially benchmarked in two of the most important text mining challenges where it obtained very successful results.

Applications of temporal information extraction systems include the improvement of summarisation, question answering, retrieving and filtering information. Although the complexity of temporal information is far from being tamed by automatic computer systems, this thesis provides a stepping stone towards that goal.

This thesis has provided the following contributions:

ManTIME An automatic text mining pipeline for the TIE of texts written in English, which integrates a rule-based temporal expression normaliser (NorMA). The system adopts different strategies for the identification and normalisation phases. The former is based on a ML sequence labelling classifier, Conditional Random Field (CRF), which learns a model based on linguistic attributes extracted during the pre-processing phase. The predictions generated by the CRFs are adjusted by means of an a posteriori label adjuster component. It is a data-driven pipeline which fixes erroneous sequence labels and provides a statistically significant improvement in terms of identification performance. The normalisation part is tackled using NorMA, a rule-based

system which extends a state-of-the-art system [178] with 40 new regular expression-based rules.

ManTIME has been officially benchmarked at TempEval-3 and its best run ranked 5th as best performing ML-based TIE system.

ManTIME is freely available on-line[1] as open-source code and can also be used through its online web interface.

Clinical ManTIME The ManTIME extension for clinical narratives, which includes Clinical NorMA. The system has been ported to the clinical domain and trained on the i2b2/2012 data [169]. This operation had a different impact on the identification and normalisation components. In the identification case, ManTIME has been adapted to read the annotated data in the i2b2 format and then it has been re-trained. Since the clinical definition of event is broader than the one used in the general domain (typically just verbs), a ML component per each type of event has been used. The results have been merged by using some heuristics. The clinical normalisation system extends NorMA: it adds 66 new rules which cover typical clinical temporal expressions and medical Latin abbreviations with temporal meaning.

The methodology has been officially benchmarked at i2b2/2012 where the best submitted run in the Temporal Expression Extraction (TEE) task ranked 1st, and the best run in the event extraction task ranked 5th.

Feature type analysis An extensive analysis has been performed in order to investigate the TEE performance variation with respect to all the feature types used in the literature. The analysis focusses on 93 different linguistic attributes which have been gathered by harvesting the literature in TIE. The attributes have been categorised in 4 different types: morpho-lexical, syn-

[1] http://www.cs.man.ac.uk/~filannim/projects/mantime

tactic, gazetteer-based and WordNet-based. The attributes have been implemented in the ManTIME system and a model selection has been performed on different arrangements of the before mentioned types. The results show that the morpho-lexical features already used in the literature are sufficient to provide the best performance. Moreover, adding other types of attributes to the optimal set does not improve the performance. Finally, the use of WordNet-based attributes has determined a detriment of performance. This conclusion is far from suggesting not to use WordNet since our experiment was constrained by the way WordNet has been previously used in the literature. Better ways of using the same resource may still be investigated. The analysis has required several weeks of computation in a distributed environment and represents, to the best of my knowledge, the largest feature study in TIE.

Silver data investigation As part of TempEval-3 [181], silver data were made available to the participants. Silver data are large corpora annotated by using state-of-the-art systems rather than human experts. The availability of this new resource in the field arise the question of whether or not those data helped to train better TIE systems. The runs submitted to the challenge included variations of the system trained on gold data, silver data and both. The results in the identification phase show that larger silver data set, as used in the experiment, does not lead to better performance.

Temporal footprint prediction A *temporal footprint* is the set of all the temporal expressions (dates and times) referred to a particular entity. We investigated whether the use of a TEE system on Wikipedia pages would allow us to automatically estimate persons' life span on the time line. An ad-hoc error measure has been proposed, along with four different methodologies, one

of which uses a state-of-the-art TEE system. The results indicate that the length of the page is an important factor to determine which technique to apply. For short pages, a simple TEE approach based on a regular expression matching provided better results than those based on HeidelTime. On the contrary, the TEE system provided a lower prediction error for longer pages. The analysis of the erroneously predicted life spans shed light on some methodological limitations: the assumption of normal distribution and the size of the validation set for the parameters optimisation.

Although the methodology has been tested on people's life spans, mainly for testing convenience, it can be applied to a multitude of different types of entities (e.g. companies, historical events, artifacts, etc.), for which the temporal spans are not immediately available as structured information.

The prediction system can be tested at `http://www.cs.man.ac.uk/~filannim/projects/temporal_footprints`, where the open source code is also available.

Temporal orientation A methodology that allows to predict the temporal orientation of search engines' queries by using TIE techniques. We investigated the role of each feature set with respect to the best performing ML-based classifier: Random Forests. The study shows that by including TIE-based features it is possible to improve the overall classification performance.

8.2 Future work

The work presented in this thesis attempted to answer specific research questions in the field of TIE. At the same time, it opens some new questions which are worth investigating for the advancement of the field.

8.2.1 Data-driven temporal expression normalisation

The ISO-TimeML standard [133] is the temporal annotation schema of reference in the community. It defines what temporal information is and specifies how to annotate it. Such specifics influence the applicability of data-driven approaches and their limits [135]. The normalisation problem has not yet been tackled with data-driven approaches because there are no annotated data (either in the general and clinical domain), that provide the necessary level of detail to make algorithms learn the normalisation task.

According to the annotation standards (ISO-TimeML and i2b2 annotation guidelines agree on this point), the attribute VALUE of a temporal expression is meant to be its ISO 8601 representation (i.e. *tomorrow* $\xrightarrow{value}$ `17-08-2015`, *August 2010* $\xrightarrow{value}$ `08-2010`, *every two days* $\xrightarrow{value}$ `P2D`).

In the case of deictic and anaphoric temporal expressions, which refer to an external point in time, they are characterised by having two semantics: local and global [107]. The local semantics is unrelated to its reference time and corresponds to the temporal meaning of the expression. For example, the local semantics of the expression *the next day* is *1 day after the reference time*. The global semantics, on the other hand, can be determined only when the reference time is known. If such expression was found in a document written on the 21st of February 2013, then its global semantics would have been *22-02-2013*.

To put it simply, deictic and anaphoric temporal expressions are normalised in a two-step process, which is completely hidden in the current annotation standards. In fact, annotators have been asked to provide the ISO 8601 representation of temporal expressions, which corresponds to the global semantics only. For example, the expression *three days after* in the sentence *A missing couple have been* ***[found]***$_{EVENT}$ *in a crashed car {three days after}*$_{TIMEX}$ *the* ***[accident]***$_{EVENT}$ *was first* ***[reported]***$_{EVENT}$ *to police.* has always the same local temporal meaning (*three days*

after the event time), but infinite global semantics depending on the event time (when the accident has been first reported to police).

The current annotation standards do not affect the fully-qualified temporal expressions (see Section 2.2), since for them local and global semantics coincide, meaning that their normalisation is invariant with respect to reference times. For this reason, expressions like *21st July 1978* will always be normalised in *21-07-1978*, no matter what the context is.

By annotating the relation between a temporal expression and its reference time, corpora can be used to learn how to select the appropriate speech, reference and event time [143] using data-driven approaches. Modern normalisation systems cope with this by taking into account an external parameter, called *utterance time*. This parameter is a date that is draconianly assumed to always correspond to the Document Creation Time (DCT) or in some other cases to the previously normalised date in the document [165, 92].

8.2.2 Alternative temporal expression normalisation metrics

The error measure currently adopted for the temporal expression normalisation task (ISO 8601) is accuracy, expressed as the ratio between correctly normalised expressions and the total number of expressions. An expression is considered correctly normalised when its ISO 8601 representation (value attribute) is exactly equal, character-by-character, to the gold one.

Although accuracy provides an estimation of error, it does not take into account the temporal interpretation, resulting in a very strict and sometimes wrong error measure. This happens in the following cases:

- less specific annotation: “2011-04-18” normalised as “2011-04-XX”.
- more specific annotation: “FUTURE_REF” normalised as “2017”.

- same temporal meaning, but different representation: “P24H” normalised as “P1D”.

For example, the ISO 8601 expression “P24H” represents a duration of 24 hours, whereas “P1D” represents a duration of 1 day. These expressions are conveying the same temporal meaning, though using two different representations. This is the case in which the accuracy measure as specified so far will consider the prediction wrong.

Also, the binary nature of the normalisation error (correct/incorrect) prevents to discriminate between serious and soft normalisation errors. Referring to the previous example, predicting “P1D” is arguably better than predicting “21-07-1985”.

The challenge here is to design a more temporally sounded error measure for the ISO 8601 standard. Ideally, such measure should be expressed on a continuous interval, where higher values of error correspond to pair of temporal expressions which are temporally very different from each other. I believe such measure should also take into account the different types of temporal expressions (dates, times, durations and sets) and provide a unified way to deal with errors among the possible combinations. In particular, how the error is computed when a date is wrongly normalised as a duration?

8.3 A long way to the top

The problem of interpreting temporal information is much deeper than we can appreciate and represent with the current annotation schemas. Its main source of challenges stands in the resilience of natural language. The main purpose of this section is re-scaling the temporal information extraction problem in the light of its fully linguistic complexity, rather than its Computer Science (CS) simplification.

