A Corpus of Clinical Narratives Annotated with Temporal Information

Lucian Galescu Nate Blaylock
Florida Institute for Human and Machine Cognition (IHMC)
Pensacola, Florida, USA
{lgalescu, blaylock}@ihmc.us

ABSTRACT

Clinical reports often include descriptions of events in the patient's medical history, as well as explicit or implicit temporal information about these events. We are working towards applying deep Natural Language Processing tools towards understanding such narratives. This requires both the extraction and classification of the relevant events, and the placing of those events in time, or at least in relation to one another. Although several corpora of news data exist that have been annotated using the TimeML schema, similar corpora of clinical reports are not readily available.

In this paper we report on the design of a small corpus and the annotation schema we developed, based on data from the fourth i2b2/VA challenge. These data include, among others, annotations for medical problems, tests, and treatments in clinical reports from several healthcare institutions. We have selected a subset of clinical reports and added annotations similar to those used in the TempEval tasks for the annotation of events, time expressions and temporal relations for the news domain. The annotations have been made freely available to the research community.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*

General Terms

Algorithms

Keywords

Clinical narrative, corpora, temporal relations, TimeML

1. INTRODUCTION

Clinical records are filled with information about patients' medical histories, treatments, diagnoses, and so forth. This information has the potential to influence many areas of medicine, including comparative effectiveness research and clinical decision making. A large amount of this information, however, is currently locked in unstructured text reports. We are interested in using deep natural language processing (NLP) to extract this information into structured knowledge bases.

Although much work has been done in the field of clinical information extraction (see [4] for an overview), relatively little has been done on extracting temporal information from clinical texts. For many fields, including evidence-based medicine, knowledge of not only the patient's medical history, but also the *timeline* of events is critical, e.g., to see the progression of the disease and the efficacy of treatments.

We are currently building an NLP system that will extract timelines of medical events from clinical notes [3].

In order to evaluate this system, we have created a corpus of annotated history of present illness (HPI) sections of a set of clinical notes. In this paper, we describe the data set annotated along with the annotation scheme we developed for clinical texts, based on the time markup language—TimeML [6]. We then discuss the annotations, describe related work, and conclude by mentioning our future planned work.

2. DATA SET

The 2010 i2b2/VA challenge [9] was organized to provide a shared data set for the extraction and classification of clinical problems, treatments, and tests, as well as assertion information on these and event-event relations.

The i2b2-2010 training dataset consists of 349 normalized, de-identified discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Center, as well as discharge summaries and progress notes from University of Pittsburgh Medical Center (UPMC). Unfortunately, however, we found that not all reports from the dataset are usable for temporal relation extraction, due to obfuscation from de-identification in reports from several of the sources.

Although the exact de-identification protocols for the i2b2 data are not published, it appears that several different approaches were taken. For example, the Beth Israel data has all dates changed with other dates in the same format as the original; it is not clear if relative order and durations have been preserved. The UPMC data has dates replaced with expressions of the form ***DATE[Nov 16 2007]*; the actual dates included in these expressions seem consistent and thus would allow the text to be reconstituted. Other transformations of this kind involve age (e.g., ***AGE[in the 50s]*), places and names. These expressions render the text difficult to parse using natural language parsers; reconsti-

IHI'12 January 28–30, 2012, Miami, Florida, USA.

tuting the text is an option, but for the current exercise we didn't pursue this line.

Therefore, we chose as the basis for constructing our corpus the data subset provided by Partners Healthcare because the temporal information in those files was not altered structurally by the de-identification procedures used. Source hospitals are known for each report in the i2b2 training data, but not for those in the test data. Thus, our initial dataset consisted of the 97 discharge summaries in the i2b2 training data from this source.

From each of these reports, we extracted text from the sections titled "History of Present Illness" (HPI), as these consist of text in narrative form that is very rich in events and temporal expressions. This resulted in a data set of 44 sections with 410 sentences and 7704 tokens.

