Tuning HeidelTime for identifying time expressions in clinical texts in English and French

Thierry Hamon
LIMSI-CNRS, BP133, Orsay
Université Paris 13
Sorbonne Paris Cité, France
hamon@limsi.fr

Natalia Grabar
CNRS UMR 8163 STL
Université Lille 3
59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Abstract

We present work on tuning the Heideltime system for identifying time expressions in clinical texts in English and French languages. The main amount of the method is related to the enrichment and adaptation of linguistic resources to identify Timex3 clinical expressions and to normalize them. The test of the adapted versions have been done on the i2b2/VA 2012 corpus for English and a collection of clinical texts for French, which have been annotated for the purpose of this study. We achieve a 0.8500 F-measure on the recognition and normalization of temporal expressions in English, and up to 0.9431 in French. Future work will allow to improve and consolidate the results.

1 Introduction

Working with unstructured narrative texts is very demanding on automatic methods to access, formalize and organize the information contained in these documents. The first step is the indexing of the documents in order to detect basic facts which will allow more sophisticated treatments (*e.g.*, information extraction, question/answering, visualization, or textual entailment). We are mostly interested in indexing of documents from the medical field. We distinguish two kinds of indexing: conceptual and contextual.

Conceptual indexing consists in finding out the mentions of notions, terms or concepts contained in documents. It is traditionally done thanks to the exploitation of terminological resources, such as MeSH (NLM, 2001), SNOMED International (Côté et al., 1993), SNOMED CT (Wang et al., 2002), etc. The process is dedicated to the recognition of these terms and of their variants in documents (Nadkarni et al., 2001; Mercer and Di Marco, 2004; Bashyam and Taira, 2006; Schulz and Hahn, 2000; Davis et al., 2006).

The purpose of contextual indexing is to go further and to provide a more fine-grained annotation of documents. For this, additional information may be searched in documents, such as polarity, certainty, aspect or temporality related to the concepts. If conceptual indexing extracts and provides factual information, contextual indexing is aimed to describe these facts with more details. For instance, when processing clinical records, the medical facts related to a given patient can be augmented with the associated contextual information, such as in these examples:

(1) *Patient has the stomach aches.*

(2) *Patient denies the stomach aches.*

(3) *After taking this medication, patient started to have the stomach aches.*

(4) *Two weeks ago, patient experienced the stomach aches.*

(5) *In January 2014, patient experienced the stomach aches.*

In example (1), the information is purely factual, while it is negated in example (2). Example (3) conveys also aspectual information (the medical problem has started). In examples (4) and (5), medical events are positioned in the time: relative (*two weeks ago*) and absolute (*in January 2014*). We can see that the medical history of patient can become more precise and detailed thanks to such contextual information. In this way, factual information related to the stomach aches of patient may receive these additional descriptions which make each occurrence different and non-redundant. Notice that the previous I2B2 contests[1] addressed the information extraction tasks related to different kinds of contextual information.

[1] *https://www.i2b2.org/NLP*

Temporality has become an important research field in the NLP topics and several challenges addressed this taks: ACE (ACE challenge, 2004), SemEval (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), I2B2 2012 (Sun et al., 2013). We propose to continue working on the extraction of temporal information related to medical events. This kind of study relies on several important tasks when processing the narrative documents : identification and normalization of linguistic expressions that are indicative of the temporality (Verhagen et al., 2007; Chang and Manning, 2012; Strötgen and Gertz, 2012; Kessler et al., 2012), and their modelization and chaining (Batal et al., 2009; Moskovitch and Shahar, 2009; Pustejovsky et al., 2010; Sun et al., 2013; Grouin et al., 2013). The identification of temporal expressions provides basic knowledge for other tasks processing the temporality information. The existing available automatic systems such as HeidelTime (Strötgen and Gertz, 2012) or SUTIME (Chang and Manning, 2012) exploit rule-based approaches, which makes them adaptable to new data and areas. During a preliminary study, we tested several such systems for identification of temporal relations and found that HeidelTime has the best combination of performance and adaptability. We propose to exploit this automatic systems, to adapt and to test it on the medical clinical documents in two languages (English and French).

In the following of this study, we introduce the corpora (Section 2) and methods (Section 3). We then describe and discuss the obtained results (Section 4.2) and conclude (Section 5).