In the following excerpt, taken from an article published by the BBC News on 18th March 2015, events and temporal expressions are highlighted according to what the ISO-TimeML standard expects.

> {18 March 2015}$_{\text{TIMEX}}$, Vancouver. The director of Google's self-drive car project has [**revealed**]$_{\text{EVENT}}$ {this morning}$_{\text{TIMEX}}$ his motivation for ensuring that the technology [**is**]$_{\text{EVENT}}$ standard on roads within {five years}$_{\text{TIMEX}}$. Chris Urmson [**told**]$_{\text{EVENT}}$ delegates at the TED conference that his eldest son [**was**]$_{\text{EVENT}}$ 11-years-old and [**due to take**]$_{\text{EVENT}}$ his driving test in "{four and a half years}$_{\text{TIMEX}}$". "My team are [**committed to making sure**]$_{\text{EVENT}}$ that doesn't happen," he [**said**]$_{\text{EVENT}}$. "Some 1.2 million people are [**killed**]$_{\text{EVENT}}$ on the roads around the world {each year}$_{\text{TIMEX}}$. That number [**is**]$_{\text{EVENT}}$ equivalent to a jet [**falling out**]$_{\text{EVENT}}$ of the sky {every day}$_{\text{TIMEX}}$." The incremental changes some car-makers are [**introducing**]$_{\text{EVENT}}$ [**are**]$_{\text{EVENT}}$ not enough, he [**said**]$_{\text{EVENT}}$. "That is not to say that driver-assistance cars won't [**be**]$_{\text{EVENT}}$ useful but if we are really [**going to make**]$_{\text{EVENT}}$ changes to our cities, [**get rid**]$_{\text{EVENT}}$ of parking lots, we [**need**]$_{\text{EVENT}}$ self-drive cars," he [**said**]$_{\text{EVENT}}$.

The excerpt includes several narrative devices typically used in English written texts, such as direct speech (quoted sentences), indirect speech (like the second sentence in the excerpt), anaphoric and cataphoric references ('his' refers to 'The director of Google's self-drive car project', which, in turn, refers to 'Chris Urmson'), and more generally deixis: time deixis (e.g. the word *now* in "It is raining {now}$_{\text{TIMEX}}$." refers to a time which is relative to the time of utterance) and discourse deixis (e.g. the word *that* in "My team are committed to making sure that doesn't happen" refers to a problem mentioned before).

While people typically cope with such complexity by using background knowledge, computers need to disentangle such linguistic devices in order to work out the temporal flow of the facts narrated in the excerpt. Most of the complexity of the task lies in this. Simply highlighting events and temporal expressions is not enough to extract the temporal flow of events, but it takes us a bit closer.

For the sake of discussion, consider the following questions in reference to the example provided before:

- when precisely did "this morning" happen? 10:00am, 11:00am? in which time zone?
- is the expression "five years" an approximate or precise duration?
- when Chris said "each year" did he mean the 18th March of "each year"?
- what is the meaning of "every day"? When does such period start? When does it end?
- If Chris succeed in his goal (presumably on 18th March 2020), how could his son do not do the driving test (on 18th September 2019)? By "five years" did he meant less than "four years and half"?

Yet all the those questions, people are perfectly comfortable in understanding such piece of text.

8.4 Final conclusions

This thesis explored methods for the extraction of temporal information from texts. The task of highlighting temporally relevant portions of text is accomplished at a satisfactory level by using data-driven approaches. On the other hand, predicting the temporal meaning of those expressions and consequently anchoring them on

the time-line is more challenging, and has not yet been done with data-driven approaches.

This thesis provides a method for the extraction of temporal expressions and events from texts written in English. The methodology can be applied to any domain, provided the availability of annotated data. The method relies on ML models to recognise temporal expressions and events, whereas the normalisation phase is accomplished through an ad-hoc rule-based system. Eventually, the predictions are automatically adjusted via a data-driven component. The method has been officially benchmarked in two of the key text mining challenges where it obtained successful results.

Applications of temporal information extraction systems include the improvement of summarisation, question answering, retrieving and filtering information. Although the complexity of temporal information is far from being tamed by automatic computer systems, this thesis provides a stepping stone towards that goal.

Bibliography

[1] Steven Abney. *Semisupervised Learning for Computational Linguistics*, chapter 2, pages 14–15. Chapman & Hall/CRC, 1st edition, 2007.

[2] Sisay Fissaha Adafre and Maarten de Rijke. Feature engineering and post-processing for temporal expression recognition using conditional random fields. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pages 9–16, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[3] Klaus-Peter Adlassnig, Carlo Combi, Amar K Das, Elpida T Keravnou, and Giuseppe Pozzi. Temporal representation and reasoning in medicine: research directions and challenges. *Artificial intelligence in medicine*, 38(2):101–113, 2006.

[4] David Ahn, Sisay Fissaha Adafre, and Maarten de Rijke. Towards task-based temporal extraction and recognition. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2005. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[5] David Ahn, Joris van Rantwijk, and Maarten de Rijke. A cascaded machine learning approach to interpreting temporal expressions. In *HLT-NAACL*, pages 420–427, 2007.

[6] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

[7] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[8] Gabor Angeli, Christopher D. Manning, and Daniel Jurafsky. Parsing time: learning to interpret time expressions. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 446–455, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[9] Gabor Angeli and Jakob Uszkoreit. Language-independent discriminative parsing of temporal expressions. In *ACL (1)*, pages 83–92, 2013.

[10] Apache. OpenNLP, 2010.

[11] Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[12] Alan R. Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[13] Javier Arroyo and Carlos Maté. Introducing interval time series: Accuracy measures. *COMPSTAT 2006, proceedings in computational statistics*, pages 1139–1146, 2006.

[14] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[15] Gérard Becher, Françoise Clerin-Debart, and Patrice Enjalbert. A model for time granularity in natural language. In *Proceedings of the Fifth International Workshop on Temporal Representation and Reasoning*, pages 29–, Washington, DC, USA, 1998. IEEE Computer Society.

[16] S. Bethard, L. Derczynski, J. Pustejovsky, and M. Verhagen. Clinical Tempeval. *ArXiv e-prints*, March 2014.

[17] Steven Bethard. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 10–14, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics, Association for Computational Linguistics.

[18] Steven Bethard and James H Martin. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154. Association for Computational Linguistics, 2006.

[19] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[20] Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. Propbank annotation guidelines. 2010.

[21] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.

[22] Svetla Boytcheva and Galia Angelova. A workbench for temporal event information extraction from patient records. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 48–58. Springer, 2012.

[23] Matteo Brucato, Leon Derczynski, Hector Llorens, Kalina Bontcheva, and Christian S Jensen. Recognising and interpreting named temporal expressions. In *RANLP*, pages 113–121, 2013.

[24] David A Campbell and Stephen B Johnson. Comparing syntactic complexity in medical and non-medical corpora. In *Proceedings of the AMIA Symposium*, page 90. American Medical Informatics Association, 2001.

[25] Jaime G Carbonell and Ralf D Brown. Anaphora resolution: a multi-strategy approach. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 96–101. Association for Computational Linguistics, 1988.

[26] Xavier Carreras, Isaac Chao, Lluis Padró, and Muntsa Padró. Freeling: An open-source suite of language analyzers. In *LREC*, 2004.

[27] Tommaso Caselli, Felice dell'Orletta, and Irina Prodanof. TETI: a TimeML compliant TIMEX tagger for Italian. In *IMCSIT'09*, pages 185–192, 2009.

[28] Nate Chambers. NavyTime: Event and time ordering from raw text. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 73–77, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[29] Angel Chang and Christopher D. Manning. SUTime: Evaluation in TempEval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 78–82, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[30] Angel X. Chang and Christopher Manning. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[31] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[32] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543, 2011.

[33] Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. Chinese chunking based on conditional random fields. *NLP*, pages 149–152, 2006.

[34] Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. An empirical study of chinese chunking. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 97–104. Association for Computational Linguistics, 2006.

[35] Jinho D Choi and Martha Palmer. Transition-based semantic role labeling using predicate argument clustering. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 37–45. Association for Computational Linguistics, 2011.

[36] Noam Chomsky. Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3):113–124, 1956.

[37] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[38] Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.

[39] Paul Clough. A perl program for sentence splitting using rules. *University of Sheffield*, 2001.

[40] H Cunningham, D Maynard, K Bontcheva, and V Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[41] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562, 2011.

[42] Fatima De Carvalho. Histogrammes et indices de proximité en analyse données symboliques. *Acyes de l'école d'été sur l'analyse des données symboliques. LISE-CEREMADE, Université de Paris IX Dauphine*, pages 101–127, 1996.

[43] Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard III. Development and evaluation of a clinical note section header terminology. In *AMIA Annual Symposium proceedings*, volume 2008, page 156. American Medical Informatics Association, 2008.

[44] Leon Derczynski and Robert Gaizauskas. USFD2: Annotating temporal expresions and tlinks for tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 337–340. Association for Computational Linguistics, 2010.

