SemEval-2015 Task 6: Clinical TempEval

Steven Bethard
University of Alabama at Birmingham
Birmingham, AL 35294, USA
bethard@cis.uab.edu

Leon Derczynski
University of Sheffield
Sheffield, S1 4DP, UK
leon@dcs.shef.ac.uk

Guergana Savova
Harvard Medical School
Boston, MA 02115, USA
Guergana.Savova@
childrens.harvard.edu

James Pustejovsky, Marc Verhagen
Brandeis University
Waltham, MA 02453, USA
jamesp@cs.brandeis.edu
marc@cs.brandeis.edu

Abstract

Clinical TempEval 2015 brought the temporal information extraction tasks of past TempEval campaigns to the clinical domain. Nine sub-tasks were included, covering problems in time expression identification, event expression identification and temporal relation identification. Participant systems were trained and evaluated on a corpus of clinical notes and pathology reports from the Mayo Clinic, annotated with an extension of TimeML for the clinical domain. Three teams submitted a total of 13 system runs, with the best systems achieving near-human performance on identifying events and times, but with a large performance gap still remaining for temporal relations.

1 Introduction

The TempEval shared tasks have, since 2007, provided a focus for research on temporal information extraction (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013). Participant systems compete to identify critical components of the timeline of a text, including time expressions, event expressions and temporal relations. However, the TempEval campaigns to date have focused primarily on in-document timelines derived from news articles.

Clinical TempEval brings these temporal information extraction tasks to the clinical domain, using clinical notes and pathology reports from the Mayo Clinic. This follows recent interest in temporal information extraction for the clinical domain, e.g., the i2b2 2012 shared task (Sun et al., 2013), and broadens our understanding of the language of time beyond newswire expressions and structure.

Clinical TempEval focuses on discrete, well-defined tasks which allow rapid, reliable and repeatable evaluation. Participating systems are expected to take as input raw text such as:

> April 23, 2014: The patient did not have any postoperative bleeding so we will resume chemotherapy with a larger bolus on Friday even if there is slight nausea.

And output annotations over the text that capture the following kinds of information:

- *April 23, 2014*: TIMEX3
  - TYPE=DATE
- *postoperative*: TIMEX3
  - TYPE=PREPOSTEXP
  - CONTAINS
- *bleeding*: EVENT
  - POLARITY=NEG
  - BEFORE document creation time
- *resume*: EVENT
  - TYPE=ASPECTUAL
  - AFTER document creation time
- *chemotherapy*: EVENT
  - AFTER document creation time
- *bolus*: EVENT
  - AFTER document creation time
- *Friday*: TIMEX3
  - TYPE=DATE
  - CONTAINS
- *nausea*: EVENT
  - DEGREE=LITTLE
  - MODALITY=HYPOTHETICAL
  - AFTER document creation time

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 806–814,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

That is, the systems should identify the time expressions, event expressions, attributes of those expressions, and temporal relations between them.

2 Data

The Clinical TempEval corpus was based on a set of 600 clinical notes and pathology reports from cancer patients at the Mayo Clinic. These notes were manually de-identified by the Mayo Clinic to replace names, locations, etc. with generic placeholders, but time expressions were not altered. The notes were then manually annotated by the THYME project (`thyme.healthnlp.org`) using an extension of ISO-TimeML for the annotation of times, events and temporal relations in clinical notes (Styler et al., 2014b). This extension includes additions such as new time expression types (e.g., PREPOSTEXP for expressions like *postoperative*), new EVENT attributes (e.g., DEGREE=LITTLE for expressions like *slight nausea*), and an increased focus on temporal relations of type CONTAINS (a.k.a. INCLUDES).

The annotation procedure was as follows:

1. Annotators identified time and event expressions, along with their attributes
2. Adjudicators revised and finalized the time and event expressions and their attributes
3. Annotators identified temporal relations between pairs of events and events and times
4. Adjudicators revised and finalized the temporal relations

More details on the corpus annotation process are documented in a separate article (Styler et al., 2014a).

Because the data contained incompletely de-identified clinical data (the time expressions were retained), participants were required to sign a data use agreement with the Mayo Clinic to obtain the raw text of the clinical notes and pathology reports.[1] The event, time and temporal relation annotations were distributed separately from the text, in an open source repository[2] using the Anafora standoff format (Chen and Styler, 2013).

