Analysis of Temporal Expressions Annotated in Clinical Notes

Hegler Tissot
Federal University of Parana, Brazil
hctissot@inf.ufpr.br

Angus Roberts
University of Sheffield, UK
angus.roberts@sheffield.ac.uk

Leon Derczynski
University of Sheffield, UK
leon.derczynski@sheffield.ac.uk

Genevieve Gorrell
University of Sheffield, UK
g.gorrell@sheffield.ac.uk

Marcos Didonet Del Fabro
Federal University of Parana, Brazil
marcos.ddf@inf.ufpr.br

Abstract

Annotating the semantics of time in language is important. THYME is a recent temporal annotation standard for clinical texts. This paper examines temporal expressions in the first major corpus released under this standard. It investigates where the standard has proven difficult to apply, and gives a series of recommendations regarding temporal annotation in this important domain.

1 Introduction

Time provides a substrate for the human management of perception and action. As a pervasive element of human life, time is a primary element that allows us to observe, describe and reason about what surrounds us in the world (Caselli, 2009). As a cognitive and linguistic component for describing changes which happen through the occurrence of events, processes, and actions, time provides a way to record, order, and measure the duration of such occurrences (Bartak et al., 2013).

Understanding temporal information has become crucial for several language processing applications, such as question answering, text summarisation, information retrieval, and knowledge base population. To this end, it is important to develop strong annotation standards and corpora for temporal semantics. Challenges in developing these standards include: a) how to formally represent the elements that describe temporal concepts; and b) what procedures should be performed by an algorithm, in order to deal with the set of temporal reasoning operations that humans seem to perform relatively easily (Caselli, 2009). The sub-problem of automatic recognition of temporal expressions within natural language text is a particularly challenging and active area in computational linguistics (Pustejovsky et al., 2003).

One way of iteratively improving annotation standards and corpora is to use human annotations to test an annotation model (Pustejovsky and Moszkowicz, 2012). This paper provides an analysis of temporal expression annotation in one such corpus, in an effort to gather information on the underlying model and to improve future annotation efforts.

Our analysis is based on the corpus and standard that backed a recent shared annotation exercise in SemEval (Semantic Evaluation) 2015[1]. SemEval is a series of evaluations that aims to verify the effectivenesses of existing approaches to semantic analysis. SemEval-2015 Task 6, Clinical TempEval (Bethard et al., 2015), was a temporal information extraction task over the clinical domain, using clinical notes and pathology reports, focused on identification of spans and features for time expressions (TIMEX), and based on specific annotation guidelines. Clinical TempEval temporal expression

[1] http://alt.qcri.org/semeval2015/

results[2] were given in terms of Precision, Recall and F1-score for identifying spans and classes of temporal expressions. The identification of expressions should be based on a set of provided guidelines.

The clarity of guidelines, skill of annotators and quality of annotated resource can be estimated by measuring agreement between annotators. It is recommended that the target inter-annotator agreement for linguistic resources be at or above 0.90 (Hovy et al., 2006). Clinical TempEval's timex annotations had an IAA of 0.80 (or 0.79) (Styler et al., 2014), suggesting that these can be improved.

To investigate the quality of the dataset and annotation standard in Clinical TempEval, we have used a rule-based system using JAPE (Cunningham et al., 2011) based as closely as possible on the annotation guidelines, and referring to the corpus for guidance in edge cases. When evaluated using the Clinical TempEval scoring software, this system obtained good Recall (0.795 for timex spans and 0.756 for timex classes) but low precision ranging from 0.29 to 0.49. These results are low compared to the state of the art on other temporally annotated corpora.

In order to discover the reason for the low precision, we analysed the differences between our system and the manually-annotated Clinical TempEval corpus. Our analysis demonstrated how difficult it is to create a manually annotated Gold Standard for time expressions and why this problem is still open in computational linguistics. The analysis is based on a methodology composed of six steps, from manual annotation of the input data, to finding and classifying the time expressions, and finally to a classification of the discrepancies found.

This article is organised as follows: Section 2 describes the methodology we used to perform the analysis; Section 3 describes the analysis results and Section 4 gives a series of recommendations in order to guide future temporal annotation in the clinical domain. Section 5 refers to the related work, and Section 6 concludes with final considerations and future work.

