SemEval-2010 Task 13: Evaluating Events, Time Expressions, and Temporal Relations (TempEval-2)

James Pustejovsky Computer Science Department Brandeis University Waltham, Massachusetts, USA jamesp@cs.brandeis.edu

Abstract

We describe the TempEval-2 task which is currently in preparation for the SemEval-2010 evaluation exercise. This task involves identifying the temporal relations between events and temporal expressions in text. Six distinct subtasks are defined, ranging from identifying temporal and event expressions, to anchoring events to temporal expressions, and ordering events relative to each other.

1 Introduction

Newspaper texts, narratives and other such texts describe events which occur in time and specify the temporal location and order of these events. Text comprehension, even at the most general level, involves the capability to identify the events described in a text and locate these in time. This capability is crucial to a wide range of NLP applications, from document summarization and question answering to machine translation. As in many areas of NLP, an open evaluation challenge in the area of temporal annotation will serve to drive research forward.

The automatic identification of all temporal referring expressions, events, and temporal relations within a text is the ultimate aim of research in this area. However, addressing this aim in a first evaluation challenge was deemed too difficult and a staged approach was suggested. The 2007 SemEval task, TempEval (henceforth TempEval-1), was an initial evaluation exercise based on three limited tasks that were considered realistic both from the perspective of assembling resources for development and testing and from the perspective of developing systems capable of addressing the tasks. Marc Verhagen

Computer Science Department Brandeis University Waltham, Massachusetts, USA marc@cs.brandeis.edu

We are now preparing TempEval-2, a temporal evaluation task based on TempEval-1. TempEval-2 is more elaborate in two respects: (i) it is a multilingual task, and (ii) it consists of six subtasks rather than three.

2 TempEval-1

TempEval-1 consisted of three tasks:

- A. determine the relation between an event and a timex in the same sentence;
- B. determine the relation between an event and the document creation time;
- C. determine the relation between the main events of two consecutive sentences.

The data sets were based on TimeBank (Pustejovsky et al., 2003; Boguraev et al., 2007), a handbuilt gold standard of annotated texts using the TimeML markup scheme.¹ The data sets included sentence boundaries, TIMEX3 tags (including the special document creation time tag), and EVENT tags. For tasks A and B, a restricted set of events was used, namely those events that occur more than 5 times in TimeBank. For all three tasks, the relation labels used were BEFORE, AFTER, OVER-LAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE.² For a more elaborate description of TempEval-1, see (Verhagen et al., 2007; Verhagen et al., 2009).

¹See www.timeml.org for details on TimeML, Time-Bank is distributed free of charge by the Linguistic Data Consortium (www.ldc.upenn.edu), catalog number LDC2006T08.

²Which is different from the set of 13 labels from TimeML. The set of labels for TempEval-1 was simplified to aid data preparation and to reduce the complexity of the task.

There were six systems competing in TempEval-1: University of Colorado at Boulder (CU-TMP); Language Computer Corporation (LCC-TE); Nara Institute of Science and Technology (NAIST); University of Sheffield (USFD); Universities of Wolverhampton and Allicante (WVALI); and XEROX Research Centre Europe (XRCE-T).

The difference between these systems was not large, and details of system performance, along with comparisons and evaluation, are presented in (Verhagen et al., 2009). The scores for WVALI's hybrid approach were noticeably higher than those of the other systems in task B and, using relaxed scoring, in task C as well. But for task A, the highest scoring systems are barely ahead of the rest of the field. Similarly, for task C using strict scoring, there is no system that clearly separates itself from the field. Interestingly, the baseline is close to the average system performance on task A, but for other tasks the system scores noticeably exceed the baseline. Note that the XRCE-T system is somewhat conservative in assigning TLINKS for tasks A and B, producing lower recall scores than other systems, which in turn yield lower f-measure scores. For task A, this is mostly due to a decision only to assign a temporal relation between elements that can also be linked by the syntactic analyzer.

3 TempEval-2

The set of tasks chosen for TempEval-1 was by no means complete, but was a first step towards a fuller set of tasks for temporal parsing of texts. While the main goal of the division in subtasks was to aid evaluation, the larger goal of temporal annotation in order to create a complete temporal characterization of a document was not accomplished. Results from the first competition indicate that task A was defined too generally. As originally defined, it asks to temporally link all events in a sentence to all time expressions in the same sentence. A clearer task would have been to solicit local anchorings and to separate these from the less well-defined temporal relations between arbitrary events and times in the same sentence. We expect both inter-annotator agreement and system performance to be higher with a more precise subtask. Thus, the set of tasks used in TempEval-1 is far from complete and the tasks

could have been made more restrictive. As a result, inter-annotator agreement scores lag, making precise evaluation more challenging.

