TempEval2:
Evaluating Events, Time Expressions and Temporal Relations

James Pustejovsky, Marc Verhagen, Xue Nianwen, Robert Gaizauskas,
Mark Hepple, Frank Schilder, Graham Katz, Roser Saurí, Estela Saquete,
Tommaso Caselli, Nicoletta Calzolari, Kiyong Lee, and Seohyun Im

SemEval Task Proposal

1 Introduction

In SemEval-2007, the TempEval task was added as a new task focused on temporal relations. In the task proposal, the usefulness of such a task was motivated as follows:

> "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. 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.

We now propose 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 TempEval2

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, a hand-built 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, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE.[2]

3 TempEval2

The six proposed tasks for TempEval-2 are:

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". However, we are investigating whether for some tasks the more precise set of TimeML relations could be used.

The relation labels used are the same as for TempEval-1: BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE. Task participants can choose to do do all task, focus on the time expression task, focus on the event task or focus on the four temporal relation tasks. In addition, participants choose one of the five languages for which we provide data: English, Italian, Chinese, Spanish and Korean.

3.1 Resources

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

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

[2] Which 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.

The data for the five languages are prepared independently from each other. We do not provide a parallel corpus. However, annotation specifications and guidelines for the five languages will be developed in conjunction with each 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, the data are developed by Brandeis University under two 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 Evaluation Methodology

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