2 Methodology

SemEval-2015 Task 6 (Clinical TempEval) was a temporal information extraction task over the clinical domain, using clinical notes and pathology reports for cancer patients provided by Mayo Clinic.[3] Clinical TempEval focuses on identification of: spans and features for timexes, event expressions, and narrative container relations. For time expressions, participants identified expression spans within the text and their corresponding classes: DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET.[4]

Participating systems had to annotate timexes according to the guidelines for the annotation of times, events and temporal relations in clinical notes – THYME Annotation Guidelines (Styler et al., 2014) –, which is an extension of ISO TimeML (Pustejovsky et al., 2010) developed by the THYME project.[5] Further, ISO TimeML extends two other guidelines: a) TimeML Annotation Guidelines (Sauri et al., 2006), and b) TIDES 2005 Standard for the Annotation of Temporal Expressions (Ferro et al., 2005).

For Clinical TempEval two datasets were provided. The first was a training dataset comprising 293 documents with a total number 3818 annotated time expressions. The second dataset comprised 150 documents with a total of 2078 annotations. This was used for evaluation and was then made available to participants, after evaluations were completed. Each annotation identified the span and class of each timex. Table 1 show the number of annotated timex by class in each dataset.

In order to understand why our system achieved such low Precision in the final Clinical TempEval results, we performed an extensive analysis of the manually annotated time expressions provided for that task, following the steps described below:

- **Manual annotations**: we tabulate all the manually annotated timexes from the Clinical TempEval corpus, listing the timex string, the timex partial sentence (including two previous and following timex tokens), the timex span (begin and end offset boundaries), and the timex class.

[2] http://alt.qcri.org/semeval2015/task6/index.php?id=results

[3] http://www.mayoclinic.org

[4] There was no time normalisation task in Clinical TempEval

[5] http://thyme.healthnlp.org/

Table 1: Time expressions per dataset

Class	Training	Evaluation
DATE	2583	1422
TIME	117	59
DURATION	433	200
SET	218	116
QUANTIFIER	162	109
PREPOSTEXP	305	172
Total	**3818**	**2078**

- **System result**: we created a similar list with the timexes identified by our system.
- **Matches & Similarities**: we compared the manual annotations with our system result to identify a) those timexes that match in terms of span and class, b) those that are similar in terms of span (at least one overlapping character), and c) those that do not have a corresponding entry.
- **Guideline reference**: For each timex that did not match, we identified the guideline, topic and section corresponding to the inconsistency.
- **Agreements & Disagreements**: we set as an "annotation agreement" each timex that a) had the exact same span and class in both manual annotated corpus and our system result, and b) complied with the annotation guidelines – an "annotation disagreement" happened when one of the previous conditions failed.
- **Found expressions**: We checked in the corpus, using a mixture of word lists and simple patterns, for additional timexes that were neither manually annotated as part of the reference corpus, nor identified by our system. We refer to the combined set of (a) manually annotated expressions, (b) expressions automatically identified by our system, and (c) these additional expressions additionally found, as the "found expressions". We will refer to this combined set of found expressions in Section 3.

3 Annotation Analysis

We analysed the annotated datasets provided by Clinical TempEval following the methodology described in Section 2. We considered 4 types of disagreements: a) inconsistency on the annotated span and class; b) non-markable expressions; c) frequent expressions; and d) quantifiers. Each of these is explained below.

3.1 Analysis of Span and Class

When comparing the guidelines against the manually annotated corpus we can observe some inconsistencies concerning the span and the class feature of a timex. We can expect to see a degree of error in any manually annotated corpus; however, we find similar divergences occurring repeatedly. Table 2 summarises all the expression types we analysed, detailing the number of annotation agreements and disagreements, as well as the total number of expressions found in the corpus.

According to TimeML Annotation Guidelines (section 2.2.3), expressions which refer to a time of the day, should be annotated as a class TIME, even if in a very indefinite way (as periods of the day, e.g., "last night" and "the morning of January 31"). From a total of 107 expressions referring to a period of the day, 89 were annotated in the corpus (more than 80%). However, we observed 51 were not annotated as a TIME, but mainly as a DATE class (less than 50% of total number of found expressions).

THYME Guidelines exemplify in section 4.2.6 that temporal granularities denoting a frequency must be annotated as a SET, for example *"monthly", "weekly", "a day", "per day", "a week", "per minute"*. However, 55% of such expressions were incorrectly annotated as DATE or QUANTIFIER (44 disagreements according to the guidelines).

Table 2: Timex class and span inconsistences

Kind of expression	Annotation Agreements	Annotation Disagreements	Found Expressions
Periods of the day	38	51	107
Temporal granularity as frequency	11	44	80
Explicit times	18	26	445
DATE modified to DURATION	35	60	95
DURATION from explicit DATEs	11	8	19
Total	**113**	**189**	**746**

Explicit times of the day should be annotated as a timex of class TIME (section 2.2.3 of TimeML guideline). This should be the case even if such expressions appear isolated in the text (e.g., *"1:33 pm"*) or within a more complex expression together with a date (e.g., *"04-Oct-2010 09:44"*). Less than 10% of the expressions denoting time were manually annotated. Of these, almost 60% represent annotation disagreements as a timex of class DATE instead of TIME.