The overall goal of temporal tagging of a text is to provide a temporal characterization of a set of events that is as complete as possible. If the annotation graph of a document is not completely connected then it is impossible to determine temporal relations between two arbitrary events because these events could be in separate subgraphs. Hence, for the current competition, TempEval-2, we have enriched the task description to bring us closer to creating such a temporal characterization for a text. We have enriched the TempEval-2 task definition to include six distinct subtasks:

- A. Determine the extent of the time expressions in a text as defined by the TimeML TIMEX3 tag. In addition, determine value of the features TYPE and VAL. The possible values of TYPE are TIME, DATE, DURATION, and SET; the value of VAL is a normalized value as defined by the TIMEX2 and TIMEX3 standards.
- B. Determine the extent of the events in a text as defined by the TimeML EVENT tag. In addition, determine the value of the features TENSE, ASPECT, POLARITY, and MODALITY.
- C. Determine the temporal relation between an event and a time expression in the same sentence. For TempEval-2, this task is further restricted by requiring that either the event syntactically dominates the time expression or the event and time expression occur in the same noun phrase.
- D. Determine the temporal relation between an event and the document creation time.
- E. Determine the temporal relation between two main events in consecutive sentences.
- F. Determine the temporal relation between two events where one event syntactically dominates the other event. This refers to examples like "she *heard* an *explosion*" and "he *said* they *postponed* the meeting".

The complete TimeML specification assumes the temporal interval relations as defined by Allen (Allen, 1983) in Figure 1.

	A EQUALS B
┝┻┥ ┝┺┥	A is BEFORE B; B is AFTER A
	A MEETS B; B is MET BY A
	A OVERLAPS B; B is OVERLAPPED BY A
	A STARTS B; B is STARTED BY A
	A FINISHES B; B is FINISHED BY A
	A is DURING B; B CONTAINS A

Figure 1: Allen Relations

For this task, however, we assume a reduced subset, as introduced in TempEval-1: BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE. However, we are investigating whether for some tasks the more precise set of TimeML relations could be used.

Task participants may choose to either do all tasks, focus on the time expression task, focus on the event task, or focus on the four temporal relation tasks. In addition, participants may choose one or more of the five languages for which we provide data: English, Italian, Chinese, Spanish, and Korean.

3.1 Extent of Time Expression

This task involves identification of the EXTENT, TYPE, and VAL of temporal expressions in the text. Times can be expressed syntactically by adverbial or prepositional phrases, as shown in the following:

(1) a. on Thursday

- b. November 15, 2004
- c. Thursday evening
- d. in the late 80's
- e. Later this afternoon
- f. yesterday

The TYPE of the temporal extent must be identified. There are four temporal types that will be distinguished for this task;

- (2) a. Time: *at 2:45 p.m.*
 - b. Date: January 27, 1920, yesterday
 - c. Duration two weeks
 - d. Set: every Monday morning

The VAL attribute will assume values according to an extension of the ISO 8601 standard, as enhanced by TIMEX2.

(3) November 22, 2004

```
<TIMEX3 tid="t1" type="DATE" value="2004-11-22"/>
```

3.2 Extent of Event Expression

The EVENT tag is used to annotate those elements in a text that describe what is conventionally referred to as an *eventuality*. Syntactically, events are typically expressed as inflected verbs, although event nominals, such as "crash" in *killed by the crash*, should also be annotated as EVENTS.

In this task, event extents must be identified and tagged with EVENT, along with values for the features TENSE, ASPECT, POLARITY, and MODALITY. Examples of these features are shown below:

(4) should have *bought*

```
<EVENT id="e1" pred="BUY" pos="VERB"
tense="PAST" aspect="PERFECTIVE"
modality="SHOULD" polarity="POS"/>
```

(5) did not *teach*

```
<EVENT id="e2" pred="TEACH" pos="VERB"
tense="PAST" aspect="NONE"
modality="NONE" polarity="NEG"/>
```

The specifics on the definition of event extent will follow the published TimeML guideline (cf. timeml.org).

3.3 Within-sentence Event-Time Anchoring

This task involves determining the temporal relation between an event and a time expression in the same sentence. This was present in TempEval-1, but here, in TempEval-2, this problem is further restricted by requiring that the event either syntactically dominates the time expression or the event and time expression occur in the same noun phrase. For example, the following constructions will be targeted for temporal labeling:

- (6) Mary taught_{e1} on Tuesday morning_{t1} OVERLAP(e1,t1)
- (7) They cancelled the *evening*_{t2} $class_{e2}$ OVERLAP(e2,t2)

3.4 Neighboring Sentence Event-Event Ordering

In this task, the goal is to identify the temporal relation between two main events in consecutive sentences. This task was covered in the previous competition, and includes pairs such as that shown below:

(8) The President $spoke_{e1}$ to the nation on Tuesday on the financial crisis. He had $conferred_{e2}$ with his cabinet regarding policy the day before. AFTER(e1,e2)