[45] Leon Derczynski, Hector Llorens, and Estela Saquete. Massively increasing TIMEX3 resources: A transduction approach. *ArXiv e-prints*, March 2012.

[46] Gaël Harry Dias, Mohammed Hasanuzzaman, Stéphane Ferrari, and Yann Mathet. TempoWordNet for sentence time tagging. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 833–838, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[47] Jay Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

[48] Jason M Eisner. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 340–345. Association for Computational Linguistics, 1996.

[49] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[50] Ronen Feldman and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, December 2006.

[51] Oscar Ferrández, Brett R South, Shuying Shen, F Jeffrey Friedlin, Matthew H Samore, and Stéphane M Meystre. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *Journal of the American Medical Informatics Association*, 20(1):77–83, 2013.

[52] Lisa Ferro, Inderjeet Mani, Beth Sundheim, and George Wilson. TIDES Temporal Annotation Guidelines - Version 1.0.2. Technical report, The MITRE Corporation, McLean, Virginia, June 2001.

[53] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[54] Michele Filannino. Temporal expression normalisation in natural language texts. *CoRR*, abs/1206.2010, 2012.

[55] Michele Filannino, Gavin Brown, and Goran Nenadic. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[56] Michele Filannino and Goran Nenadic. Mining temporal footprints from Wikipedia. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 7–13, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[57] Michele Filannino and Goran Nenadic. Using machine learning to predict temporal orientation of search engines' queries in the temporalia challenge. In *NTCIR-11, EVIA 2014 (NII Testbeds and Community for Information Access Research)*, 2014.

[58] Michele Filannino and Goran Nenadic. Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. *Data & Knowledge Engineering*, 100:19–33, 2015.

[59] Elena Filatova and Eduard Hovy. Assigning time-stamps to event-clauses. In *Proceedings of the Workshop on Temporal and Spatial Information Processing - Volume 13*, TASIP '01, pages 13:1–13:8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[60] Jenny Rose Finkel, Alex Kleeman, and Christopher D Manning. Efficient, feature-based, conditional random field parsing. In *ACL*, volume 46, pages 959–967, 2008.

[61] Alice Gaby. The thaayorre think of time like they talk of space. *Frontiers in Psychology*, 3(300), 2012.

[62] Lucian Galescu and Nate Blaylock. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 715–720, New York, NY, USA, 2012. ACM.

[63] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora*, volume 71, 1998.

[64] Jasdeep Gill, Tim Chearman, Mike Carey, Sukhjinder Nijjer, and Frank Cross. Presenting patient data in the electronic care record: the role of timelines. *JRSM Short Rep*, 1(4):29, 2010.

[65] Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association*, 20(5):820–827, 2013.

[66] TimeML Working Group et al. Guidelines for temporal expression annotation for english for tempeval 2010, 2009.

[67] Claire Grover, Richard Tobin, Beatrice Alex, and Kate Byrne. Edinburgh-LTG: TempEval-2 system description. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 333–336, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[68] H Gurulingappa, M Hofmann-Apitius, and J Fluck. Concept identification and assertion classification in patient health records. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[69] Isabelle Guyon. A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories*, pages 1–11, 1997.

[70] Isabelle Guyon, John Makhoul, Richard Schwartz, and Vladimir Vapnik. What size test set gives good error rate estimates? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):52–64, 1998.

[71] Eun Young Ha, Alok Baikadi, Carlyle Licata, and James C Lester. NCSU: Modeling temporal relations with Markov logic and lexical ontology. In

Proceedings of the 5th International Workshop on Semantic Evaluation, pages 341–344. Association for Computational Linguistics, 2010.

[72] Udo Hahn, Katrin Tomanek, Elena Beisswanger, and Erik Faessler. A proposal for a configurable silver standard. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 235–242. Association for Computational Linguistics, 2010.

[73] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[74] George Hripcsak, Noémie Elhadad, Yueh-Hsia Chen, Li Zhou, and Frances P Morrison. Using empiric semantic correlation to interpret temporal assertions in clinical texts. *Journal of the American Medical Informatics Association*, 16(2):220–227, 2009.

[75] ISO. *ISO 8601:2004 Data elements and interchange formats. Information interchange. Representation of dates and times.*, 2005.

[76] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1148–1158, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[77] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.

[78] Hideo Joho, Adam Jatowt, and Roi Blanco. NTCIR Temporalia: A test collection for temporal information access research. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 845–850, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[79] Hideo Joho, Adam Jatowt, Roi Blanco, H. Naka, and S. Yamamoto. Overview of NTCIR-11 Temporal Information Access (Temporalia) Task. In *Proceedings of the NTCIR-11 Conference*, 2014.

[80] Hyuckchul Jung and Amanda Stent. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 20–24, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[81] Dan Jurafsky and James H Martin. *Speech & language processing*. Pearson Education India, 2000.

[82] Ning Kang, Zubair Afzal, Bharat Singh, Erik M Van Mulligen, and Jan A Kors. Using an ensemble system to improve concept extraction from clinical records. *Journal of biomedical informatics*, 45(3):423–428, 2012.

[83] Ning Kang, Erik M van Mulligen, and Jan A Kors. Training text chunkers on a silver standard corpus: can silver replace gold? *BMC bioinformatics*, 13(1):17, 2012.

[84] Immanuel Kant, Paul Guyer, and Allen W Wood. *Critique of pure reason*. Cambridge University Press, 1998.

[85] Tadao Kasami. An efficient recognition and syntax analysis algorithm for context-free languages. Technical report, DTIC Document, 1965.

[86] Dimitar Kazakov and Suresh Manandhar. Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning*, 43(1-2):121–162, 2001.

[87] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics, 2009.

[88] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[89] Oleksandr Kolomiyets and Marie-Francine Moens. KUL: Data-driven approach to temporal parsing of newswire articles. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 83–87, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[90] Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. JU_CSE_TEMP: a first step towards evaluating events, time expressions and temporal relations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 345–350. Association for Computational Linguistics, 2010.

[91] Anup Kumar Kolya, Amitava Kundu, Rajdeep Gupta, Asif Ekbal, and Sivaji Bandyopadhyay. JU_CSE: A CRF based approach to annotation of temporal expression, event and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 64–72, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[92] Aleksandar Kovačević, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, 20(5):859–866, 2013.

[93] Erdal Kuzey and Gerhard Weikum. Extraction of temporal facts and events from Wikipedia. In *Proceedings of the 2nd Temporal Web Analytics Workshop*, pages 25–32. ACM, 2012.

[94] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[95] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764 – 766, 2013.

[96] Min Li and Jon Patrick. Extracting temporal information from electronic patient records. In *AMIA Annual Symposium Proceedings*, volume 2012, page 542. American Medical Informatics Association, 2012.

[97] Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Methods of information in medicine*, 32(4):281–291, 1993.

[98] Xiao Ling and Daniel S. Weld. Temporal information extraction. In *Proceedings of the AAAI 2010 Conference*, pages 1385 – 1390, Atlanta, Georgia, USA, July 11-15 2010. Association for the Advancement of Artificial Intelligence.

[99] Hector Llorens, Leon Derczynski, Robert J. Gaizauskas, and Estela Saquete. TIMEN: An open temporal expression normalisation resource. In *LREC*, pages 3044–3051, 2012.

[100] Hector Llorens, Borja Navarro, and Estela Saquete. Using semantic networks to identify temporal expressions from semantic roles. In *Proceedings of the International Conference RANLP-2009*, pages 219–224, Borovets, Bulgaria, September 2009. Association for Computational Linguistics.

[101] Hector Llorens, Estela Saquete, and Borja Navarro. TIPSem (english and spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[102] Hector Llorens, Estela Saquete, and Borja Navarro-Colorado. Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management*, 49(1):179 – 197, 2013.

[103] Hector Llorens, Naushad UzZaman, and James Allen. Merging temporal

annotations. In *Temporal Representation and Reasoning (TIME), 2012 19th International Symposium on*, pages 107–113. IEEE, 2012.

[104] Inderjeet Mani and George Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76. Association for Computational Linguistics, 2000.

[105] Christopher D Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer, 2011.

[106] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[107] P. Mazur and R. Dale. LTIMEX: Representing the local semantics of temporal expressions. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 201–208, 2011.

[108] Pawet Mazur and Robert Dale. Wikiwars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922. Association for Computational Linguistics, 2010.

[109] David McClosky. Any domain parsing: automatic domain adaptation for natural language parsing. 2010.

[110] David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, and Christopher D Manning. Combining joint models for biomedical event extraction. *BMC bioinformatics*, 13(Suppl 11):S9, 2012.

[111] David McClosky, Mihai Surdeanu, and Christopher D Manning. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meet-*

ing of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 1626–1635. Association for Computational Linguistics, 2011.

[112] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530. Association for Computational Linguistics, 2005.

[113] Joanna McGregor, Caroline Brooks, Padmaja Chalasani, Jude Chukwuma, Hayley Hutchings, Ronan A Lyons, and Keith Lloyd. Research the health informatics trial enhancement project (HITE): Using routinely collected primary care data to identify potential participants for a depression trial. *Trials*, 11:39, 2010.