3. ANNOTATION

The TempEval competition series [10, 11] focuses on the extraction of events, temporal expressions and the relations between them for general natural language text, based on the Time Markup Language (TimeML) [6]. Our annotation scheme is based on a subset the TempEval2 challenge. In particular, we annotate clinical events (based on the previous i2b2 annotation), time expressions, and relations between the two classes within sentences.

In this section, we first describe the event annotations, time expression annotations, and temporal relation annotations. We then discuss details of the resulting corpus of temporal narratives.

3.1 Clinical Event Annotation

The i2b2 data contains concept annotations for medical problems, tests, and treatments, which are annotated as a string of words (bracketing), along with one of the three given categories. We refer the interested reader to [9] for a more detailed description of this annotation. We built on the i2b2 annotation scheme by tagging all the i2b2 concepts as EVENTs in TimeML. No additional event annotation was made for this corpus.

We note that the i2b2 annotation was over medical *concepts*, whereas TimeML annotations (and temporal relations in general) are over *events*. Although this causes a conceptual mismatch, we decided on this strategy for several reasons. First, i2b2 data is among a very rarefied set of clinical text data which is publicly available. Lack of available data, especially shared data, is a big hindrance for research in this area. Second, this allows a task which essentially builds on the i2b2-2010 tasks, making this more compatible for researchers who have already worked on i2b2 data (such as ourselves). Lastly, this allowed us to more quickly annotate the data for our planned system evaluation.

That said, this approach led to several differences between our treatment of events, and those used in TempEval-2. First, under the TempEval-2 guidelines we would have to consider all events explicitly mentioned in the text, whether medically relevant or not. Thus, there is a whole set of non-medical events that are not annotated in our corpus.

Second, in TempEval-2 events can be represented via any grammatical construction, first and foremost verb phrases, but also noun phrases, and even certain prepositional phrases and adjectives; in contrast, i2b2 concepts can only be noun phrases. Some of these are event nominalizations (such as *intubation*) which would be tagged as events in TempEval-2. Others (such as *blood pressure*) are not events in and of themselves, but rather entities which participate in some event. During annotation of temporal relations (described below), the annotators treated these non-event concepts as if they referred to the event they were most closely associated with. Thus, in the case of medical problems such as *diabetes mellitus*, the event considered was the patient's having the condition. Similarly, for a test reference like *<EVENT>Her blood pressure</EVENT> was measured at 240/120*, the annotators considered the event of measuring the blood pressure, even though only *Her blood pressure* is annotated.

3.2 Time Expression Annotation

In TimeML, temporal expressions are annotated using the TIMEX3 tag. We followed the TempEval-2 guidelines for TIMEX3 annotation[1], that the extent of a TIMEX3 should be as small as possible. Thus, the full extent of a TIMEX3 cannot start with a preposition (e.g., *in*, *of*, *on*, etc.), or a subordinating conjunction (e.g., *after*, *before*, *since*, *until*, *when*, *while*, *as soon as*, etc.). Postmodifiers of the time expression (e.g., prepositional phrases and dependent clauses) are not considered part of the time expression. We departed from the general TimeML annotation schema in that we avoided annotating adjectival expressions such as *recent*, *current*, etc., and their adverbial counterparts. Although these are legitimate temporal expressions, we chose not to annotate them at this time because they often appear as pre-modifiers inside events (e.g., *recent upper respiratory infections*) and we wanted to avoid overlaps between event tags and temporal tags.

Four kinds of temporal expressions are identified, and marked using the **type** attribute:

- **DATE**: representing a calendar date (e.g., *03/20/99*, *yesterday*, *last fall*, *about six months ago*, etc.);
- **TIME**: representing a time of day (e.g., *5:30 p.m.*, *this morning*, etc.);
- **DURATION**: representing a span of time (e.g., *three weeks*, *the past couple of months*, *several days*, etc.); and
- **SET**: representing a set of times (e.g., *monthly*, *approximately every three months*, *per day*, etc.).

Although the TimeML specification provides a number of attributes for describing features of the various types of time expressions, we followed the simplified guidelines of TempEval-2, and annotated just the **type** and the **value** attributes for these expressions.

