Temporal Expressions in Swedish Medical Text – A Pilot Study

Sumithra Velupillai
Department of Computer and Systems Sciences
Stockholm University
Sweden
sumithra@dsv.su.se

Abstract

One of the most important features of health care is to be able to follow a patient's progress over time and identify events in a temporal order. We describe initial steps in creating resources for automatic temporal reasoning of Swedish medical text. As a first step, we focus on the identification of temporal expressions by exploiting existing resources and systems available for English. We adapt the HeidelTime system and manually evaluate its performance on a small subset of Swedish intensive care unit documents. On this subset, the adapted version of HeidelTime achieves a precision of 92% and a recall of 66%. We also extract the most frequent temporal expressions from a separate, larger subset, and note that most expressions concern parts of days or specific times. We intend to further develop resources for temporal reasoning of Swedish medical text by creating a gold standard corpus also annotated with events and temporal links, in addition to temporal expressions and their normalised values.

1 Introduction

One of the most important features of health care is to be able to follow patient progress over time and identify clinically relevant events in a temporal order. In medical records, temporal information is stored with explicit timestamps, but it is also documented in free text in the clinical narratives. To meet our overall goal of building accurate and useful information extraction systems in the health care domain, our aim is to build resources for temporal reasoning in Swedish clinical text. For instance, in the example sentence ***MR-undersökningen av skallen*** *igår visade att den vä-sidiga förändringen i thalamus minskat i volym.* ("The **MRI-scan of the scull** yesterday showed that the left (abbreviated) side change in thalamus has decreased in volume"), a temporal reasoning system should extract the **event** (**MRI-scan of the scull**) and the temporal expression (yesterday), and be able to normalise the time expression to a specific date and classify the temporal relation.

In this pilot study we focus on the identification of temporal expressions, utilising existing resources and systems available for English. A temporal expression is defined as any mention of dates, times, durations, and frequencies, e.g. "April 2nd", "10:50am", "five hours ago", and "every 2 hours". When successfully identifying such expressions, subsequent anchoring in time is made possible.

Although English and Swedish are both Germanic languages, there are some differences that are important to take into account when adapting existing solutions developed for English to Swedish, e.g. Swedish is more inflective and is more compounding than English.

The purpose of this study is to initiate our work on temporal reasoning for Swedish, and to evaluate existing solutions adapted to Swedish. These are our first steps towards the creation of a reference standard that can be used for evaluation of future systems.

2 Background

Temporal reasoning has been the focus of several international natural language processing (NLP) challenges in the general domain such as ACE[1], TempEval-2 and 3 (Verhagen et al., 2010; UzZaman et al., 2013), and in the clinical domain through the 2012 i2b2 challenge (Sun et al., 2013). Most previous work has been performed on En-

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 88–92,
Baltimore, Maryland USA, June 26-27 2014.

glish documents, but the TempEval series have also included other languages, e.g. Spanish. For temporal modelling, the TimeML (Pustejovsky et al., 2010) guidelines are widely used. The TimeML standard denotes events (EVENT), temporal expressions (TIMEX3) and temporal relations (TLINK).

For English, several systems have been developed for all or some of these subtasks, such as the TARSQI Toolkit (Verhagen et al., 2005) and SUTime (Chang and Manning, 2012). Both these tools are rule-based, and rely on regular expressions and gazetteers. The TARSQI Toolkit has also been developed for the clinical domain: MedTTK (Reeves et al., 2013).

In other domains, and for other languages, HeidelTime (Strötgen and Gertz, 2012) and TIMEN (Llorens et al., 2012) are examples of other rule-based systems. These are also developed to be easily extendable to new domains and languages. HeidelTime ranked first in the TempEval-3 challenge on TIMEX3:s, resulting in an $F1$ of 77.61 for the task of correctly identifying and normalising temporal expressions.

HeidelTime was also used in several participating systems in the i2b2 challenge (Lin et al., 2013; Tang et al., 2013; Grouin et al., 2013) with success. Top results for correctly identifying and normalising temporal expressions in the clinical domain are around 66 $F1$ (Sun et al., 2013). The system has also been adapted for French clinical text (Hamon and Grabar, 2014).

3 Methods

The HeidelTime system was chosen for the initial development of a Swedish temporal expression identifier. Given that its architecture is designed to be easily extendible for other languages as well as domains, and after reviewing alternative existing systems, we concluded that it was suitable for this pilot study.

3.1 Data

We used medical records from an intensive care unit (ICU) from the Stockholm EPR Corpus, a clinical database from the Stockholm region in Sweden[2] (Dalianis et al., 2012). Each medical record (document) contains all entries (notes) about one patient a given day. The document contains notes written by both physicians and nurses. They also contain headings (e.g. *Dagantecknning* ("Daily note"), *Andning* ("Breathing")) and timestamps for when a specific note/heading was recorded in the medical record system. These are excluded in this analysis.