[1]The details of this process are described at `http://thyme.healthnlp.org/`

[2]`https://github.com/stylerw/thymedata`

	Train	Dev
Documents	293	147
EVENTS	38890	20974
TIMEX3s	3833	2078
TLINKS with TYPE=CONTAINS	11176	6173

Table 1: Number of documents, event expressions, time expressions and narrative container relations in the training and development portions of the THYME data. (Dev is the Clinical TempEval 2015 test set.)

The corpus was split into three portions: Train (50%), Dev (25%) and Test (25%). For Clinical TempEval 2015, the Train portion was used for training and the Dev portion was used for testing. The Test portion was not distributed, and was reserved as a test set for a future iteration of the shared task. Table 1 shows the number of documents, event expressions (EVENT annotations), time expressions (TIMEX3 annotations) and narrative container relations (TLINK annotations with TYPE=CONTAINS attributes) in the Train and Dev portions of the corpus.

3 Tasks

A total of nine tasks were included, grouped into three categories:

- Identifying time expressions (TIMEX3 annotations in the THYME corpus) consisting of the following components[3]:
  - The spans (character offsets) of the expression in the text
  - Class: DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET
- Identifying event expressions (EVENT annotations in the THYME corpus) consisting of the following components:
  - The spans (character offsets) of the expression in the text
  - Contextual Modality: ACTUAL, HYPOTHETICAL, HEDGED or GENERIC
  - Degree: MOST, LITTLE or N/A
  - Polarity: POS or NEG
  - Type: ASPECTUAL, EVIDENTIAL or N/A

[3]Normalized time values (e.g. 2015-02-05) were originally planned, but annotation was not completed in time.

- Identifying temporal relations between events and times, focusing on the following types:
  - Relations between events and the document creation time (BEFORE, OVERLAP, BEFORE-OVERLAP or AFTER), represented by DOCTIMEREL annotations in the THYME corpus
  - Narrative container relations (Pustejovsky and Stubbs, 2011) between events and/or times, represented by TLINK annotations with TYPE=CONTAINS in the THYME corpus

The evaluation was run in two phases:

1. Systems were given access only to the raw text, and were asked to identify time expressions, event expressions and temporal relations
2. Systems were given access to the raw text and the manual event and time annotations, and were asked to identify only temporal relations

4 Evaluation Metrics

All of the tasks were evaluated using the standard metrics of precision (P), recall (R) and F_1:

$$P = \frac{|S \cap H|}{|S|}$$

$$R = \frac{|S \cap H|}{|H|}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

where S is the set of items predicted by the system and H is the set of items manually annotated by the humans. Applying these metrics to the tasks only requires a definition of what is considered an "item" for each task.

- For evaluating the spans of event expressions or time expressions, items were tuples of (begin, end) character offsets. Thus, systems only received credit for identifying events and times with exactly the same character offsets as the manually annotated ones.
- For evaluating the attributes of event expressions or time expressions – Class, Contextual Modality, Degree, Polarity and Type – items were tuples of (begin, end, value) where begin and end are character offsets and value is the value that was given to the relevant attribute. Thus, systems only received credit for an event (or time) attribute if they both found an event (or time) with the correct character offsets and then assigned the correct value for that attribute.
- For relations between events and the document creation time, items were tuples of (begin, end, value), just as if it were an event attribute. Thus, systems only received credit if they found a correct event and assigned the correct relation (BEFORE, OVERLAP, BEFORE-OVERLAP or AFTER) between that event and the document creation time. Note that in the second phase of the evaluation, when manual event annotations were given as input, precision, recall and F_1 are all equivalent to standard accuracy.
- For narrative container relations, items were tuples of ((begin$_1$, end$_1$), (begin$_2$, end$_2$)), where the begins and ends corresponded to the character offsets of the events or times participating in the relation. Thus, systems only received credit for a narrative container relation if they found both events/times and correctly assigned a CONTAINS relation between them.