[114] Gary H Merrill. The meddra paradox. In *AMIA annual symposium proceedings*, volume 2008, page 470. American Medical Informatics Association, 2008.

[115] Andrei Mikheev. Tagging sentence boundaries. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 264–271. Association for Computational Linguistics, 2000.

[116] Timothy A Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana K Savova. Discovering narrative containers in clinical text. *ACL 2013*, page 18, 2013.

[117] Ruslan Mitkov. *Anaphora resolution*. Routledge, 2014.

[118] Danielle L Mowery, Henk Harkema, and Wendy W Chapman. Temporal annotation of clinical text. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 106–107. Association for Computational Linguistics, 2008.

[119] Dat PT Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. Relation extraction from wikipedia using subtree mining. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1414. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[120] Joakim Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*. Citeseer, 2003.

[121] Philip V Ogren, Guergana Savova, James D Buntrock, and Christopher G Chute. Building and evaluating annotated corpora for medical NLP systems. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1050. American Medical Informatics Association, 2006.

[122] Chris D. Paice. Another stemmer. *SIGIR Forum*, 24(3):56–61, November 1990.

[123] David D Palmer. A trainable rule-based algorithm for word segmentation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 321–328. Association for Computational Linguistics, 1997.

[124] Francesco Palumbo and CarloN. Lauro. A PCA for interval-valued data based on midpoints and radii. In H. Yanai, A. Okada, K. Shigemasu,

Y. Kano, and J.J. Meulman, editors, *New Developments in Psychometrics*, pages 641–648. Springer Japan, 2003.

[125] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[126] Jon D Patrick, Dung HM Nguyen, Yefeng Wang, and Min Li. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association*, 18(5):574–579, 2011.

[127] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

[128] Martin Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[129] Jordi Poveda, Mihai Surdeanu, and Jordi Turmo. A comparison of statistical and rule-induction learners for automatic tagging of time expressions in english. In *In Proc. of the 14th International Symposium on Temporal Representation and Reasoning (TIME 2007*, pages 141–149. IEEE, 2007.

[130] Jordi Poveda, Mihai Surdeanu, and Jordi Turmo. An analysis of bootstrapping for the recognition of temporal expressions. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 49–57, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[131] Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. The penn discourse treebank 2.0 annotation manual. 2007.

[132] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003.

[133] James Pustejovsky, Lee Kiyong, Harry Bunt, Laurent Romary, et al. ISO-TimeML: An international standard for semantic annotation. In *LREC 2010*, 2010.

[134] James Pustejovsky and Amber Stubbs. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics, 2011.

[135] James Pustejovsky and Amber Stubbs. *Natural language annotation for machine learning*. "O'Reilly Media, Inc.", 2012.

[136] John Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[137] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[138] Günter Radden. Spatial time in the west and the east. *Space and Time in Language. Frankfurt: Peter Lang*, 2011.

[139] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[140] Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. 2007.

[141] Dietrich Rebholz-Schuhmann, Antonio José Jimeno-Yepes, Erik M van Mulligen, Ning Kang, Jan A Kors, David Milward, Peter T Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, et al. The CALBC silver standard corpus for biomedical named entities - a study in harmonizing the contributions from four independent named entity taggers. In *LREC*, 2010.

[142] Ruth M Reeves, Ferdo R Ong, Michael E Matheny, Joshua C Denny, Dominik Aronsky, Glenn T Gobbel, Diane Montella, Theodore Speroff, and Steven H Brown. Detecting temporal expressions in medical narratives. *International journal of medical informatics*, 82(2):118–127, 2013.

[143] Hans Reichenbach. Elements of symbolic logic. 1980.

[144] Erin Renshaw, Christopher JC Burges, and Ran Gilad-Bachrach. Selective classifiers for part-of-speech tagging. 2014.

[145] Stefan Rigo and Alberto Lavelli. MulTiSEX - a multi-language timex sequential extractor. In *Temporal Representation and Reasoning (TIME), 2011 Eighteenth International Symposium on*, pages 163–170, 2011.

[146] Livio Robaldo, Tommaso Caselli, Irene Russo, and Matteo Grella. From italian text to timeml document via dependency parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 177–187. Springer, 2011.

[147] Anisa Rula, Matteo Palmonari, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Jens Lehmann, and Lorenz Bühmann. Hybrid acquisition of temporal scopes for RDF data. In *Proc. of the Extended Semantic Web Conference 2014*, 2014.

[148] N Sager, Carol Friedman, E Chi, C Macleod, S Chen, and S Johnson. The analysis and processing of clinical narrative. *MedInfo*, 2:1101–5, 1986.

[149] Wany Sampaio, Chris Sinha, and Vera Da Silva Sinha. *Mixing and mapping: Motion, path, and manner in Amondawa*. na, 2009.

[150] E. Saquete, O. Ferrández, S. Ferrández, P. Martínez-Barco, and R. Muñoz. Combining automatic acquisition of knowledge with machine learning approaches for multilingual temporal recognition and normalization. *Information Sciences*, 178(17):3319 – 3332, 2008.

[151] E. Saquete, R. Muñoz, and P. Martínez-Barco. Event ordering using TERSEO system. *Data & Knowledge Engineering*, 58(1):70 – 89, 2006. Application of natural language to information systems (NLDB04).

[152] Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. Evita: a robust event recognizer for qa systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 700–707. Association for Computational Linguistics, 2005.

[153] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[154] Richard H Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. *Summit on translational bioinformatics*, 2009:116, 2009.

[155] Allen G Schick, Lawrence A Gordon, and Susan Haka. Information overload: A temporal approach. *Accounting, Organizations and Society*, 15(3):199–220, 1990.

[156] Frank Schilder and Andrew McCulloh. Temporal information extraction from legal documents. *Annotating, Extracting and Reasoning about Time and Events*, (05151), 2005.

[157] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[158] Andrea Setzer. *Temporal information in newswire articles: an annotation scheme and corpus study*. PhD thesis, University of Sheffield, September 2001.

[159] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

[160] David Shenk. *Data Smog: Surviving the Information Glut*. Harper San Francisco, 1998.

[161] Carlota S Smith. The syntax and interpretation of temporal expressions in english. *Linguistics and philosophy*, 2(1):43–99, 1978.

[162] Carlota S Smith. Tense and temporal interpretation. *Lingua*, 117(2):419–436, 2007.

[163] Sunghwan Sohn, Kavishwar B Wagholikar, Dingcheng Li, Siddhartha R Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *Journal of the American Medical Informatics Association*, 20(5):836–842, 2013.

[164] Richard Sproat, William Gale, Chilin Shih, and Nancy Chang. A stochastic finite-state word-segmentation algorithm for chinese. *Computational linguistics*, 22(3):377–404, 1996.

[165] Jannik Strötgen and Michael Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[166] Jannik Strötgen and Michael Gertz. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *LREC*, volume 12, pages 3746–3753, 2012.

[167] Jannik Strötgen and Michael Gertz. A baseline temporal tagger for all languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 541–547, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[168] Jannik Strötgen, Julian Zell, and Michael Gertz. HeidelTime: Tuning english and developing spanish resources for tempeval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evalu-*

ation (SemEval 2013), pages 15–19, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[169] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013.

[170] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association*, 20(5):814–819, 2013.

[171] Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. Coupled temporal scoping of relational facts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 73–82, New York, NY, USA, 2012. ACM.

[172] Katrin Tomanek, Joachim Wermter, and Udo Hahn. Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 49–57, 2007.

[173] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Juníchi Tsujii. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392, 2005.

[174] Özlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 2009.

[175] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[176] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[177] N. UzZaman, H. Llorens, and J. Allen. Evaluating temporal information understanding with temporal question answering. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 79–82, 2012.

[178] Naushad UzZaman and James Allen. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[179] Naushad UzZaman and James F Allen. Trios-timebank corpus: Extended timebank corpus with help of deep understanding of text. In *LREC*. Citeseer, 2010.

[180] Naushad UzZaman and James F. Allen. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 351–356, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[181] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint*

*Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[182] Marc Verhagen. Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2):211–241, 2005.

[183] Marc Verhagen. *Drawing TimeML Relations with TBox*. Springer, 2007.

[184] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. SemEval-2007 Task 15: TempEval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80, Prague, 2007.

[185] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179, 2009.

[186] Marc Verhagen, Inderjeet Mani, Roser Sauri, Robert Knippen, Seok Bae Jang, Jessica Littman, Anna Rumshisky, John Phillips, and James Pustejovsky. Automating temporal annotation with TARSQI. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 81–84. Association for Computational Linguistics, 2005.

[187] Marc Verhagen and James Pustejovsky. Temporal processing with the tarsqi toolkit. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, pages 189–192. Association for Computational Linguistics, 2008.

[188] Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[189] Kavishwar B. Wagholikar, Kathy L. MacLaughlin, Michael R. Henry, Robert A. Greenes, Ronald A. Hankey, Hongfang Liu, and Rajeev Chaudhry. Clinical decision support with automated text processing for cervical cancer screening. *Journal of the American Medical Informatics Association*, 19(5):833–839, 2012.

[190] Wikipedia. Wikipedia manual of style, dates and numbers - chronological items, July 2014.