Three subsets from this ICU dataset were used: 1) two randomly selected documents were used for analysing and identifying domain specific time expressions and regular expressions to be added in the adaptation of HeidelTime (development set), 2) a random sample of ten documents was used for manual analysis and evaluation (test set), and 3) a set of 100 documents was also extracted for the purpose of empirically studying the types of temporal expressions found in the data by the adapted system (validation set).

3.2 Adaptation of HeidelTime and Evaluation

The available resources (keywords and regular expression rules) in the HeidelTime system were initially translated automatically (Google translate[3]) and manually corrected. Regular expressions were modified to handle Swedish inflections and other specific traits. An initial analysis on two separate, randomly selected ICU notes (development set) was performed, as a first step in adapting for both the Swedish language and the clinical domain.

Results on the system performance were manually evaluated on the test set by one computational linguistics researcher by analysing system outputs: adding annotations when the system failed to identify a temporal expression, and correcting system output errors. A contingency table was created for calculating precision, recall and $F1$, the main outcome measures. Moreover, the top most frequent temporal expressions found by the system on a separate set were extracted (validation set), for illustration and analysis purposes.

4 Results

We report general statistics for the ICU corpus, results from the adaptation and evaluation of HeidelTime for Swedish (HTSwe) on the test set, and the most frequent temporal expressions found by HTSwe in a separate set of 100 ICU documents (validation set).

[2] Study approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5

[3] http://translate.google.se

4.1 Data: ICU corpus

General statistics for the test set used in this study is shown in Table 1. On average, each document consists of 54.6 sentences, and each sentence contains on average 8.7 tokens (including punctuation). We observe that some sentences are very short (min = 1), and there is great variability in length, as can be seen through the standard deviation.

	#	min - max	avg $\pm$ std
Sentences /document	540	35 - 80	54.6$\pm$14.1
Tokens /sentence	4749	1 - 52	8.7$\pm$5.7

Table 1: General statistics for the test set (ten ICU documents) used in this study. Minimum, maximum, average and standard deviation for sentences per document and tokens (including punctuation) per sentence.

4.2 Adaptation and evaluation of HeidelTime: HTSwe

The main modifications required in the adaptation of HeidelTime to Swedish (HTSwe) involved handling definite articles and plurals, e.g. adding *eftermiddag(en)?(ar)?(na)?* ("afternoon", "the afternoon"/"afternoons"/"the afternoons"). From the analysis of the small development set, some abbreviations were also added, e.g. *em* ("afternoon"). Regular expressions for handling typical ways dates are written in Swedish were added, e.g. "020812" and "31/12 -99" (day, month, year). In order to avoid false positives, a rule for handling measurements that could be interpreted as years (e.g. *1900 ml*) was also added (a negative rule).

Results from running HTSwe on the test set are shown in Table 2. HTSwe correctly identified 105 temporal expressions, but missed 55 expressions that should have been marked, and classified 9 expressions erroneously. In total, there are 160 TIMEX3s. Overall performance was 92% precision, 65% recall and $F1$ = 77%.

The main errors were due to faulty regular expressions for times, e.g. *13-tiden* ("around 13 PM) and missing keywords such as *dygn* ("day" - a word to indicate a full day, i.e. 24 hours) and *lunchtid* ("around lunch"). Some missing keywords were specific for the clinical domain, e.g. *efternatten* ("the after/late night", typical for shift indication). There were also some partial errors. For instance, *i dag* ("today") was only included with the spelling *idag* in the system, thus generating a TIMEX3 output only for *dag*.

	TIMEX3 Annotator	**Other** Annotator	$\sum$
TIMEX3 HTSwe	105	9	114
Other HTSwe	55	4580	4635
$\sum$	160	4589	4749

Table 2: Contingency table, TIMEX3 annotations by the annotator and the adapted HeidelTime system for Swedish (HTSwe) on the test set. "Other" means all other tokens in the corpus. These results yield a precision of 92%, a recall of 66%, and $F1$ = 77% for HTSwe.

On the validation set, 168 unique time expressions were found by the system, and 1,178 in total. The most frequent expressions all denote parts of days, e.g. *idag* ("today"), *nu* ("now"), and *natten* ("the night"), see Table 3. Specific times (mostly specific hours) were also very common. Thus, there were many translated expressions in the HeidelTime system that never occurred in the data.

TIMEX3	**N**	%
idag ("today")	164	14%
nu ("now")	132	11%
natten ("the night")	117	10%
morgonen ("the morning")	96	8%
em ("afternoon", abbreviated)	82	7%
kvällen ("the evening")	74	6%
igår ("yesterday")	49	4%
fm ("morning", abbreviated)	34	3%
morgon ("morning")	30	3%
natt ('night")	26	2%
Total	1178	100%

Table 3: Most frequent (top ten, descending order) TIMEX3s found by HTSwe on the validation set (100 ICU documents). Total = all TIMEX3:s found by HTSwe in the entire validation set. There were 168 unique TIMEX3s in the validation set.