Section 4.2.3 of THYME Guidelines state that words like "since", "during" and "until" preceding a timex of class DATE should modify the timex class to DURATION. However, in almost 65% of such modified timexes, we found that this rule was not followed, and that the timex was presented as a DATE.

Additionally, in the same section, one can find that two dates can be used to construct a DURATION timex (e.g., "December 2009 through March 2010"). However, because each one represents a single point in time, they should both be separately annotated as DATE rather than DURATION.

3.2 Non-Markable Expressions

The guidelines are clear about a diverse set of non-markable expressions. The TIDES Guidelines have a specific section (3.2) to describe what should not be annotated as a timex, including prepositions and subordinating conjunctions, specific duration and frequency expressions, and proper names. Table 3 lists time expressions found in the provided corpus that are non-markable expressions according to the guidelines.

Table 3: Non-markable time expression

Expression	Annotation Disagreements	Found Expressions
Words "Date/Time"	63	359
Non-quantifiable durations	43	185
Prepositions as triggers	130	1248
Total	**236**	**1792**

There is no reference in the guidelines to annotating the words "Date" and "Time" as a timex when they are not part of a more complex expression, as such isolated words cannot be normalised. In expressions like "Date/Time=Mar 3, 2010", it is expected that "Mar 3, 2010" should be annotated as a DATE, but not the words "Date" and "Time" as time expressions of class DATE and TIME respectively. We found 359 occurrences of such words in 217 different documents, from which 63 of them were incorrectly annotated as DATE and TIME (17.5%).

Non-quantifiable durations are not markable, as they refer to some vague duration (interval) of time, including expressions like "duration", "for a long time", "some time", and "an appropriate amount of time". On the other hand, temporal expressions denoting imprecise amount of time should be annotated as a timex (e.g., *"many days", "few hours"*). We found 185 non-quantifiable duration expressions, from which 43 were incorrectly annotated as a timex with class DURATION (almost 25% of disagreement).

Prepositions which introduce noun phrases are never triggers for time expressions and they can never appear as the syntactic head of an annotated expression. In around 10% of those kind of expressions found in the corpus, time expressions were incorrectly annotated including the head preposition (*"in", "on", "at", "during", "after", "since", "until"*). Some examples include "until July", "on Monday", "in the last year".

3.3 Frequent Expressions

We observed that some expressions tend to appear more often than others in the Clinical TempEval datasets. Most of these are a timex of class SET. A SET is defined (section 4.2.6 of THYME Guidelines) as an expression which comprises a quantifier (optional) and an interval to represent a frequency (mandatory). "Three times weekly", "monthly" and "1/day" are considered as a SET, but not "twice" which is considered as a QUANTIFIER.

We selected a set of the most significant expressions, in terms of the number of occurrences, in order to compare the number of manually annotated expressions against the number of expressions which we found within the text. The expressions were organized in 7 groups:

- Present reference expressions of class DATE *"current(ly)", "recent(ly)", "now", "present(ly)"*;
- Past reference expressions of class DATE *"previous(ly)", "the past"*;
- Explicit years *"2009", "2010"*;
- Precise and imprecise expressions of class DURATION *"24-hour", "2 hours", "six-months", "years"*;
- SETs comprising number of times and frequency *"one-time daily", "two times a day", "twice-a-day", "twice-daily", "three times a day", "four times a day"*;
- SETs comprising only frequencies *"every 6 hours", "every 4 hours", "every evening", "every morning", "every bedtime"*;
- SETs following the pattern "999 /min" – such expressions are part of measurements as in *"Pulse Rate=88 /min"* or *"Resp Rate=16 /min"*.

Table 4 shows how many times each expression was manually annotated and how many times we found it within the corpus (number of found occurrences). Considering all of the selected expressions for this analysis, only 23.3% of such expressions were manually annotated. Considering only SET expressions, the percentage of manually annotated expression is even lower (8.5%).

Table 4: Frequent expressions

Expression	**Manually Annotated**	**Found Expressions**
DATE: present reference	372	836
DATE: past reference	52	117
DATE: explicit years	55	91
DURATION: precise and imprecise	22	114
SET: times and frequency	20	1087
SET: frequency	0	216
SETs: *999 /min*	114	266
Total	**635**	**2727**

3.4 Quantifiers

A special type of timex of class QUANTIFIER was introduced in the THYME Annotation Guidelines. These are used to identify expressions such as "twice", "four times", and "three incidents" which represent the number of occurrences of an EVENT. However, the THYME Guidelines do not make it clear whether or not the words that identify the event itself should be part of the timex span.