[191] David S Wishart. Drugbank and its relevance to pharmacogenomics. *Pharmacogenomics*, 9:1166–1162, 2008.

[192] Fei Wu, Raphael Hoffmann, and Daniel S. Weld. Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 731–739, New York, NY, USA, 2008. ACM.

[193] Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, I Eric, and Chao Chang. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):849–858, 2013.

[194] Hui Yang, Irena Spasic, John A Keane, and Goran Nenadic. A text mining approach to the prediction of disease status from clinical discharge summaries. *Journal of the American Medical Informatics Association*, 16(4):596–600, 2009.

[195] Daniel H Younger. Recognition and parsing of context-free languages in time n^3. *Information and control*, 10(2):189–208, 1967.

[196] Vanni Zavarella and Hristo Tanev. FSS-TimEx for TempEval-3: Extracting temporal information from text. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 58–63, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[197] Qing Zeng and James J Cimino. Automated knowledge extraction from the UMLS. In *Proceedings of the AMIA Symposium*, page 568. American Medical Informatics Association, 1998.

[198] Ran Zhao, Quang Xuan Do, and Dan Roth. A robust shallow temporal reasoning system. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstration Session*, pages 29–32. Association for Computational Linguistics, 2012.

[199] Li Zhou and George Hripcsak. Temporal reasoning with medical data? A review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202, 2007.

Appendices

Appendix A

Temporal expression normalisation in natural language texts

This chapter is directly adapted from the following paper:

- Michele Filannino. Temporal expression normalisation in natural language texts. *CoRR*, abs/1206.2010, 2012

It is the short and preliminary version of the paper presented in Chapter 3:

- Michele Filannino, Gavin Brown, and Goran Nenadic. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics

Abstract

Automatic annotation of temporal expressions is a research challenge of great interest in the field of information extraction. In this report, I describe a novel rule-based architecture, built on top of a pre-existing system, which is able to normalise temporal expressions detected in English texts. Gold standard temporally-annotated resources are limited in size and this makes research difficult. The proposed system outperforms the state-of-the-art systems with respect to TempEval-2 Shared Task (VALUE attribute) and achieves substantially better results with respect to the pre-existing system on top of which it has been developed. I will also introduce a new free corpus consisting of 2822 unique annotated temporal expressions. Both the corpus and the system are freely available on-line[1].

A.1 Method

The contribution of this paper is twofold. Firstly, I will illustrate a temporal expression corpus explicitly designed for the normalisation phase. Then I will describe the software architecture of a new normaliser built on top of a pre-existing one.

A.1.1 Temporal expressions corpus

Gold-standard temporally-annotated resources are very limited in general domain [45], and even less in specific ones like medical, clinical and biological [62]. In the last decade, different sources of annotated temporal expressions have been developed. Because of the rapid evolution of this research field, usually the sources

[1] http://www.cs.man.ac.uk/~filannim/

differ even with respect to the annotation guidelines. This leads to the existence of different corpora not entirely compatible to each other.

The main difference among them consists in the tag used to annotate temporal expressions: TIMEX2 against TIMEX3. These two tags reflect totally different way of annotating the same temporal expressions leading to the impossibility of using both corpora at the same time.

I created a corpus of temporal expressions collecting all TIMEX3 tags in four different corpora: AQUAINT[2], TimeBank 1.2[3], WikiWars[4] and TRIOS TimeBank v0.1[5]. I extracted from each document all the possible temporal expressions and for each one I also saved the related document creation time, the type (*DATE*, *TIME*, *SET* or *DURATION*) and the normalisation provided by the human annotators. Then I compacted the corpus removing possible duplicates. With the expression *duplicates* I refer to completely identical tuples, i.e. same text, same normalisation, same utterance time and same type.

I obtained a corpus of 2822 unique annotated temporal expressions. The Table A.2 shows an excerpt of the corpus. Further information about the distribution of temporal expression types in it is provided in Table A.1.

The corpus is freely available [6] in CSV format using a tabulation character as delimiter.

A.1.2 Temporal expressions normaliser

I built a new normaliser on top of the one freely available from University of Rochester[7]: TRIOS. It is a rule-based normaliser and it has been proved to provide

[2]http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31/
[3]http://www.timexportal.info/corpora-timebank12
[4]http://www.timexportal.info/wikiwars
[5]http://www.cs.rochester.edu/u/naushad/trios-timebank-corpus
[6]http://www.cs.man.ac.uk/~filannim/timex3s_corpus.csv
[7]http://www.cs.rochester.edu/u/naushad/temporal

Timex type	Frequency
DATE	2307
DURATION	416
TIME	71
SET	28
TOTAL	2822

Table A.1: Distribution of TIMEX3 tags in the corpus.

the second best performance in TempEval-2 Shared Task [179]. All the rules are in the form of regular expressions in a switch architecture: the activation of one of them excludes the activation of all the others.

I introduced a top layer with three new kinds of rules: extension, manipulation and post-manipulation rules.

The extension rules are just new rules that cover non-expected cases and are checked immediately before the pre-existing rules. If a temporal expression do not activate any of the extension rule, it goes into TRIOS. For example, some of these rules are used to normalise expressions of festivities dates such as *"Thanksgiving day"* or *"Saint Patrick's day"*.

The manipulation rules have been introduced to turn particular well-known expressions into an easier form before TRIOS processes them. Once one of these rules is activated, the original temporal expression is transformed into a reduced one that is easier to normalise properly for the pre-existing set of rules. After the transformation, the new temporal expression is taken in input by TRIOS for the normalisation task.

Lastly, I used the post-manipulation rules to solve some deficiencies in the normaliser by adding further information lost by TRIOS and finally improving the performance. In this case the temporal expression is evaluated through the extension rules or the original set. At the end of the normalisation process the

Temporal expression	Type	Value	Utterance
...			
more than two years	DURATION	P2Y	20110926
much of 2010	DATE	FUTURE_REF	20110926
nearly a month	DATE	P1M	20110926
nearly an hour	DURATION	PT1H	19910225
nearly forty years	DURATION	P40Y	1919980120
nearly four years ago	DATE	1994	19980227:081300
nearly three years	DURATION	P3Y	19891030
nearly two months	DURATION	P2M	19980306:131900
nearly two months afterwards	DATE	FUTURE_REF	20110926
nearly two weeks ago	DATE	1989-WXX	19891030
nearly two years	DURATION	P2Y	19980301:141100
next day	DATE	2011-09-27	20110926
...			

Table A.2: Brief excerpt of the corpus.

result is enriched with further information. For example, I used these rules to add information about seasons which are not considered in TRIOS at all.

In the end, I introduced 32 new regular expression patterns: 16 extension rules, 12 manipulation rules and 4 post-manipulation rules. The entire system is freely available online[8] under GNU licence[9].

A.2 Evaluation

I evaluated the normalisation system using the new corpus previously described as a training set and then I measured the performances with respect to the TempEval-2 Shared Task test set. This offered me the possibility of comparing my normaliser with all the others evaluated in that challenge.

In order to measure the difference between TRIOS and my extension I also tested both of them by using the new corpus. It is important to notice that TRIOS has been trained on the same data provided in the new corpus. For this reason a comparison between these systems is legitimate.

In both cases, the evaluation procedure is based on counting. Because the normalisation task is aimed at providing the right TYPE attribute and the right VALUE attribute, the evaluation is carried out by counting how many times the system provides the same value with respect to the human ones. It is important to emphasise that every value provided by the system that differs form the human one for at least one character is considered error.

If this method is quite reasonable for TYPE attribute, it might be too restrictive for VALUE attribute. Some practical examples could be of help to explain the problem.

[8] http://www.cs.man.ac.uk/~filannim/timex_normaliser.zip
[9] http://www.gnu.org/licenses/gpl.html

- The human annotation of a certain timex is *{type: "DATE", value: "FUTURE_REF"}* whereas the system provides a the more specific annotation *{type: "DATE", value: "2013-09-XX"}*.

- The system provides an annotation that is less specific than that provided by humans. For example, it happens when the human-annotation is *{type: "DATE", value: "2011-04-18"}* and the system provide *{type: "DATE", value: "2011-04-XX"}*.

In all these cases the annotations are considered completely wrong. Even when the system provides a partially wrong annotation, e.g. *{type: "DATE", value: "2011-04-23"}* for a human annotation of *{type: "DATE", value: "2011-04-18"}*, considering it a complete wrong result may be too strict because year and month are correct however. This fact has justified the investigation of other measurement metrics [180].

A.2.1 Results

The normalisation results with respect to TempEval-2 Shared Task are shown in Table A.3. The new TRIOS extension outperforms each system in the normalisation of VALUE attributes and performs competitively in the normalisation of TYPE attributes.

The table already shows that the normalisation of value attributes is slightly harder than that of type attributes. The extension of TRIOS outperformed the original system of 2.81% for TYPE attribute and 9.13% for VALUE attribute.

I randomly sub-sampled (400 temporal expressions) the original corpus 10 times and I measured the performances with TRIOS and my extension. I conducted a statistical analysis on the results and I proved that the difference is statistically significant (Willcoxon test), respectively $p = 0.00586$ and $p = 0.0001621$.