For attributes, an additional metric measures how accurately a system predicts the attribute values on just those events or times that the system predicted. The goal here is to allow a comparison across systems for assigning attribute values, even when different systems produce very different numbers of events and times. This is calculated by dividing the F_1 on the attribute by the F_1 on identifying the spans:

$$A = \frac{\text{attribute } F_1}{\text{span } F_1}$$

For the narrative container relations, additional metrics were included that took into account *temporal closure*, where additional relations can be deterministically inferred from other relations (e.g., A CON-

TAINS B and B CONTAINS C, so A CONTAINS C):

$$P_{\text{closure}} = \frac{|S \cap \text{closure}(H)|}{|S|}$$

$$R_{\text{closure}} = \frac{|\text{closure}(S) \cap H|}{|H|}$$

$$F_{\text{closure}} = \frac{2 \cdot P_{\text{closure}} \cdot R_{\text{closure}}}{P_{\text{closure}} + R_{\text{closure}}}$$

These measures take the approach of prior work (UzZaman and Allen, 2011) and TempEval 2013 (UzZaman et al., 2013), following the intuition that precision should measure the fraction of system-predicted relations that can be verified from the human annotations (either the original human annotations or annotations inferred from those through closure), and that recall should measure the fraction of human-annotated relations that can be verified from the system output (either the original system predictions or predictions inferred from those through closure).

5 Baseline Systems

Two rule-based systems were used as baselines to compare the participating systems against.

memorize For all tasks but the narrative container task, a memorization-based baseline was used.

To train the model, all phrases annotated as either events or times in the training data were collected. All exact character matches for these phrases in the training data were then examined, and only phrases that were annotated as events or times greater than 50% of the time were retained. For each phrase, the most frequently annotated type (event or time) and attribute values for instances of that phrase were determined.

To predict with the model, the raw text of the test data was searched for all exact character matches of any of the memorized phrases, preferring longer phrases when multiple matches overlapped. Wherever a phrase match was found, an event or time with the memorized (most frequent) attribute values was predicted.

closest For the narrative container task, a proximity-based baseline was used. Each time expression was predicted to be a narrative container, containing only the closest event expression to it in the text.

6 Participating Systems

Three research teams submitted a total of 13 runs:

BluLab The team from Stockholm University and University of Utah participated in all tasks, using supervised classifiers with features generated by the Apache clinical Text Analysis and Knowledge Extraction System (cTAKES)[4]. Their runs differed in whether and how many rules were used to constrain the search for narrative container relations.

KPSCMI The team from Kaiser Permanente Southern California participated in the time expression tasks. Their runs compared an extended version of the rule-based HeidelTime[5] system (run 1) with systems based on supervised classifiers (run 2-3).

UFPRSheffield The team from Universidade Federal do Paraná and University of Sheffield participated in the time expression tasks. Their runs compared in-house rule-based systems (the Hynx runs) to systems based on supervised classifiers (the SVM runs).

7 Human Agreement

We also give two types of human agreement on the task, measured with the same evaluation metrics as the systems:

ann-ann Inter-annotator agreement between the two independent human annotators who annotated each document. This is the most commonly reported type of agreement, and often considered to be an upper bound on system performance.

adj-ann Inter-annotator agreement between the adjudicator and the two independent annotators. This is usually a better bound on system performance in adjudicated corpora, since the models are trained on the adjudicated data, not on the individual annotator data.

[4] https://ctakes.apache.org

[5] https://code.google.com/p/heideltime/

	span			span + class			
Team	P	R	F1	P	R	F1	A
Baseline: memorize	0.743	0.372	0.496	0.723	0.362	0.483	0.974
BluLab: run 1-3	**0.797**	0.664	**0.725**	**0.778**	0.652	**0.709**	0.819
KPSCMI: run 1	0.272	0.782	0.404	0.223	0.642	0.331	0.948
KPSCMI: run 2	0.705	0.683	0.694	0.668	0.648	0.658	0.948
KPSCMI: run 3	0.693	0.706	0.699	0.657	0.669	0.663	0.973
UFPRSheffield-SVM: run 1	0.732	0.661	0.695	0.712	0.643	0.676	0.977
UFPRSheffield-SVM: run 2	0.741	0.655	0.695	0.723	0.640	0.679	0.950
UFPRSheffield-Hynx: run 1	0.479	0.747	0.584	0.455	0.709	0.555	0.952
UFPRSheffield-Hynx: run 2	0.494	0.770	0.602	0.470	0.733	0.573	0.951
UFPRSheffield-Hynx: run 3	0.311	0.794	0.447	0.296	0.756	0.425	0.951
UFPRSheffield-Hynx: run 4	0.311	**0.795**	0.447	0.296	**0.756**	0.425	0.952
UFPRSheffield-Hynx: run 5	0.411	**0.795**	0.542	0.391	**0.756**	0.516	**0.978**
Agreement: ann-ann	-	-	0.690	-	-	0.644	0.933
Agreement: adj-ann	-	-	0.774	-	-	0.747	0.965