In order to understand the way in which QUANTIFIERs and associated EVENTs should be annotated, we examined their occurrence in the Clinical TempEval corpus. We listed all non-numerical words that we found either (a) annotated as part of the QUANTIFIER span or (b) immediately after the QUANTIFIER span. Our reasoning was that these represented the repeated EVENT.

Those 20 most frequent EVENT words found in this way are detailed in Table 5. In the table, we compare the number of manually annotated QUANTIFIERs associated with these EVENTs in the reference corpus, with the number of all QUANTIFIERs that we could find, where they were related to the same kind of EVENT. For example, if the reference corpus included a QUANTIFIER annotation for "twice" in the expression "twice before colonoscopy", then we looked for all occurrences of QUANTIFIER expressions associated with "colonoscopy". Only 11.6% of the QUANTIFIERs that we found were manually annotated in reference the corpus.

Table 5: Words related to quantifiers

Related word	**Manually Annotated**	**Found Expressions**
tablet	5	1135
unit	3	117
cycle	51	65
"drinking" words*	44	53
session	4	44
pack	19	29
colonoscopy	4	27
fraction	14	22
treatment	5	16
bowel	8	16
episode	7	11
stool	7	10
beat	5	7
occasion	5	5
Total	**181**	**1557**

* "Drinking" words include "cup", "glass", "beer", "can", "drink", "bottle", and "beverage".

Note that the THYME Annotation Guidelines explicitly exclude numeric quantifiers of objects as opposed to events, excluding for example "two units of blood". However, we included those words in our analysis as they were used as a referenced EVENT to annotate QUANTIFIERs in the corpus, usually followed by an expression which identifies frequency (e.g., "1 TABLET by mouth every evening").

4 Recommendations

The analysis given in the previous section has led us to think about the way in which manual temporal expression annotation efforts are conducted. We venture to make a number of recommendations, hoping that these will at least be considered in future manual annotation efforts. We discuss our recommendations below.

Annotation guidelines should clearly state the full set of rules defining what should or should not be annotated, and how. For THYME, the annotators had to piece together several guidelines to figure out what to annotate. This is a potential source of error. Training in the use of multiple sets of guidelines could be considered as an alternative.

Examples are a valuable aid to annotators. Although examples are given in the THYME guidelines, the number could be expanded. In the CLEF Project for example (Roberts et al., 2009), each time an annotator raised a question, and each time persistent differences between annotators were found, new examples were added to the guidelines to re-enforce the point raised.

In creating the THYME gold standard, multiple annotators and an adjudication process were used. A potential source of error with this approach is that where all annotators have a low recall and adjudication focuses only on resolving disputes, the resulting recall can be no greater than the union of the two. This casts doubt on the veracity of inter-annotator agreement (Fleiss et al., 1981) as an indicator of the accuracy of annotation of a corpus.

This last point raises the potential merit of using a high recall rule-based system to prepare a corpus, creating annotations for review by human annotators. Some constructs and guidelines can be represented by simple, unambiguous rules, and where this is the case, the rules will most likely outperform the human annotator in terms of recall. We feel that in such high recall cases, the disadvantage of the approach, that there tends to be a poor correction of missing spans, would be outweighed by the increased number of annotations found.

5 Related Work

TimeML (Pustejovsky et al., 2003) is an expressive language for temporal information annotation, designed to connect the processes of temporal analysis of a text with a representation and formal meaning of time. It is a specification language for event and temporal expressions in natural language text able to capture distinct phenomena in temporal markup, to anchor events to temporally denoting expressions, and to order relative event expressions.

The development of temporal annotation standards and corpora has a long history. Of note is the TimeBank corpus (Pustejovsky et al., 2003), which contains 183 news articles annotated with temporal information, events, times and temporal links between events and times. This corpus was developed in multiple iterations, and prior analyses of the annotated data and the annotation standard aided the evolution of both. For example, Boguraev and Ando (2007) presented an extensive analysis of the TimeBank reference corpus in terms of development support of TimeML-compliant analytics, which helped advance the state of the art in temporal annotation. Indeed, iterative application of an annotation standard and examination of the resulting annotated data are critical steps in the MATTER development cycle, used for construction annotation standards (Pustejovsky, 2006; Pustejovsky and Stubbs, 2012).

