# 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

**James Pustejovsky**
Brandeis University
Waltham, MA 02453
USA
jamesp@cs.brandeis.edu

**Marc Verhagen**
Brandeis University
Waltham, MA 02453
USA
marc@cs.brandeis.edu

## Abstract

We describe the Clinical TempEval task which is currently in preparation for the SemEval-2015 evaluation exercise. This task involves identifying and describing events, times and the relations between them in clinical text. Six discrete subtasks are included, focusing on recognising mentions of times and events, describing those mentions for both entity types, identifying the relation between an event and the document creation time, and identifying narrative container relations.

## 1  Summary

ClinicalTempEval will bring the temporal information extraction tasks of previous TempEvals to the clinical domain, using clinical notes of colon cancer patients from the Mayo Clinic. Our definitions of events and times are based on those in TimeML (Pustejovsky et al., 2003a; Pustejovsky et al., 2010). ClinicalTempEval will provide the following temporal annotation sub-tasks:

- TS: identifying the spans of time expressions
- ES: Identifying the spans of event expressions
- TA: identifying the attributes of time expressions (type = DATE / TIME / DURATION / QUANTI-FIER / PREPOSTEXP / SET; value = TIMEX3.val) – see (Ferro et al., 2005) for additional details of the TIMEX specification
- EA: identifying the attributes of event expressions (type = NA/ASPECTUAL/EVIDENTIAL; polarity = POS / NEG; degree = NA / MOST / LITTLE; modality = ACTUAL / HEDGED / HYPO-THETICAL / GENERIC)
- DR: identifying the relation between an event and the document creation time (docTimeRel = BE-FORE / OVERLAP / AFTER / BEFORE-OR-OVERLAP)
- CR: identifying narrative container relations (CONTAINS a.k.a. INCLUDES)

These sub-tasks are largely common to previous similar exercises. They will be presented in three scenarios, detailed in Section 5.

## 2  Motivation

The TempEval task has, since 2007, provided a focus for research on temporal information extraction (Verhagen et al., 2009; Verhagen et al., 2010; UzZaman et al., 2013). The automatic identication of all temporal referring expressions, events and temporal relations within a text is the ultimate aim of research in this area. As a result of previous TempEvals, we have discovered much about temporal information extraction, identifying the difficulties in the area and contributing new resources.

TempEvals information extraction exercises have been presented as discrete, well-defined tasks, with automatic and quantitative evaluation of approaches a key part. We continue this format, looking at

the most difficult parts of temporal IE in a formal manner. This makes evaluation rapid, reliable and repeatable, focused on the key parts of the problem instead of potentially harder-to-assess situations like, for example, free-form question answering or event-based summarisation.

ClinicalTempEval extends TempEval to the clinical domain. This move is for two key reasons. First, concentrating on newswire constrains our understanding of time in language to a particular range of expressions, event types and timeframes. Second, there is great interest in temporal information extraction in this domain, and great utility to be had in solving it. For example, in 2012 a traditional clinical NLP challenge (i2b2) ran their shared task on just this problem (Sun et al., 2013). We have already acquired annotations over sharable clinical texts for the task.

## 3 Data

The THYME project[1] is currently annotating times, events and temporal relations in clinical notes following guidelines derived from ISO TimeML for the THYME project. (The i2b2 guidelines were derived from the THYME guidelines and are essentially a subset of the THYME guidelines.) The annotation pipeline is as follows:

1. Annotators identify time and event expressions, along with their attributes (for events, attributes include the temporal relation to the document creation time)
2. Adjudicators revise and finalize the time and event expressions and their attributes
3. Annotators identify temporal relations between pairs of events and between events and times
4. Adjudicators revise and finalize the temporal relations

Currently, 232 notes from 87 patients have been annotated, with over 30000 events, 2500 times and 9000 narrative container relations (Pustejovsky and Stubbs, 2011; Miller et al., 2013). We anticipate around 600 notes from 200 patients by the time the TempEval 2015 training data is released.