In most cases, the **value** attribute is set by computing the exact date or time, duration, from the text. For certain expressions such as *now* and *today*, by convention we use the token PRESENT_REF. There were a few cases where we didn't assign a value; most of them were ambiguous anaphoric temporal expression, such as *at that time*, for which an unique referent could not be identified; to be sure, when a unique referent was identified and we could compute a value, we assigned it to the anaphoric expression as well. We also encountered an example of a holiday name, *Martin Luther King Day*; in such cases it is customary to not assign

[1]See `http://timeml.org/tempeval2` for the TempEval-2 annotation guidelines.

```
<EVENT class="OCCURRENCE" eid="e4">This operation</EVENT> was performed
for <TIMEX3 tid="t2" type="DURATION" value="P2M">two months</TIMEX3> of
<EVENT class="OCCURRENCE" eid="e5">increased rest pain</EVENT> .

<TLINK eventID="e4" relType="OVERLAP-OR-AFTER" relatedToTime="t2"/>
<TLINK eventID="e5" relType="DURING" relatedToTime="t2"/>
```

Figure 1: Example of annotated sentence

	Sections	Sentences	Tokens	Events	TIMEX3s	TLINKs
dev	5	55	1156	111	25	54
tst	39	355	6548	698	185	367
Total	44	410	7704	809	210	421

Table 1: Summary statistics for the corpus

values if they cannot be computed from the context of the document, without referring to cultural knowledge.

In addition to the mark-up for temporal expressions found in text, we added one non-consuming TIMEX tag at the top of each document as an annotation of the admission date, which, for clinical notes is similar in function to the Document Creation Time used in typical TimeML annotation schemes for news data. We expect this tag, which was added automatically, to be of crucial importance to any temporal inference that may be carried out, since references to the admission date are quite frequent in the texts comprising this corpus.

3.3 Temporal Relation Annotation

Temporal relations between events, times, as well as between events and times are represented in TimeML via the TLINK tag. For this version of the corpus we decided to annotate only one set of temporal relations, those between events and time expressions in the same sentence. This corresponds to the first task (Task A) in TempEval-1; note that a similar task in TempEval-2 (Task C) further restricts this task by requiring that the event either dominates syntactically, or appears in the same clause as the time expression, thus ensuring a tighter connection between the two.

The set of relations we used is the same used in the TempEval tasks: BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE. A TLINK between an event and a time expression is marked as BEFORE if the event happened before the time or time interval denoted by the time expression. If there is non-empty overlap between an event and a time expression, the TLINK between them is marked OVERLAP. If an event happens immediately before a time, or if it marks the beginning (or is begun by) a time, the relation between the event and that time is marked BEFORE-OR-OVERLAP. The AFTER and OVERLAP-OR-AFTER relations are the inverse of the BEFORE and BEFORE-OR-OVERLAP relations, respectively.

One question that came up when annotating temporal relations was that of what kind of evidence/knowledge was needed for assessing the temporal extent of a medical condition. For example, consider a case where the clinical note states that the patient was admitted to the hospital today for lymphoma. We know she has lymphoma today, but did she have lymphoma yesterday? What about 2 weeks ago? The answer is that we can assume, but don't know for sure. To handle these questions, we adopted a minimalist approach, by requiring the annotators to only code according to the evidence presented in the text.

3.4 An Example

In Figure 1 we show an example of a fully annotated sentence from our corpus (slightly simplified to enhance readability). As described above, *This operation* and *increased rest pain* — which in the i2b2 corpus are tagged as a **treatment** and a **problem**, respectively — are both marked as EVENT occurrences. There is one temporal expression, *two months*, which is marked as a DURATION. The event of having increased rest pain occurred during the two-month period, and thus the relation between the two is marked with the type DURING. The operation occurred immediately after (or, it ended) the two-month period, therefore it is marked as OVERLAP-OR-AFTER.

3.5 Corpus Details

The corpus included 44 sections with 410 sentences and 7704 tokens. There were 809 events, including 522 problems, 160 treatments and 127 tests.