Within the previous SemEval evaluation (UzZaman et al., 2013), the TempEval-3 Task "A" examined temporal information extraction and normalisation using the complete set of TimeML temporal relations. Most of the participant systems achieved over 0.70 in Precision and Recall, and best approaches achieved 0.82 and 0.77 for strict F1-score on identifying span and value of timexes. TempEval and TempEval-2 (Verhagen et al., 2009, 2010) also included temporal annotation tasks, of which both were followed by informative analyses of the corpora and participant results (Lee and Katz, 2009; Derczynski, 2013), which led to a better understanding of the task as framed in these exercises.

Other researchers have annotated temporal information in clinical text. For example, the CLEF Project (Roberts et al., 2009) semantically annotated a corpus to assist in the extraction of clinical information from text. It used two different schemas to annotate a) clinical entities and relations between them, and b) time expressions and their temporal relations with the clinical entities in the text. The i2b2 Natural Language Processing Challenge for Clinical Records focused on the temporal relations in clinical narratives, attracting 18 participating teams to analyse discharge summaries, annotating time expressions, events, and relations between them (Sun et al., 2013).

6 Conclusions and Future Work

Adapting annotation of temporal semantics to clinical notes is a significant and challenging task. This paper detailed the results of a principled analysis of expert manual annotations of temporal expressions in the THYME schema over a corpus of clinical notes. Discrepancies between annotations and the guidelines were found in multiple categories. The spans or temporal expressions were not always correct. Ambiguity remained regarding the correct timex class, as happened also in TimeML. Wording in the guidelines was sometimes misinterpreted leading to non-markable timexes being annotated. Finally, as in

TimeML, confusion appeared around the annotation of complex SET-type timexes and their quantifiers. This data-driven analysis and its findings should help guide future temporal annotation efforts in the clinical domain.

Acknowledgments

We would like to thank the Mayo Clinic for permission to use the THYME corpus, and CAPES, which is partially financing this work. This work received funding from the European Union's Seventh Framework Programme (grant No. 611233, PHEME).

References

Bartak, R., R. Morris, and K. Venable (2013). *An Introduction to Constraint-Based Temporal Reasoning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.

Bethard, S., L. Derczynski, J. Pustejovsky, and M. Verhagen (2015). SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics.

Boguraev, B. and R. K. Ando (2007). Effective use of TimeBank for TimeML analysis. In *Annotating, extracting and reasoning about time and events*, pp. 41–58. Springer.

Caselli, T. (2009). *Time, Events and Temporal Relations: an Empirical Model for Temporal Processing of Italian Texts*. Ph. D. thesis, Università di Pisa, Pisa, Italy.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters (2011). *Text Processing with GATE (Version 6)*. GATE.

Derczynski, L. (2013). *Determining the Types of Temporal Relations in Discourse*. Ph. D. thesis, University of Sheffield.

Ferro, L., L. Gerber, I. Mani, B. Sundheim, and G. Wilson (2005). TIDES 2005 standard for the annotation of temporal expressions. Technical report, The MITRE Corporation.

Fleiss, J. L., B. Levin, and M. C. Paik (1981). The Measurement of Interrater Agreement. pp. 212–236.

Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60. ACL.

Lee, C. M. and G. Katz (2009). Error analysis of the tempeval temporal relation identification task. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pp. 138–145. ACL.

Pustejovsky, J. (2006). Unifying linguistic annotations: A TimeML case study. In *Proceedings of Text, Speech, and Dialogue Conference*.

Pustejovsky, J., J. Castano, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz (2003). TimeML: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5)*.

Pustejovsky, J., P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo (2003, March). The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, Lancaster, pp. 647–656.

Pustejovsky, J., K. Lee, H. Bunt, and L. Romary (2010). ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. ELRA.

Pustejovsky, J. and J. Moszkowicz (2012). The role of model testing in standards development: The case of ISO-Space. In *LREC*, pp. 3060–3063.

Pustejovsky, J. and A. Stubbs (2012). *Natural language annotation for machine learning*. O'Reilly Media, Inc.

Roberts, A., R. J. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics 42*(5), 950–966.

Sauri, R., J. Littman, R. Gaizauskas, A. Setzer, and J. Pustejovsky (2006). TimeML Annotation Guidelines, Version 1.2.1.

Styler, W., S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, and J. Pustejovsky (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics 2*, 143–154.

Sun, W., A. Rumshisky, and O. Uzuner (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc 20*(5), 806–813.

UzZaman, N., H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 1–9. ACL.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, J. Moszkowicz, and J. Pustejovsky (2009). The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation 43*(2), 161–179.

Verhagen, M., R. Sauri, T. Caselli, and J. Pustejovsky (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 57–62. ACL.