The normalisation results with respect to the new corpus are shown in Table A.4.

A.2.2 Error analysis

The original TRIOS normaliser made 1023 value mistakes and 402 type mistakes while its extension respectively made 779 and 323. Through an accurate analysis of the errors, I found plenty of human annotations that seemed to be wrong at first impression. Once I analysed the same annotations taking into account the entire sentence from which each expression had been extracted, I found that the human annotations were actually right. Some examples are shown in Table A.5.

This leads to the conclusion that further improvements are possible only if I consider also the resolution of anaphoric expressions. To do this, it will be necessary to consider a wider window for each temporal expression that takes into account at least the entire sentence in which each temporal expression is located.

A.3 Conclusions

I introduced a new rule-based normaliser of temporal expressions and I showed that it resulted in better performances than the current state-of-the-art system with

	type	value
Edinburgh	0.84	0.63
HeidelTime	0.96	0.85
KUL	0.91	0.55
TERSEO	**0.98**	0.65
TipSem	0.92	0.65
TRIOS	0.94	0.76
TRIOS extension	0.95	**0.86**

Table A.3: Results obtained from TempEval-2 test set.

	type	value
TRIOS	0.8572	0.6257
TRIOS extension	**0.8853**	**0.7170**

Table A.4: Results obtained from the corpus.

	human	**system**
25	1999-04-25	n/a
last year	1988-Q2	1988
three years before	FUTURE_REF	PAST_REF
the summer of 1862	FUTURE_REF	1862-SU
the weekend	P2D	PRESENT_REF

Table A.5: Some errors made by the normaliser.

respect to TempEval-2 Shared Task. I also illustrated the corpus of temporal expressions for normalisation and its purpose. I made both, the normaliser and the corpus, freely available on-line (GNU public licence apply).

A.3.1 Future work

The work presented in this report is the product of a preliminary study in the field of information extraction. The results presented in this report clearly show the necessity of coping with anaphoric temporal expression to substantially enhance the performances of normalisation phase. Currently, the normalisation task takes into account only the temporal expressions, without considering a wider window, such as the entire sentence or a pre-defined number of words after and before the expression. This is required in order to cope with anaphoric expressions.

My long-term goal is to develop novel temporal expressions extraction techniques and use them in clinical domain. Because of the lack of pre-annotated clinical data, I will explore the use of semi-supervised machine learning approaches for the identification phase.

Acknowledgements

I would like to thank Naushad UzZaman from the University of Rochester to have shared his normaliser with the scientific community. I would also like to acknowledge the support of UK Engineering and Physical Science Research Council in the form of doctoral training grant.

Appendix B

Chapter 4 Supplementary Data

This appendix is the supplementary material of Chapter 4, which accompanies the following paper:

- Aleksandar Kovačević, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, 20(5):859–866, 2013

B.1 Results on the training data

This section provides the results of applying different system runs on a subset of the test data. This dataset (referred to as "batch 2") contained 95 narratives and was realized after the first half of the training data ("batch 1", also 95 narratives). The Conditional Random Field (CRF)-models were trained on batch 1 only.

B.2 Error analysis

We provide error analyses of the results obtained by the best performing runs on the training data.

TE recognition. An error analysis identified interesting challenges. For example, around 20% of false positives (FPs) were due to typical date "patterns" used to represent other medical information (e.g. "25/52/70" is an arterial blood gas test result). A significant chunk of false positives (20%) are ambiguous temporal expressions (e.g. "that time", "x 3", "daily", "per day") that are not always annotated as TEs in the gold standard: for example, only 48% of mentions of "[this|the|that] time" were annotated as TEs; similarly, "daily" has only a 68% precision hit rate. On the other hand, false negatives (FNs) included specific TE mentions such as "time of delivery" and "day of transfer", or highly ambiguous mentions such as "now", which were excluded.

TE normalization. The majority of the normalization errors were due to the limited coverage of the rules (e.g. "the course of the night"), the presence of typos (e.g. "the following mornig") and ambiguities (e.g. "this time"). Another source of mistake was a wrong reference time attached to a TE. In addition to occasional errors in the gold standard annotations (e.g. "2017-09-15" normalized as "2019-09-15"), some errors were recorded because of a different normalization code used when compared to the gold standard although the values were equiv-

	Identification						Normalization		
	Strict matching			Lenient matching					
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	Type (%)	Value (%)	Modifier (%)
Run 1	82.58	82.22	82.40	90.59	90.30	**90.44**	82.38	69.25	81.68
Run 2	76.81	**83.00**	79.78	86.04	**93.17**	89.46	**84.86**	**72.05**	**83.93**
Run 3	**83.79**	81.37	**82.56**	**91.61**	89.05	90.31	81.37	68.40	80.67

Table B.1: Temporal expression extraction: micro-averaged results on the training data (batch 2, 95 narratives with 1288 temporal expressions).

	Identification						Normalization	
	Strict matching			Lenient matching				
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	Polarity (%)	Modality (%)
Run 1	**82.56**	74.68	78.42	**89.96**	81.39	85.46	75.19	78.39
Run 2	82.54	**74.71**	**78.43**	89.95	**81.42**	**85.47**	**75.22**	**78.42**

Table B.2: Event extraction: micro-averaged results on the training data (batch 2, 95 narratives with 8903 events).

alent in the temporal sense (e.g. value: PT24H (24 hours) vs. value: P1D (one day)). Furthermore, some errors were due to a non-standardized approach when normalizing expressions such as "postoperative day XX": in some cases, the day of the referent event (e.g. the day of operation) would be day 0, sometimes day 1. This has led to a potential one-day difference between the annotations and the system's predictions.

Event recognition. The errors made by the event recognition module generally fall into three categories. The first category comprises FNs due to the lack of representative features or training data. This is most evident in broadly-scoped classes such as Occurrence and Clinical Department. The segment "gravida 4, para 1" is an example of an FN (for the Occurrence category) where both terms were infrequent in the training data. The second error group is due to our token-level tagging approach, where the CRF contextual features do not always capture enough information. For example, the word "stable" produced a number of FPs because it was mostly annotated as an Occurrence ("stable", "remained stable", "stable condition"), but also as Evidential ("relatively stable", "stable vital signs"), Problem ("stable bleed"), Test ("stable hemodynamics") and Treatment ("a stable dose"). The third error group contains sometimes inconsistent annotations in the gold standard. For example, verb "noted" has been annotated as Evidential 56 times in 40 documents in the gold standard, but we could not explain nine false positives (in just one document).

B.3 Feature impact analysis

The impact that particular groups of features have on the event and temporal expression recognition have been explored in detail. Each of the feature groups (except of the section type) has been removed from the training set, the respective

CRF models were built and applied to the test data. The feature impact results of our second submission (run 2) are presented in tables A3 and A4 (lenient matching), and tables A5 and A6 (strict matching).

Lexical features. The lexical features, in general, are beneficial for the process of extraction of both event and temporal expressions. When this group of features is removed from the event models, there is an overall drop of 4% in precision, recall an F-measure. The impact of lexical features on temporal expression recognition is significantly higher (drop in precision of 8% and a 14% decrease in recall, which results in 11% F-measure drop).Temporal features: As expected, the temporal dictionary feature group only impacts the temporal expression recognition. A 9% drop in recall when this feature group is removed proves that, using a hand-crafted dictionary of temporal terms broadens the scope of TE mentions recognised by the model. A slight increase in precision, without this group, can be explained by the absence of false positives generated when temporal dictionary terms were not annotated as such in the gold standard (due to the dependency on the context or inconsistency in annotation).

Semantic role features. The positive effect of semantic roles on temporal expression recognition reported in the literature 48 is also confirmed, but they do not make any notable difference to event recognition.

Domain features. Surprisingly, the domain features did not have any significant impact on event recognition. A further analysis revealed that the use of the 2010 data reduced the impact of these features; without the additional training data, these features help considerably (data not shown). The semantic features did, however, influence the temporal expression recognition, having a significant impact on recall (a 9% drop) and a slight impact on precision (a 1% drop). This indicates that, as expected, the presence of medical events (problems, test, treatments etc.) at the sentence level is closely related to the presence of clinical temporal expressions.

Frequency and event co-occurrence features. When removed, both frequency and co-occurring events features have similar impact on the Evidential category: the precision is increased (by 2%) while recall drops (by around 3%). The frequency features seem to be beneficial for the recognition of Occurrences. By removing these features, the precision decreases by 3% while recall stays the same. Our assumption that mentions of Occurrence events is linked to other event types was correct, since the recall drops by 4% when the CRF co-occurrence features are removed.

	Events			Temporal expressions		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
All features	**89.35**	**85.32**	**87.29**	**85.38**	**88.05**	**86.70**
No lexical features	87.13	83.88	85.48	75.05	86.75	80.48
No temporal dictionary features	90.19	83.01	86.45	74.89	90.03	81.76
No semantic role features	90.26	82.96	86.45	77.20	89.55	82.92
No semantic features	90.29	82.60	86.27	76.92	89.41	82.70

Table B.3: Event and temporal expression recognition: feature impact analysis of the CRF models on the test data (lenient matching, micro-averaged measures).