Table 2: System performance and annotator agreement on TIMEX3 tasks: identifying the time expression's span (character offsets) and class (DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET). The best system score from each column is in bold. The three BluLab runs are combined because they all have identical performance (since they only differ in their approach to narrative container relations).

Precision and recall are not reported in these scenarios since they depend on the arbitrary choice of one annotator as the "human" (H) and the other as the "system" (S).

Note that since temporal relations between events and the document creation time were annotated at the same time as the events themselves, agreement for this task is only reported in phase 1 of the evaluation. Similarly, since narrative container relations were only annotated after events and times had been adjudicated, agreement for this task is only reported in phase 2 of the evaluation.

8 Evaluation Results

8.1 Time Expressions

Table 2 shows results on the time expression tasks. The BluLab system achieved the best F_1 at identifying time expressions, 0.725. The other machine learning systems (KPSCMI run 2-3 and UFPRSheffield-SVM run 1-2) achieved F_1 in the 0.690-0.700 range. The rule-based systems (KPSCMI run 1 and UFPRSheffield-Hynx run 1-5) all achieved higher recall than the machine learning systems, but at substantial costs to precision. All systems outperformed the memorization baseline in terms of recall, and all machine-learning systems outperformed it in terms of F_1, but only the BluLab system outperformed the baseline in terms of precision.

The BluLab system also achieved the best F_1 for predicting the classes of time expressions, though this is primarily due to achieving a higher F_1 at identifying time expressions in the first place. UFPRSheffield-Hynx run 5 achieved the best accuracy on predicting classes for the time expressions it found, 0.978, though on this metric it only outperformed the memorization baseline by 0.004.

Across the time expression tasks, systems did not quite achieve performance at the level of human agreement. For the spans of time expressions, the top system achieved 0.725 $F1$, compared to 0.774 adjudicator-annotator F_1, though almost half of the systems exceeded the lower annotator-annotator F_1 of 0.690. For the classes of time expressions, the story was similar for F_1, though several models exceeded the adjudicator-annotator accuracy of 0.965 on just the time expressions predicted by the system.

8.2 Event Expressions

Table 3 shows results on the event expression tasks. The BluLab system outperformed the memorization baseline on almost every metric on every task. The

	span			span + modality				span + degree			
Team	P	R	F1	P	R	F1	A	P	R	F1	A
Baseline: memorize	0.876	0.810	0.842	0.810	0.749	0.778	0.924	0.871	0.806	0.838	**0.995**
BluLab: run 1-3	**0.887**	**0.864**	**0.875**	**0.834**	**0.813**	**0.824**	**0.942**	**0.882**	**0.859**	**0.870**	0.994
Agreement: ann-ann	-	-	0.819	-	-	0.779	0.951	-	-	0.815	0.995
Agreement: adj-ann	-	-	0.880	-	-	0.855	0.972	-	-	0.877	0.997

	span + polarity				span + type			
Team	P	R	F1	A	P	R	F1	A
Baseline: memorize	0.800	0.740	0.769	0.913	**0.846**	0.783	0.813	**0.966**
BluLab: run 1-3	**0.868**	**0.846**	**0.857**	**0.979**	0.834	**0.812**	**0.823**	0.941
Agreement: ann-ann	-	-	0.798	0.974	-	-	0.773	0.944
Agreement: adj-ann	-	-	0.869	0.988	-	-	0.853	0.969

Table 3: System performance and annotator agreement on EVENT tasks: identifying the event expression's span (character offsets), contextual modality (ACTUAL, HYPOTHETICAL, HEDGED or GENERIC), degree (MOST, LITTLE or N/A), polarity (POS or NEG) and type (ASPECTUAL, EVIDENTIAL or N/A). The best system score from each column is in bold.

one exception was the semantic type of the event, where the memorization baseline had a better precision and also a better accuracy on the classes of the events that it identified.