3.5 Sentence Event-DCT Ordering

This task was also included in TempEval-1 and requires the identification of the temporal order between the matrix event of the sentence and the Document Creation Time (DCT) of the article or text. For example, the text fragment below specifies a fixed DCT, relative to which matrix events from the two sentences are ordered:

(9) DCT: MARCH 5, 2009

a. Most troops will $leave_{e1}$ Iraq by August of 2010. AFTER(e1,dct)

b. The country $defaulted_{e2}$ on debts for that entire year. BEFORE(e2,dct)

3.6 Within-sentence Event-Event Ordering

The final task involves identifying the temporal relation between two events, where one event syntactically dominates the other event. This includes examples such as those illustrated below.

- (10) The students $heard_{e1}$ a fire $alarm_{e2}$. OVERLAP(e1,e2)
- (11) He said_{e1} they had postponed_{e2} the meeting. AFTER(e1,e2)

4 Resources and Evaluation

4.1 Data

The development corpus will contain the following data:

- 1. Sentence boundaries;
- 2. The document creation time (DCT) for each document;
- 3. All temporal expressions in accordance with the TimeML TIMEX3 tag;
- 4. All events in accordance with the TimeML EVENT tag;
- 5. Main event markers for each sentence;
- 6. All temporal relations defined by tasks C through F.

The data for the five languages are being prepared independently of each other. We do not provide a parallel corpus. However, annotation specifications and guidelines for the five languages will be developed in conjunction with one other. For some languages, we may not use all four temporal linking tasks. Data preparation is currently underway for English and will start soon for the other languages. Obviously, data preparation is a large task. For English and Chinese, the data are being developed at Brandeis University under three existing grants.

For evaluation data, we will provide two data sets, each consisting of different documents. DataSet1 is for tasks A and B and will contain data item 1 and 2 from the list above. DataSet2 is for tasks C though F and will contain data items 1 through 5.

4.2 Data Preparation

For all languages, annotation guidelines are defined for all tasks, based on version 1.2.1 of the TimeML annotation guidelines for English³. The most notable changes relative to the previous TimeML guidelines are the following:

• The guidelines are not all presented in one document, but are split up according to the seven TempEval-2 tasks. Full temporal annotation has proven to be a very complex task, splitting it into subtasks with separate guidelines for

³See http://www.timeml.org.

each task has proven to make temporal annotation more manageable.

- It is not required that all tasks for temporal linking (tasks C through F) use the same relation set. One of the goals during the data preparation phase is to determine what kind of relation set makes sense for each individual task.
- The guidelines can be different depending on the language. This is obviously required because time expressions, events, and relations are expressed differently across languages.

Annotation proceeds in two phases: a dual annotation phase where two annotators annotate each document and an adjudication phase where a judge resolves disagreements between the annotators. We are expanding the annotation tool used for TempEval-1, making sure that we can quickly annotate data for all tasks while making it easy for a language to define an annotation task in a slightly different way from another language. The Brandeis Annotation Tool (BAT) is a generic web-based annotation tool that is centered around the notion of annotation tasks. With the task decomposition allowed by BAT, it is possible to flexibly structure the complex task of temporal annotation by splitting it up in as many sub tasks as seems useful. As such, BAT is well-suited for TempEval-2 annotation. Comparison of annotation speed with tools that do not allow task decomposition showed that annotation with BAT is up to ten times faster. Annotation has started for Italian and English.

For all tasks, precision and recall are used as evaluation metrics. A scoring program will be supplied for participants.

5 Conclusion

In this paper, we described the TempEval-2 task within the SemEval 2010 competition. This task involves identifying the temporal relations between events and temporal expressions in text. Using a subset of TimeML temporal relations, we show how temporal relations and anchorings can be annotated and identified in five different languages. The markup language adopted presents a descriptive framework with which to examine the temporal aspects of natural language information, demonstrating in particular, how tense and temporal information is encoded in specific sentences, and how temporal relations are encoded between events and temporal expressions. This work paves the way towards establishing a broad and open standard metadata markup language for natural language texts, examining events, temporal expressions, and their orderings.

References

- James Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.
- Bran Boguraev, James Pustejovsky, Rie Ando, and Marc Verhagen. 2007. Timebank evolution as a community resource for timeml parsing. *Language Resource and Evaluation*, 41(1):91–115.
- James Pustejovsky, David Day, Lisa Ferro, Robert Gaizauskas, Patrick Hanks, Marcia Lazo, Roser Saurí, Andrew See, Andrea Setzer, and Beth Sundheim. 2003. The TimeBank Corpus. *Corpus Linguistics*, March.
- Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In Proc. of the Fourth Int. Workshop on Semantic Evaluations (SemEval-2007), pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.
- Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*.