2 Material

Corpora composed of training and test sets are the main material we work with. The corpora are in two languages, English and French, and has comparable sizes. All the processed corpora are deidentified. Corpora in English are built within the I2B2 2012 challenge (Sun et al., 2013). The training corpus consists of 190 clinical records and the test corpus of 120 records. The reference data contain annotations of temporal expressions according to the Timex3s guidelines: date, duration, frequency and time (Pustejovsky et al., 2010). Corpora in French are built on purpose of this study. The clinical documents are issued from a French hospital. The training corpus consists of 182 clinical records and the test corpus of 120 records. 25 documents from the test set are annotated to provide the reference data for evaluation.

3 Method

HeidelTime is a cross-domain temporal tagger that extracts temporal expressions from documents and normalizes them according to the Timex3 annotation standard, which is part of the markup language TimeML (Pustejovsky et al., 2010). This is a rule-based system. Because the source code and the resources (patterns, normalization information, and rules) are strictly separated, it is possible to develop and implement resources for additional languages and areas using HeidelTime's rule syntax. HeidelTime is provided with modules for processing documents in several languages, *e.g.* French (Moriceau and Tannier, 2014). In English, several versions of the system exist, such as general-language English and scientific English.

HeidelTime uses different normalization strategies depending on the domain of the documents that are to be processed: news, narratives (*e.g.* Wikipedia articles), colloquial (*e.g.* SMS, tweets), and scientific (*e.g.* biomedical studies). The *news* strategy allows to fix the document creation date. This date is important for computing and normalizing the relative dates, such as *two weeks ago* or *5 days later*, for which the reference point in time is necessary: if the document creation date is *2012/03/24*, *two weeks ago* becomes *2012/03/10*.

Our method consists of three steps: tuning HeidelTime to clinical data in English and French (Section 3.1), evaluation of the results (Section 3.2), and exploitation of the computed data for the visualization of the medical events (Section 3.3).

3.1 Tuning HeidelTime

While HeidelTime proposes a good coverage of the temporal expressions used in general language documents, it needs to be adapted to specialized areas. We propose to tune this tool to the medical domain documents. The tuning is done in two languages (English and French). Tuning involves three aspects:

1. The most important adaptation needed is related to the enrichment and encoding of linguistic expressions specific to medical and especially clinical temporal expressions, such as *post-operative day #*, *b.i.d.* meaning *twice a day*, *day of life*, etc.

2. The admission date is considered as the reference or starting point for computing relative dates, such as *2 days later*. For the identification of the admission date, specific preprocessing step is applied in order to detect it within the documents;

3. Additional normalizations of the temporal expressions are done for normalizing the durations in approximate numerical values rather than in the undefined 'X'-value; and for external computation for some durations and frequencies due to limitations in HeidelTime's internal arithmetic processor.

3.2 Evaluating the results

HeidelTime is tuned on the training set. It is evaluated on the test set. The results generated are evaluated against the reference data with:

- precision $\mathcal{P}$: percentage of the relevant temporal expressions extracted divided by the total number of the temporal expressions extracted;
- recall $\mathcal{R}$: percentage of the relevant temporal expressions extracted divided by the number of the expected temporal expressions;
- APR: the arithmetic average of the precision and recall values $\frac{\mathcal{P}+\mathcal{R}}{2}$;
- F-measure $\mathcal{F}$: the harmonic mean of the precision and recall values $\frac{\mathcal{P}*\mathcal{R}}{\mathcal{P}+\mathcal{R}}$.

3.3 Exploiting the results

In order to judge about the usefulness of the temporal information extracted, we exploit it to build the timeline. For this, the medical events are associated with normalized and absolute temporal information. This temporal information is then used to order and visualize the medical events.

4 Experiments and Results

4.1 Experiments

The experiments performed are the following. Data in English and French are processed. Data in two languages are processed by available versions of HeidelTime: two existing versions (general language and scientific language) and the medical version created thanks to the work performed in this study. Results obtained are evaluated against the reference data.