The annotation was a multi-step process carried out by the two authors. In the first step, we selected 5 random sections to be used as development data (marked as **dev** in Table 1). Both annotators marked up independently the temporal expressions and temporal relations according to TimeML guidelines. Our final annotation schema, as described above, was by and large the result of the reconciliation between the two annotators' mark-up.

We computed the *inter-annotator agreement* (IAA) using the F-score metric (the harmonic mean between precision and recall, using one annotator's data as key and the other's as response). For the five development sections we obtained an IAA of 75.86% on TIMEX3 tags and 73.04% on TLINK tags (using exact match).

During the second step, the annotators marked up only temporal expressions for the 39 remaining sections (marked as **tst** in Table 1). Then the two annotations were reconciled an a final version of the TIMEX3 annotation was thus obtained. We computed an IAA of 83.95% for this step.

During the third and final step, TLINKs for the 39 sections were annotated based on the reconciled version of the TIMEX3 annotation. All the intra-sentential event-time

	Count	Percentage
DATE	105	50.0%
TIME	16	7.6%
DURATION	76	36.2%
SET	13	6.2%
Total	210	100.0%

Table 2: Summary statistics for temporal expressions

	Count	Percentage
OVERLAP	268	63.7%
BEFORE	25	5.9%
AFTER	23	5.5%
BEFORE-OR-OVERLAP	11	2.6%
OVERLAP-OR-AFTER	34	8.1%
VAGUE	60	14.3%
Total	421	100.0%

Table 3: Summary statistics for temporal relations

TLINK tags were created automatically, and annotators only marked up the type of the TLINK. On this task we obtained an IAA of 74.39%.

Table 1 has summary statistics for the final, reconciled version of the corpus. Table 2 contains detail about the types of temporal expressions encountered. Of note, DATEs constitute half of all temporal expressions, with DURATIONs covering most of the remainder.

Finally, in Table 3 we show details about the types of intra-sentential TLINKs found in the corpus. Note that close to two thirds of them have the type OVERLAP. Also, OVERLAP-OR-AFTER relations far outnumber the ones marked BEFORE-OR-OVERLAP. We think these statistics reflect the content and style of the HPI sections, since most of what they deal with is when health problems started started and what happened afterwards (effects, interventions, etc.).

The resulting annotated corpus (the HPI TimeML Corpus) has been made available for public release and can be freely downloaded.[2] As explained above, the annotation is made *on top of* the 2010 i2b2/VA data. Due to licensing and IRB restrictions, we are unable to distribute the 2010 i2b2/VA data ourselves (i2b2's release of the underlying data to the research community is planned for November 2011 [9]). Instead, we have released our TimeML annotations in the form of a stand-off annotation, together with scripts for reconstructing the inline TimeML mark-up from the text files in the i2b2 distribution (which interested parties can obtain separately from i2b2).

4. DISCUSSION

Although this is a relatively small corpus, several questions came up during annotation, which we consider interesting problems for future work. First, how much world knowledge should be used in annotation? Consider the case of a note mentioning that a patient has a history of Parkinson's disease and myocardial infarction. For Parkinson's, with world knowledge we know that this is a long-term event that is incurable. If the patient has a history of it, he still has it. With a heart attack, we know this is a short-term event. If a patient has a history of it, the same heart attack event is not occurring now. Both of these events would have a different temporal relation with the time expression *today*. In this annotation, we decided that "common" world knowledge about medical concepts could be used to infer temporal relations, but no specialized medical knowledge could be used (except for what could be inferred from the text itself).

There were also two cases in the data where it appeared the wrong dates were used. One stated that a past procedure had been performed on a date that was far after the discharge date! This raises the question of how such mistakes should be annotated. In both cases we decided to annotate the TIMEX3 value to match the actual text, not what we may have inferred as being the correct date. TLINKs between events and these dates were all marked as VAGUE.

In a few cases both annotators found that a temporal link between an event and a temporal expression in the same sentence did not make sense; in particular, such was the case with frequency expressions (e.g., *per day*) when linked to any events other than the ones to which the temporal frequency applies. For consistency, we deferred a decision on what to do with these cases, and kept the temporal links in the mark-up, with the type VAGUE. We expect that a complete and consistent treatment of such cases will require further refinement of the TimeML schema, rather than simple adjustments to our annotation schema for this corpus; for an example of such a refinement towards better handling of recurrent events and event quantification we refer to [1].