5 Discussion and Conclusion

We perform an initial study on automatic identification of temporal expressions in Swedish clinical

text by translating and adapting the HeidelTime system, and evaluating performance on Swedish ICU records. Results show that precision is high (92%), which is promising for our future development of a temporal reasoning system for Swedish. The main errors involve regular expressions for time and some missing keywords; these expressions will be added in our next iteration in this work. Our results, $F1 = 77\%$, are lower than state-of-the-art systems for English clinical text, where the top-performing system in the 2010 i2b2 Challenge achieved 90% $F1$ for TIMEX3 spans (Sun et al., 2013). However, given the small size of this study, results are encouraging, and we have created a baseline system which can be used for further improvements.

The adaptation and translation of HeidelTime involved extending regular expressions and rules to handle Swedish inflections and specific ways of writing dates and times. Through a small, initial analysis on a development set, some further additions and modifications were made, which led to the correct identification of common TIMEX3s present in this type of document. A majority of the expressions translated from the original system was not found in the data. Hence, it is worthwhile analysing a small subset to inform the adaptation of HeidelTime.

The ICU notes are an interesting and suitable type of documentation for temporal reasoning studies, as they contain notes on the progress of patients in constant care. However, from the results it is evident that the types of TIMEX3 expressions are rather limited and mostly refer to parts of days or specific times. Moreover, as recall was lower (66%), there is clearly room for improvement. We plan to extend our study to also include other report types.

5.1 Limitations

There are several limitations in this study. The corpus is very small, and evaluated only by one annotator, which limits the conclusions that can be drawn from the analysis. For the creation of a reference standard, we plan to involve at least one clinician, in order to get validation from a domain expert, and to be able to calculate inter-annotator agreement. The size of the corpus will also be increased. We have not evaluated performance on TIMEX3 normalisation, which, of course, is crucial for an accurate temporal reasoning system. For instance, we have not considered the category *Frequency*, which is essential in the clinical domain to capture e.g. medication instructions and dosages. Moreover, we have not annotated and evaluated *events*. This is perhaps the most important part of a temporal reasoning system. We plan to utilise existing named entity taggers developed in our group as a pre-annotation step in the creation of our reference standard. The last step involves annotating temporal links (TLINK) between events and TIMEX3:s. We believe that part-of-speech (PoS) and/or syntactic information will be a very important component in an end-to-end system for this task. We plan to tailor an existing Swedish PoS tagger, to better handle Swedish clinical text.

5.2 Conclusion

Our main finding is that it is feasible to adapt HeidelTime to the Swedish clinical domain. Moreover, we have shown that the parts of days and specific times are the most frequent temporal expressions in Swedish ICU documents.

This is the first step towards building resources for temporal reasoning in Swedish. We believe these results are useful for our continued endeavour in this area. Our next step is to add further keywords and regular expressions to improve recall, and to evaluate TIMEX3 normalisation. Following that, we will annotate events and temporal links.

To our knowledge, this is the first study on temporal expression identification in Swedish clinical text. All resulting gazetteers and guidelines in our future work on temporal reasoning in Swedish will be made publicly available.

Acknowledgments

The author wishes to thank the anonymous reviewers for invaluable comments on this manuscript. Thanks also to Danielle Mowery and Dr. Wendy Chapman for all their support. This work was partially funded by Swedish Research Council (350-2012-6658) and Swedish Fulbright Commission.

References

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck,

Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In Pierre Nugues, editor, *Proc. 4th SLTC, 2012*, pages 17–18, Lund, October 25-26.

Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. 2013. Eventual situations for timeline extraction from clinical reports. *JAMIA*, 20:820–827.

Thierry Hamon and Natalia Grabar. 2014. Tuning HeidelTime for identifying time expressions in clinical texts in English and French. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 101–105, Gothenburg, Sweden, April. Association for Computational Linguistics.

Yu-Kai Lin, Hsinchun Chen, and Randall A. Brown. 2013. MedTime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, 46:20–28.

Hector Llorens, Leon Derczynski, Robert Gaizauskas, and Estela Saquete. 2012. TIMEN: An Open Temporal Expression Normalisation Resource. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

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. 2013. Detecting temporal expressions in medical narratives. *International Journal of Medical Informatics*, 82:118–127.

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.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *JAMIA*, 20:828–835.

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, Inderjeet Mani, Roser Sauri, Robert Knippen, Seok Bae Jang, Jessica Littman, Anna Rumshisky, John Phillips, and James Pustejovsky. 2005. Automating Temporal Annotation with TARSQI. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, pages 81–84, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.