4.2 Results

We added several new rules to HeidelTime (164 in English and 47 in French) to adapt the recognition of temporal expressions in medical documents. Some cases are difficult to annotate. For instance, it is complicated to decide whether some expressions are concerned with dates or durations. The utterance like *2 years ago* (*il y a 2 ans*) is considered to indicate the date. The utterance like *since 2010* (*depuis 2010*) is considered to indicate the duration, although it can be remarked that the beginning of the duration interval marks the beginning of the process and its date. Another complex situation appears with the relative dates:

- as already mentioned, date like *2 years ago* (*il y a 2 ans*) are to be normalized according to the reference time point;
- a more complex situation appears with expressions like *the day of the surgery* (*le jour de l'opération*) or *at the end of the treatment by antiobiotics* (*à la fin de l'antibiothrapie*), for which it is necessary first to make the reference in time of the other medical event before being able to define the date in question.

In Table 1, we present the evaluation results for English. On the training corpus, with the general language version and the scientific version of HeidelTime, we obtain F-measure around 0.66: precision (0.77 to 0.79) is higher than recall (0.56). The values of F-measure and APR are identical. The version we adapted to the medical language provides better results for all the evaluation measures used: F-measure becomes then 0.84, with precision up to 0.85 and recall 0.84. This is a good improvement of the automatic tool which indicates that specialized areas, such as medical area, use indeed specific lexicon and constructions. Interestingly, on the test corpus, the results decrease for the general language and scientific versions of HeidelTime, but increase for the medical version of HeidelTime, with F-measure 0.85. During the I2B2 competition, the maximal F-measure obtained was 0.91. With F-measure 0.84, our system was ranked 10/14 on the English data. Currently, we improve these previous results.

In Table 2, we present the results obtained on the French test corpus (26 documents). Two versions of HeidelTime are applied: general language, that is already available, and medical, that has been developed in the presented work. We can

Versions of HeidelTime	Training				Test			
	$\mathcal{P}$	$\mathcal{R}$	*APR*	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	*APR*	$\mathcal{F}$
general language	0.7745	0.5676	0.6551	0.6551	0.8000	0.5473	0.6499	0.6499
scientific	0.7877	0.5676	0.6598	0.6598	0.8018	0.5445	0.6486	0.6486
medical	0.8478	0.8381	0.8429	0.8429	0.8533	0.8467	0.8500	0.8500

Table 1: Results obtained on training and test sets in English.

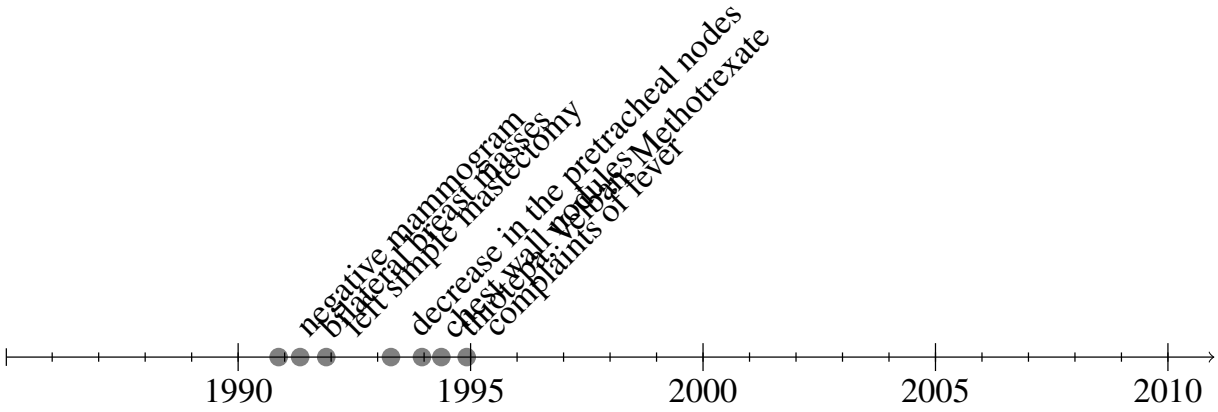

Figure 1: Visualization of temporal data.

Versions of HeidelTime	Test		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
general language	0.9030	0.9341	0.9183
medical	0.9504	0.9341	0.9422

Table 2: Results obtained on test set in French.

observe that the adapted version suits better the content of clinical documents and improves the F-measure values by 3 points, reaching up to 0.94.

The main limitation of the system is due to the incomplete coverage of the linguistic expressions (*e.g. au cours de, mensuel* (*during, monthly*)). Among the current false positives, we can find ratios (*2/10* is considered as date, while it means lab results), polysemous expressions (*Juillet* in *rue du 14 Juillet* (*14 Juillet street*)), and segmentation errors (*few days* detected instead of *the next few days*). These limitations will be fixed in the future work.