5. RELATED WORK

There has been recent interest in the clinical NLP community in temporal relation extraction ([12, 8], inter alia). We are, however, aware of only a handful of annotation schemas and annotated datasets for this task. Additionally, none of the datasets of which we are aware are publicly available, which makes the direct comparison of extraction approaches practically impossible.

Harkema et al. [2] describe an annotation schema for temporal information based on TimeML. The annotation, however, only concentrates on investigation events, such as X-rays and CT scans. The annotation was done on 446 clinical documents.

Mowery et al. [5] annotated a set of 24 clinical reports with temporal information. Clinical conditions in the corpus were annotated directly with temporal information (as opposed to the TimeML approach where relations between events and explicit time expressions are annotated). The conditions were annotated with an explicit time and date, if surmisable, or as relative to other clinical events.

Probably the closest effort to our own work is [7], which defined a clinical annotation schema based on TimeML and used it to annotate a 5000K token corpus. This approach more closely mirrored TimeML in that all events (medical or otherwise) were natively tagged (not just the medically relevant noun phrases as with our approach). Additionally, each event was annotated with a "tense" of *past*, *present*, or *future*, which referred not to the tense of the verb, but rather the temporal occurrence of the event with respect to the time of the patient encounter with the clinician. Additional information such as modality, aspect, and event-event temporal relations were also annotated.

[2]See http://cs.rochester.edu/research/speech/hpi-timeml

6. CONCLUSION AND FUTURE WORK

In this paper we have described an annotated dataset of temporal relations from clinical texts. We built on the dataset from the 2010 i2b2/VA challenge to include temporal expressions and intra-sentential relations between time expressions and medical events. The resulting annotation has been made available to the research community. We hope that this will provide a common dataset to evaluate automated approaches to timeline extraction from clinical texts. In the near future, we plan to use this corpus to evaluate our own clinical timeline extraction system.

Acknowledgments

This work was supported by Award RC2CA1488332 from the National Cancer Center and the H. Lee Moffitt Cancer Center and Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY.

7. REFERENCES

[1] H. Bunt and J. Pustejovsky. Annotation of temporal and event quantification. In *Proceedings of the Fifth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5)*, pages 15–22, Jan. 2010.

[2] H. Harkema, A. Setzer, R. Gaizauskas, M. Hepple, R. Power, and J. Rogers. Mining and Modelling Temporal Clinical Data. In S. J. Cox, editor, *Proceedings of the 4th UK e-Science All Hands Meeting*, Nottingham, UK, 2005.

[3] H. Jung, J. Allen, N. Blaylock, W. de Beaumont, L. Galescu, and M. Swift. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of the Tenth Workshop on Biomedical Natural Language Processing (BioNLP)*, Portland, Oregon, June 23–24 2011.

[4] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook of Medical Informatics*, pages 128–144, 2008.

[5] D. L. Mowery, H. Harkema, and W. W. Chapman. Temporal annotation of clinical text. In *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 106–107, Columbus, Ohio, June 2008.

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

[7] G. Savova, S. Bethard, W. Styler, J. Martin, M. Palmer, J. Masanz, and W. Ward. Towards temporal relation discovery from the clinical narrative. In *Proceedings of the 2009 AMIA Annual Symposium*, pages 568–572, 2009.

[8] C. Tao, W.-Q. Wei, H. R. Solbrig, G. Savova, and C. G. Chute. CNTRO: A semantic web ontology for temporal relation inferencing in clinical narratives. In *Proceedings of the 2010 AMIA Annual Symposium*, Washington, DC, 2010.

[9] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Inforatics Association*, 18(5), September 2011.

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

[11] M. Verhagen, R. Saur'ı, T. Caselli, and J. 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.

[12] L. Zhou, G. B. Melton, S. Parsons, and G. Hripcsak. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, 39:424–439, 2006.