The BluLab system got close to the level of adjudicator-annotator agreement on identifying the spans of event expressions (0.875 vs. 0.880 F_1), identifying the degree of events (0.870 vs. 0.877 F_1), and identifying the polarity of events (0.857 vs. 0.869 F_1), and it generally met or exceeded the lower annotator-annotator agreement on these tasks. There is a larger gap (3+ points of F_1) between the system performance and adjudicator-annotator agreement for event modality and event type, though only a small gap (<1 point of F_1) for the lower annotator-annotator agreement on these tasks.

8.3 Temporal Relations

Table 4 shows performance on the temporal relation tasks. In detecting the relations between events and the document creation time, the BluLab system substantially outperformed the memorization baseline, achieving F_1 of 0.702 on system-predicted events (phase 1) and F_1 of 0.791 on manually annotated events (phase 2). In identifying narrative container relations, the best BluLab system (run 2) outperformed the proximity-based baseline when using system-predicted events (F_{closure} of 0.123 vs. 0.106) but not when using manually annotated events (F_{closure} of 0.181 vs. 0.260). Across both phase 1 and phase 2 for narrative container relations, the top BluLab system always had the best recall, while the baseline system always had the best precision.

Annotator agreement was higher than system performance on all temporal relation tasks. For relations between events and the document creation time, adjudicator-annotator agreement was 0.761 F_1, compared to the best system's 0.702 F_1, though this system did exceed the lower annotator-annotator agreement of 0.628 F_1. For narrative container relations using manually annotated EVENTs and TIMEX3s, the gap was much greater, with adjudicator-annotator agreement at 0.672 F_{closure}, and the top system (the baseline system) at 0.260 F_{closure}. Even the lower annotator-annotator agreement of 0.475 F_{closure} was much higher than the system performance.

9 Discussion

The results of Clinical TempEval 2015 suggest that a small number of temporal information extraction tasks are solved by current state-of-the-art systems, but for the majority of tasks, there is still room for improvement. Identifying events, their degrees and their polarities were the easiest tasks for the participants, with the best systems achieving within about 0.01 of human agreement on the tasks. Systems for identifying event modality and event type were not far behind, achieving within about 0.03 of human agree-

	To document time			Narrative containers					
				Without closure			With closure		
	P	R	F1	P	R	F1	P	R	F1
Phase 1: systems are given only the raw text									
Baseline: memorize	0.600	0.555	0.577	-	-	-	-	-	-
Baseline: closest	-	-	-	**0.368**	0.061	**0.104**	**0.400**	0.061	0.106
BluLab: run 1	**0.712**	**0.693**	**0.702**	0.085	0.080	0.082	0.100	0.099	0.100
BluLab: run 2	**0.712**	**0.693**	**0.702**	0.080	**0.142**	0.102	0.094	**0.179**	**0.123**
BluLab: run 3	**0.712**	**0.693**	**0.702**	0.084	0.086	0.085	0.090	0.103	0.096
Agreement: ann-ann	-	-	0.628	-	-	-	-	-	-
Agreement: adj-ann	-	-	0.761	-	-	-	-	-	-
Phase 2: systems are given manually annotated EVENTs and TIMEX3s									
Baseline: memorize	-	-	0.608	-	-	-	-	-	-
Baseline: closest	-	-	-	**0.514**	0.170	0.255	**0.554**	0.170	**0.260**
BluLab: run 1	-	-	**0.791**	0.100	0.104	0.102	0.117	0.128	0.123
BluLab: run 2	-	-	**0.791**	0.109	**0.210**	0.143	0.140	**0.254**	0.181
BluLab: run 3	-	-	**0.791**	0.119	0.137	0.128	0.150	0.155	0.153
Agreement: ann-ann	-	-	-	-	-	0.449	-	-	0.475
Agreement: adj-ann	-	-	-	-	-	0.655	-	-	0.672

Table 4: System performance and annotator agreement on temporal relation tasks: identifying relations between events and the document creation time (DOCTIMEREL), and identifying narrative container relations (CONTAINS). The best system score from each column is in bold.

ment. Time expressions and relations to the document creation time were at the next level of difficulty, with a gap of about 0.05 from human agreement.