The PHEME project[2] which starts January 2014 involves annotation of spatial and temporal aspects of non-newswire text (with a focus on social media). This project will provide annotations for time expression values.

For Clinical TempEval, we will use splits at patient record level. This means that patient data will not leak across datasets. One half of the patient records will be used as training data, one quarter as development data, and the final quarter as the test set. This gives a test set roughly the size of half of the TimeBank corpus (Pustejovsky et al., 2003b).

## 4 Data Use Agreements

All clinical notes have been de-identified, but access to the TempEval 2015 data will still require participants to sign a data use agreement with the Mayo Clinic, to ensure that the data is handled appropriately. After the competition, the data set will be available to other researchers (though again, requiring a data use agreement).

## 5 Evaluation

We envision three different evaluation setups:

1. Only the plain text is given
2. Manually annotated event and time expression spans are given
3. Manually annotated event and time expression spans and attributes are given

Evaluation for each setup will be as follows:

1. Only the plain text is given
   - TS, ES: precision, recall and F1

---

[1] See http://projectreporter.nih.gov/project_info_description.cfm?aid=8138604&icde=10245671
[2] See http://www.pheme.eu

- TA, EA: precision, recall and F1 for each attribute, and an overall precision, recall and F1 where a time/event is marked correct only if all attributes are correct
- DR: precision, recall and F1
- CR: precision, recall and F1, and closure-based precision, recall and F1, where temporal closure is run to infer additional relations on both the system and the reference relations and scores are calculated on the post-closure relations.

2. Manually annotated event and time expression spans are given
   - TA, EA: accuracy for each attribute, and an overall accuracy where a time/event is marked correct only if all attributes are correct
   - DR: accuracy
   - CR: precision, recall and F1, and closure-based precision, recall and F1.

3. Manually annotated event and time expression spans and attributes are given
   - DR: accuracy
   - CR: precision, recall and F1, and closure-based precision, recall and F1.

## 6 Resources Required

Annotation costs are covered by the THYME project. There is already sufficient data available now to run a shared task, but annotation is ongoing, and we will make available whatever has been fully annotated and adjudicated at the time of the TempEval 2015 training data release.

The main resource that still needs to be developed is the evaluation scripts, used by the official evaluation, and also in a form that can be distributed to participants.

## 7 Baseline Systems

We will provide several baseline systems to compare against, such as:

- For TS, ES, TA and EA: If an event/time was seen in the training data and its seen in the test data, mark it as an event/time and give identical attributes to whatever it had in the training data
- For DR: the most common class and/or a memorization baseline like above
- For CR: link each event to the closest time expression in the same sentence

## 8 Organizers

- Steven Bethard <bethard@cis.uab.edu>, The University of Alabama at Birmingham
- Leon Derczynski <leon@dcs.shef.ac.uk>, The University of Sheffield
- James Pustejovsky <jamesp@cs.brandeis.edu>, Brandeis University
- Marc Verhagen <marc@cs.brandeis.edu>, Brandeis University

### Acknowledgments

### References

[Ferro et al.2005] L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. Tides 2005 standard for the annotation of temporal expressions.

[Miller et al.2013] T. A. Miller, S. Bethard, D. Dligach, S. Pradhan, C. Lin, and G. K. Savova. 2013. Discovering narrative containers in clinical text. *ACL 2013*, page 18.

[Pustejovsky and Stubbs2011] J. Pustejovsky and A. Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

[Pustejovsky et al.2003a] J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

[Pustejovsky et al.2003b] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. 2003b. The TimeBank corpus. In *Corpus linguistics*, volume 2003, page 40.

[Pustejovsky et al.2010] J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*.

[Sun et al.2013] W. Sun, A. Rumshisky, and O. Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

[UzZaman et al.2013] N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, and J. 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, pages 1–9.

[Verhagen et al.2009] M. Verhagen, 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 et al.2010] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proc. SemEval*, pages 57–62.