	Evidential			Occurrence		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
All features	**65.17**	**75.80**	**70.09**	**66.91**	**63.43**	**65.12**
No frequency features	66.82	72.10	69.36	61.14	67.82	64.31
No co-occurring event features	68.45	72.94	70.63	66.55	59.76	62.97

Table B.4: Impact analysis of frequency and co-occurring events features on the Evidential and Occurrence models for the test data (lenient matching, micro-averaged measures).

	Evidential			Occurrence		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
All features	**65.17**	**75.80**	**70.09**	**55.66**	**52.77**	**54.18**
No frequency features	66.82	72.10	69.36	55.38	49.73	52.40
No co-occurring event features	68.45	72.94	70.63	51.18	56.77	53.83

Table B.5: Impact analysis of frequency and co-occurring event features on the Evidential and Occurrence models for the test data (strict matching, micro-averaged measures).

	Events			Temporal expressions		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
All features	**81.74**	**78.05**	**79.85**	**69.78**	**72.14**	**70.94**
No lexical features	76.63	73.77	75.17	55.93	64.68	59.99
No temporal dictionary features	83.16	76.54	79.71	60.38	72.59	65.93
No semantic role features	83.38	76.63	79.86	62.03	72.04	66.66
No semantic features	83.38	76.25	79.65	61.04	70.98	65.64

Table B.6: Event and temporal expression recognition: feature impact analysis of the CRF models on the test data (strict matching, micro-averaged measures).

B.4 Clinical normalisation rules

Here we present a list of patterns used in Clinical NorMA to match temporal expressions. The following patterns have been extracted from the code and therefore they contain some references to variable names. Please refer to the source code[1] for further details.

[1] `https://github.com/filannim/clinical-norma/blob/master/clinical_norMA.py`

1	('([012]?[0-9]:?[0123456]?[0-9]? ?[a|p]\.?m\.?) ?on ?([0-9][0-9]*[/|-][0-9][0-9]*(?:[/-][0-9][0-9]*)?)$')
2	('^([0-9][0-9])[-|/]([0-9][0-9])$')
3	('([0-9][0-9]?/[0-9][0-9]?)-([0-9][0-9]?/[0-9][0-9]?)/([0-9][0-9]?[0-9]?[0-9]?)')
4	('([0-9][0-9]*[/|-][0-9][0-9]*(?:[/-][0-9][0-9]*)?) ?at ?([012]?[0-9]:?[0123456]?[0-9]? ?[a|p]\.?m\.?)$')
5	('^([0-9][0-9]?)[-|/]([0-9][0-9]?)[-|/]([0-9]\{2,4\})')
6	('^([0-9][0-9]?) ?(?:/|-) ?([0-9][0-9]?)(?: a\.?m\.?| p\.?m\.?)?$')
7	("^([0-9][0-9]?/[0-9][0-9]?) ?(?:-|to) ?([0-9][0-9]?/[0-9][0-9]?)$")
8	('^([0-9][0-9]?[0-9]?) (?:[a-zA-Z]+) (q.*)')
9	('^([0-9][0-9]?[0-9]?)? ?(days?|months?|years?|hours?|minutes?|seconds?|weeks?) ?(?:[a-z]* ?half ?) ago$')
10	('^([0-9]?[0-9]) ?(?:a\.?m\.?|anti meridian)$')
11	("([0-9][0-9]*) (?:days?|months?|years?|hours?|minutes?|weeks?) before")
12	('^([0-9][0-9]?)(?: ?)(?:occasions?)?(?:doses?)?(?:times?)?$')
13	('^([0-9]?[0-9]) ?(p\.?m\.?|pm|post meridian)')
14	('[0-9]\{2\}')[0])
15	('[0-9]\{2\}')[1])
16	('([0-9]\{4\})[-|/]([0-9][0-9]?)[-|/]([0-9][0-9]?)')
17	('^[0-9]\{6\}$')
18	('[0-9]\{8\}')
19	('[0-9]\{8\}T[0-9]\{6\}')
20	('^([0-9a-z\.-/]*) ?(?:(?:of)? the same (?:night|morning|day|period))$')
21	('^([0-9]+) ?(days?|weeks?|months?|years?|hrs?|hours?|mins?|minutes?|secs?|seconds?| d\.?|w\.?|s\.?|y\.?)$')
22	('^([0-9]+) times ?every (days?|weeks?|months?|years?|hours?|minutes?)$')
23	('^([0-9]+) ?(?:-|to) ?([0-9]+) (days?|weeks?|months?|years?|hours?|minutes?)$')
24	('^([0-9]+) ?(?:to|-) ?([0-9]+) ?(years?|yrs?|ys?|months|mns?|ms?|days?|ds?) (?:prior|ago|pta)$')
25	('^([0-9]+)(?:-?| ?)x$')

Figure B.1: Rules extracted from Clinical NorMA's code - part 1

26	('^([0-9\.]+) ?(years?|yrs?|ys?|months|mns?|ms?|weeks?|wks?|ws?|days?|ds?) (?:prior|ago|pta)$')
27	('^(?:a)? ?(an|a couple of|couple of|a'+'|'.join(dt_util.get_literal_nums())+') ?(days?|weeks?|months?|years?|hrs?|hours?|mins?|minutes?|secs?|seconds?) $')
28	('a? ?(?:few|several|lots?|bunch|much|different) (years?|weeks?|months?|days?|hours?|minutes?|seconds?)')
29	('^(a|'+'|'.join(dt_util.get_literal_nums())+')? ?(days?|months?|years?|hours?|minutes?|seconds?|weeks?) ?(?:[a-z]* ?half ?) ago$')
30	('^(alt\.? ?h\.?|every other hours?|alternis horis)$')
31	('^(?:an|a|the)? ?(?:many|lots?|several|a lot of|a number of) (days?|months?|years?|hours?|minutes?|seconds?|weeks?)$')
32	('^(?:an|a|the)? ?(?:many|lots?|several|a lot of|a number of) years? ago$')
33	('^annually$')
34	("(?:'|\')([0-9][0-9]*)")
35	('^(at dinner|dinner ?time|dinner)$')
36	('^(at lunch|lunch ?time|lunch)$')
37	('(autumn)')
38	('^(a)?(year)(-)?(ago)')
39	('(a)?(year)(-|)?(earlier)')
40	('([a-z]\{3\}) ?\([a-z0-9 \.,]+\) ?for ([0-9]+) ?days?')
41	('[a-zA-Z]* (year two)(-|)(thousand)$')
42	('^(?:[a-z]+)?half (?:an)?hour$')
43	('^(b\.?i\.?d\.?|bis in die|twice daily|twice a day|two times [for|by] day)$')
44	('^(b\.?t\.?|bed ?times?|h\.?s\.?|nocte|noct\.?)$')
45	('(columbus day)')
46	('^daily ?\(?daily$')
47	('(\d?\d[/|-]\d?\d[/|-]\d+) (?:at|, at|,) (\d?\d:\d?\d(?: ?[p|a]\.?m\.?)?)')
48	('^(dieb.?[|-]?alt.?|diebus alternis|quoque alternis die|e\.?o\.?d\.?|q\.?a\.?d\.?|q\.?o\.?d\.?)$')
49	('^(earlier)')

Figure B.2: Rules extracted from Clinical NorMA's code - part 2

50	('^(?:early|late) ?(?:on)? ?([a-zA-Z0-9 -/]+)$')
51	('^(?:early)? ?(?:post)? ?(?:-)? ?(operative|extubation) ?(?:course)$')
52	('^(?:every)([0-9]+) (days?|ds?|months?|years?|ys?|hours?|hrs?|minutes?|mns?|seconds?|secs?|weeks?|wks?)$')
53	('every ('+'|'.join(dt_util.get_literal_nums())+') (days?|ds?|months?|years?|ys?|hours?|hrs?|minutes?|mns?|seconds?|secs?|weeks?|wks?)')
54	('^(every other hours?)$')
55	('^(?:few|several|different|many|lots) times? a day$')
56	(get_string_range_numbers(1,31) + ' ' + month_string + ' ' + '.?[12][0-9][0-9][0-9]')
57	('('+'|'.join(dt_util.get_literal_nums())+') (?:days?|months?|years?|hours?|minutes?|weeks?) before')
58	('^('+'|'.join(dt_util.get_literal_nums())+')(?: ?)(?:occasions?)?(?:doses?)?(?:times?)?$')
59	('^('+'|'.join(dt_util.get_literal_nums())+') ?(?:-|to) ?('+'|'.join(dt_util.get_literal_nums())+') (days?|weeks?|months?|years?|hours?|minutes?|seconds?)$')
60	('^('+'|'.join(dt_util.get_literal_nums())+') ?(?:to|-) ?('+'|'.join(dt_util.get_literal_nums())+') ?(years?|yrs?|ys?|months|mns?|ms?|days?|ds?) (?:prior|ago|pta)$')
61	('^('+'|'.join(dt_util.get_literal_nums())+') ?(years?|yrs?|ys?|months|mns?|ms?|weeks?|wks?|ws?|days?|ds?) (?:prior|ago|pta)$')
62	('labou?r day')
63	('^(mane|in the morning)$')
64	('(martin luther king,? ?(?:jr\.?)? day)')
65	('memorial day')
66	('^(?:monday|mon|mo|tuesday|tue|tu|wednesday|wed|we|thursday|thu|th|friday|fri|fr|saturday|sat|sa|sunday|sun|su) ?(?:-|/|,) ?([0-9-/]*)$')
67	("monthly")
68	('('+month_string+') (?:through|to) ('+month_string+')(?: of)? ([0-9][0-9]*)')
69	('^(?:multiples?|many|severals?|differents?|lots?) ?(?:times?|episodes?)?$')
70	('new years? eve')
71	('(noon)')