Identifying narrative container relations was by far the most difficult task, with the best systems down by more than 0.40 from human agreement. In absolute terms, performance on narrative container relations was also quite low, with system F_{closure} scores in the 0.10-0.12 range on system-generated events and times, and in the 0.12-0.26 range on manually-annotated events and times. For comparison, in TempEval 2013, which used newswire data, F_{closure} scores were in the 0.24-0.36 range on system-generated events and times and in the 0.35-0.56 range on manually-annotated events and times (UzZaman et al., 2013). One major difference between the corpora is that the narrative container relations in the clinical domain often span many sentences, while almost all of the relations in TempEval 2013 were either within the same sentence or across adjacent sentences. Most past research systems have also focused on identifying within-sentence and adjacent-sentence relations. This focus on local relations might explain the poor performance on the more distant relations in the THYME corpus. But further investigation is needed to better understand the challenge here.

In almost all tasks, the submitted systems substantially outperformed the baselines. The one exception to this was the narrative containers task. The baseline there – which simply predicted that each time expression contained the nearest event expression to it in the text – achieved 4 times the precision of the best submitted system and consequently achieved the best F_1 by a large margin. This suggests that future systems may want to incorporate better measures of proximity that can capture some of what the baseline is finding.

While machine learning methods were overall the most successful, for time expression identification, the submitted rule-based systems achieved the best recall. This is counter to the usual assumption that rule-based systems will be more precise, and that machine learning systems will sacrifice precision to increase recall. The difference is likely that the rule-based systems were aiming for good coverage, trying to find all potential time expressions, but had too few constraints to discard such phrases in inappropriate contexts. The baseline system is suggestive

of this possibility: it has a constraint to only memorize phrases that corresponded with time expressions more than 50% of the time, and it has high precision (0.743) and low recall (0.372) as is typically expected of a rule-based system, but if the constraint is removed, it has low precision (0.126) and high recall (0.521) like the participant rule-based systems.

Clinical TempEval was the first TempEval exercise to use narrative containers, a significant shift from prior exercises. Annotator agreement in the dataset is moderate, but needs to be further improved. Similar agreement scores were found when annotating temporal relations in prior corpora (for TempEval or using TimeML), although these typically involved the application of more complex temporal relation ontologies. The narrative container approach is comparatively simple. The low annotator-adjudicator scores (i.e. below 0.90, a score generally recognized to indicate a production-quality resource) suggests that annotation is difficult independent of the number of potential temporal relation types. Difficulty may lie in the comprehension and reification of the potentially complex temporal structures described in natural language text. Nevertheless, systems did well on the DCT task, achieving high scores – similar to the pattern seen in Task D of TempEval-2, which had a comparable scoring metric.

Though the results of Clinical TempEval 2015 are encouraging, they were limited somewhat by the small number of participants in the task. There are two likely reasons for this. First, there were many different sub-tasks for Clinical TempEval, meaning that to compete in all sub-tasks, a large number of sub-systems had to be developed in a limited amount of time (six months or less). This relatively high barrier for entry meant that of the 15 research groups that managed to sign a data use agreement and obtain the data before the competition, only 3 submitted systems to compete. Second, the data use agreement process was time consuming, and more than 10 research groups who began the data use agreement process were unable to complete it before the evaluation.

In future iterations of Clinical TempEval, we expect these issues to be reduced. The next Clinical TempEval will use the current Train and Dev data as the training set, and as these data are already available, this leaves research teams with a year or more to develop systems. Furthermore, arrangements with the Mayo Clinic have been made to further expedite the data use agreement process, which should significantly reduce the wait time for new participants.

Acknowledgements

This work was partially supported by funding from R01LM010090 (THYME) from the National Library of Medicine and from the European Union's Seventh Framework Programme (grant No. 611233, PHEME).

References

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, pages 14–19, Atlanta, Georgia, June.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160, Portland, Oregon, USA, June.

William F. Styler, IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014a. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

William F. Styler, IV, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen. 2014b. THYME annotation guidelines, 2.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356, Portland, Oregon, USA, June.

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.

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.

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.