In Figure 1, we propose a visualization of the temporal data, which makes use of the temporal information extracted. In this way, the medical events can be ordered thanks to their temporal anchors, which becomes a very useful information presentation in clinical practice (Hsu et al., 2012). The visualization of unspecified expressions (*e.g. later, sooner*) is being studied. Although it seems that such expressions often occur with more specific expressions (*e.g. later that day*).

5 Conclusion

HeidelTime, an existing tool for extracting and normalizing temporal information, has been adapted to the medical area documents in two languages (English and French). It is evaluated against the reference data, which indicates that its tuning to medical documents is efficient: we reach F-measure 0.85 in English and up to 0.94 in French. More complete data in French are being annotated, which will allow to perform a more complete evaluation of the tuned version. We plan to make the tuned version of HeidelTime freely available. Automatically extracted temporal information can be exploited for the visualization of the clinical data related to patients. Besides, these data can be combined with other kinds of contextual information (polarity, uncertainty) to provide a more exhaustive picture of medical history of patients.

Acknowledgments

This work is partially performed under the grant ANR/DGA Tecsan (ANR-11-TECS-012). The authors are thankful to the CHU de Bordeaux for making available the clinical documents.

References

ACE challenge. 2004. The ACE 2004 evaluation plan. evaluation of the recognition of ace entities, ace relations and ace events. Technical report, ACE challenge. http://www.itl.nist.gov/iad/mig/tests/ace/2004.

V Bashyam and Ricky K Taira. 2006. Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005aa. In *AMIA*, pages 26–30.

Iyad Batal, Lucia Sacchi, Riccardo Bellazzi, and Milos Hauskrecht. 2009. A temporal abstraction framework for classifying clinical temporal data. In *AMIA Annu Symp Proc. 2009*, pages 29–33.

Angel X. Chang and Christopher D. Manning. 2012. SUTIME: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.

Roger A. Côté, D. J. Rothwell, J. L. Palotay, R. S. Beckett, and Louise Brochu. 1993. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield.

Neil Davis, Henk Harlema, Rob Gaizauskas, Yikun Guo, Moustafa Ghanem, Tom Barnwell, Yike Guo, and Jon Ratcliffe. 2006. Three approaches to GO-tagging biomedical abstracts. In Udo Hahn and Michael Poprat, editors, *SMBM*, pages 21 – 28, Jena, Germany.

Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. 2013. Hybrid approaches to represent the clinical patient's timeline. *J Am Med Inform Assoc*, 20(5):820–7.

William Hsu, Ricky K Taira, Suzie El-Saden, Hooshang Kangarloo, and Alex AT Bui. 2012. Context-based electronic health recond: toward patient specific healthcare. *IEEE Transactions on information technology in biomedicine*, 16(2):228–234.

Remy Kessler, Xavier Tannier, Caroline Hagge, Vronique Moriceau, and Andr Bittar. 2012. Finding salient dates for building thematic timelines. In *50th Annual Meeting of the Association for Computational Linguistics*, pages 730–739.

Robert E Mercer and Chrysanne Di Marco. 2004. A design methodology for a biomedical literature indexing tool using the rhetoric of science. In *HLT-NAACL 2004, Workshop Biolink*, pages 77–84.

Vronique Moriceau and Xavier Tannier. 2014. French resources for extraction and normalization of temporal expressions with heideltime. In *LREC*.

Robert Moskovitch and Yuval Shahar. 2009. Medical temporal-knowledge discovery via temporal abstraction. In *AMIA Annu Symp Proc*, pages 452–456.

P Nadkarni, R Chen, and C Brandt. 2001. Umls concept indexing for production databases: a feasibility study. *J Am Med Inform Assoc*, 8(1):80–91.

National Library of Medicine, Bethesda, Maryland, 2001. *Medical Subject Headings*. *www.nlm.nih.gov/mesh/meshhome.html*.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Stefan Schulz and Udo Hahn. 2000. Morpheme-based, cross-lingual indexing for medical document retrieval. *Int J Med Inform*, 58-59:87–99.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753. ELRA.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*, 20(5):806–813.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. 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. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

AY Wang, JH Sable, and KA Spackman. 2002. The snomed clinical terms development process: refinement and analysis of content. In *AMIA*, pages 845–9.