Figure B.3: Rules extracted from Clinical NorMA's code - part 3

72	('('+numbers_nl+')(-|)('+period_nl+')$')
73	('^(o\.?d\.?|omne in die|every ?day|once a day|once daily|daily|q\.?d\.?|q\.?1d\.?)$')
74	('^(o\.?m\.?)$')
75	('^(o\.?m\.?|every mornings?|omne mane)$')
76	('^(o\.?m\.?|every nights?|omne nocte)$')
77	("^o/n$")
78	('^(one)(|-)(hour)$')
79	('^(one minute)$')
80	('^(one)(|-)(month)$')
81	('^(?:one time|once|one)(?: episode)?$')
82	('^(one)(|-)(week)$')
83	('^(one)(|-)(year)$')
84	('^(o\.?p\.?d\.?|once per day|once a day)$')
85	("overnight")
86	('^p ?([0-9]+) ?([a-z]{2})$')
87	("period")
88	('^per minute$')
89	('^(post cibum|p\.?c\.?|after meals?)$')
90	('^(post meridiem|p\.?m\.?|evening|afternoon?)$')
91	('^q\.?([0-9]+)$')
92	('^q\.? ?([0-9]+)(?:to|-)([0-9]+) ?h$')
93	('^(q\.?a\.?m\.?|every ?day before noon|quaque die ante meridiem)$')
94	('^(q\.?h\.?|every hour|quaque hora)$')
95	('q\.? ?hs')
96	('^(q\.? ?h\.? ?s\.? ?|every nights? at bed ?times?|quaque hora somni)$')
97	('^(q\.?i\.?d\.? ?\(?(?:4|four) times a day ?\)|q\.?d\.?s\.?|q\.?i\.?d\.?|4 times? a day|four times? a day|quater die sumendus|quattuor in die)$')
98	('^q\.? ?('+'|'.join(dt_util.get_literal_nums())+') ?(d\.?|w\.?|mo\.?|y\.?|hours?|minutes?|h\.?)$')
99	('^(q\.?p\.?m\.?|every ?day after noon|quaque die post meridiem|q\.? ?daily|q\.? ?day|each day|per day)$')
100	('^(q\.?q\.?h\.?|every four hours?|quater quaque hora|q\.? ?four)$')

Figure B.4: Rules extracted from Clinical NorMA's code - part 4

101	('^(?:q\.?|quaque) ?([0-9]+) ?(?:d\.?|days?|dies?)$')
102	('^(?:q\.?|quaque) ?([0-9]+) ?-?(?:h\.?|hours?|hora)$')
103	('^(q\.?w\.?k\.?|every weeks?)$')
104	('^(sometime)')
105	('(spring)')
106	('^(stat.?|statim|immediately|present)$')
107	('(summer)')
108	('^(summer|winter|autumn|spring)')
109	('(summer|winter|autumn|spring)')
110	('(thanksgiving day)')
111	('^(?:the) ([0-9-/]+)$')
112	('^[the]?([0-9][0-9]*) nights?')
113	('^(?:the)? ?([0-9][0-9]?)(?:st|nd|rd|th)? ?(-|to|or) ?([0-9][0-9]?)(?:st|nd|rd|th)? (?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? ?(?:days?|nights?|afternoons?)?$')
114	('^(?:the)?([0-9][0-9]?)(?:th|st|nd|rd)? of ('+month_string+')$')
115	('^(?:the)? ?(?:admission|discharge|operation) (?:date|day|night|morning)$')
116	('^(?:the)? ?(?:a\.?m\.?|p\.?m\.?|morning|afternoon|evening) of (.+)$')
117	("^(?:the)?day (?:before|ago|prior)$")
118	('(?:the)?day (?:prior|before) to')
119	("^the days$")
120	("the early on ([0-9][0-9]*)(?:-|/)([0-9][0-9]*)(?: a\.?m\.?| p\.?m\.?)")
121	('(the fiscal)(year|month|day|week|decade)')
122	('^(the following)')
123	('^(the full)')
124	('^(?:the |her |his |their)?([0-9][0-9]*)(?:st|nd|rd|th) (?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? (?:day|night|afternoon)?$')
125	('^(?:the|her|his|their)? ?day of life ?\#? ?([0-9][0-9]*)$')
126	('^(?:the|her|his|their)? ?day of life ?\#? ?('+'|'.join(dt_util.get_literal_nums())+')$')
127	('(?:the |her |his |their)?day (?:of)?(?:the)?admission')
128	('(?:the |her |his |their)?day (?:of)?(?:the)?discharge')

Figure B.5: Rules extracted from Clinical NorMA's code - part 5

129	('(?:the |her |his |their)?day (?:of)?(?:the)?transfer')
130	('^(?:the |her |his |their)?('+'|'.join(dt_util.get_literal_nums())+') (?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? (?:day|night|afternoon)?$')
131	('^(?:the |her |his |their)?(?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? ?(?:day|night|afternoon)?$')
132	('^(?:the |her |his |their)?(?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? ?(?:day |night |afternoon)? ?(?:number|num\.?|\#)? ?([0-9][0-9]*)$')
133	('^(?:the |her |his |their)?(?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? (?:day |night |afternoon)? ?(?:number|num\.?|\#)? ?('+'|'.join(dt_util.get_literal_nums())+')$')
134	('^(?:the |her |his |their)?(?:post-|post|day)? ?(?:pod|operative|op|hospital|hsp|day|hd)(?:ly)? (?:day |night |afternoon)? ?(?:number|num\.?|\#)? ?('+'|'.join(dt_util.get_literal_nums())+'|[0-9][0-9]?) ?(?:and|to) ?('+'|'.join(dt_util.get_literal_nums())+'|[0-9][0-9]?)$')
135	('^the ('+'|'.join(dt_util.get_literal_nums())+') day$')
136	('^[the]?('+'|'.join(dt_util.get_literal_nums())+') nights?')
137	('(?:the)?('+month_string+') (?:of |in)([0-9]+)')
138	('^(?:the)?(?:morning|afternoon|night|evening)? ?(?:of|on)? ?(?:the)? ?([0-9][0-9]*)(?:rd|nd|st|th)$')
139	('^(?:the)?(?:morning|afternoon|night|evening) (?:of|on) (?:the)? ?([a-zA-Z0-9 -/]+)$')
140	('^(?:the)?night (?:before|prior)')
141	('^(?:the|on) ([012]?[0-9])(?: |-)?(?:st|nd|rd|th)?$')
142	('^(?:the|on) ('+'|'.join(dt_util.get_literal_nums())+')$')
143	('^(the past)')
144	("(?:the)?(?:past|previous\last) ([0-9][0-9]*) (days?|months?|years?|hours?|minutes?|weeks?|seconds?)")
145	('(?:the)?(?:past|previous\last) ('+'|'.join(dt_util.get_literal_nums())+') (days?|months?|years?|hours?|minutes?|weeks?|seconds?)')
146	('(the past)(year|month|day|week|decade)')
147	("(?:the)? ?same day")
148	('^(the)(summer|winter|autumn|spring)')
149	("^(?:the|that) day$")
150	('(?:the|that)? ?(?:to)?night')

Figure B.6: Rules extracted from Clinical NorMA's code - part 6

151	('^(?:the|this|that)? ?(a\.?m\.?|ante meridiem|morning|before noon)$')
152	("^the weeks?$")
153	('(the year)$')
154	('^(the)?(year)(\'s|-|)(end)$')
155	('(the)(year|week|decade|hour|month)$')
156	('^(t\.?i\.?d\.? ?\(?(?:3|three) times a day ?\)|t\.?d\.?s\.?|t\.?i\.?d\.?|3 times? a day|three times? a day|ter die sumendum|ter in die)$')
157	('^times? ([0-9]+)(?: day)?$')
158	('^times ('+'|'.join(dt_util.get_literal_nums())+')$')
159	('^(t\.?i\.?w\.?|3 times? a week|three times? a week)$')
160	('^(?:twice|two times|two)(?: episode)?$')
161	('(week)$')
162	("weekly")
163	('(winter)')
164	('^x(?:-?| ?)([0-9]+)$')
165	('^([xivdcml]+)$')
166	("year")
167	('^years?[-|]?old$')

Figure B.7: Rules extracted from Clinical NorMA's code - part 7