SemEval-2007 Task 15: TempEval Temporal Relation Identification

Marc Verhagen[†], Robert Gaizauskas[‡], Frank Schilder[⋆], Mark Hepple[‡], Graham Katz[*] and James Pustejovsky[†]

[†] Brandeis University, {marc, jamesp}@cs.brandeis.edu
[‡] University of Sheffield, {r.gaizauskas,m.hepple}@dcs.shef.ac.uk
[⋆] Thomson Legal & Regulatory, frank.schilder@thomson.com,
[*] Stanford University, egkatz@stanford.edu

Abstract

The TempEval task proposes a simple way to evaluate automatic extraction of temporal relations. It avoids the pitfalls of evaluating a graph of inter-related labels by defining three sub tasks that allow pairwise evaluation of temporal relations. The task not only allows straightforward evaluation, it also avoids the complexities of full temporal parsing.

1 Introduction

Newspaper texts, narratives and other texts describe events that occur in time and specify the temporal location and order of these events. Text comprehension, amongst other capabilities, clearly requires 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.

Recent work on the annotation of events and temporal relations has resulted in both a de-facto standard for expressing these relations and a hand-built gold standard of annotated texts. TimeML (Pustejovsky et al., 2003a) is an emerging ISO standard for annotation of events, temporal expressions and the anchoring and ordering relations between them. TimeBank (Pustejovsky et al., 2003b; Boguraev et al., forthcoming) was originally conceived of as a proof of concept that illustrates the TimeML language, but has since gone through several rounds of revisions and can now be considered a gold standard for temporal information. TimeML and TimeBank have already been used as the basis for automatic time, event and temporal relation annotation tasks in a number of research projects in recent years (Mani et al., 2006; Boguraev et al., forthcoming).

An open evaluation challenge in the area of temporal annotation should serve to drive research forward, as it has in other areas of NLP. 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 judged to be too difficult, both for organizers and participants, and a staged approach was deemed more effective. Thus we here present an initial evaluation exercise based on three limited tasks that we believe are realistic both from the perspective of assembling resources for development and testing and from the perspective of developing systems capable of addressing the tasks. They are also tasks, which should they be performable automatically, have application potential.

2 Task Description

The tasks as originally proposed were modified slightly during the course of resource development for the evaluation exercise due to constraints on data and annotator availability. In the following we describe the tasks as they were ultimately realized in the evaluation.

There were three tasks – A, B and C. For all three tasks the data provided for testing and training includes annotations identifying: (1) sentence boundaries; (2) all temporal referring expression as

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 75–80,
Prague, June 2007.

specified by TIMEX3; (3) all events as specified in TimeML; (4) selected instances of temporal relations, as relevant to the given task. For tasks A and B a restricted set of event terms were identified – those whose stems occurred twenty times or more in TimeBank. This set is referred to as the Event Target List or ETL.

TASK A This task addresses only the temporal relations holding between time and event expressions that occur within the same sentence. Furthermore only event expressions that occur within the ETL are considered. In the training and test data, TLINK annotations for these temporal relations are provided, the difference being that in the test data the relation type is withheld. The task is to supply this label.

TASK B This task addresses only the temporal relations holding between the Document Creation Time (DCT) and event expressions. Again only event expressions that occur within the ETL are considered. As in Task A, TLINK annotations for these temporal relations are provided in both training and test data, and again the relation type is withheld in the test data and the task is to supply this label.

TASK C Task C relies upon the idea of their being a main event within a sentence, typically the syntactically dominant verb. The aim is to assign the temporal relation between the main events of adjacent sentences. In both training and test data the main events are identified (via an attribute in the event annotation) and TLINKs between these main events are supplied. As for Tasks A and B, the task here is to supply the correct relation label for these TLINKs.

3 Data Description and Data Preparation

The TempEval annotation language is a simplified version of TimeML [1]. For TempEval, we use the following five tags: `TempEval`, `s`, `TIMEX3`, `EVENT`, and `TLINK`. `TempEval` is the document root and `s` marks sentence boundaries. All sentence tags in the TempEval data are automatically created using the Alembic Natural Language processing tools. The other three tags are discussed here in more detail:

- `TIMEX3`. Tags the time expressions in the text. It is identical to the `TIMEX3` tag in TimeML. See the TimeML specifications and guidelines for further details on this tag and its attributes. Each document has one special `TIMEX3` tag, the Document Creation Time, which is interpreted as an interval that spans a whole day.

- `EVENT`. Tags the event expressions in the text. The interpretation of what an event is is taken from TimeML where an event is a cover term for predicates describing situations that happen or occur as well as some, but not all, stative predicates. Events can be denoted by verbs, nouns or adjectives. The TempEval event annotation scheme is somewhat simpler than that used in TimeML, whose complexity was designed to handle event expressions that introduced multiple event instances (consider, e.g. *He taught on Wednesday and Friday*). This complication was not necessary for the TempEval data. The most salient attributes encode tense, aspect, modality and polarity information. For TempEval task C, one extra attribute is added: `mainevent`, with possible values YES and NO.

- `TLINK`. This is a simplified version of the TimeML `TLINK` tag. The relation types for the TimeML version form a fine-grained set based on James Allen's interval logic (Allen, 1983). For TempEval, we use only six relation types including the three core relations BEFORE, AFTER, and OVERLAP, the two less specific relations BEFORE-OR-OVERLAP and OVERLAP-OR-AFTER for ambiguous cases, and finally the relation VAGUE for those cases where no particular relation can be established.

As stated above the `TLINK`s of concern for each task are explicitly included in the training and in the test data. However, in the latter the `relType` attribute of each `TLINK` is set to `UNKNOWN`. For each task the system must replace the `UNKNOWN` values with one of the six allowed values listed above.

The `EVENT` and `TIMEX3` annotations were taken verbatim from TimeBank version 1.2.[2] The annota-

[1] See `http://www.timeml.org` for language specifications and annotation guidelines

[2] TimeBank 1.2 is available for free through the Linguistic Data Consortium, see `http://www.timeml.org` for more

tion procedure for TLINK tags involved dual annotation by seven annotators using a web-based annotation interface. After this phase, three experienced annotators looked at all occurrences where two annotators differed as to what relation type to select and decided on the best option. For task C, there was an extra annotation phase where the main events were marked up. Main events are those events that are syntactically dominant in the sentences.

It should be noted that annotation of temporal relations is not an easy task for humans due to rampant temporal vagueness in natural language. As a result, inter-annotator agreement scores are well below the often kicked-around threshold of 90%, both for the TimeML relation set as well as the TempEval relation set. For TimeML temporal links, an inter-annotator agreement of 0.77 was reported, where agreement was measured by the average of precision and recall. The numbers for TempEval are even lower, with an agreement of 0.72 for anchorings of events to times (tasks A and B) and an agreement of 0.65 for event orderings (task C). Obviously, numbers like this temper the expectations for automatic temporal linking.

The lower number for TempEval came a bit as a surprise because, after all, there were fewer relations to choose form. However, the TempEval annotation task is different in the sense that it did not give the annotator the option to ignore certain pairs of events and made it therefore impossible to skip hard-to-classify temporal relations.

4 Evaluating Temporal Relations

In full temporal annotation, evaluation of temporal annotation runs into the same issues as evaluation of anaphora chains: simple pairwise comparisons may not be the best way to evaluate. In temporal annotation, for example, one may wonder how the response in (1) should be evaluated given the key in (2).

(1) {A before B, A before C, B equals C}
(2) {A after B, A after C, B equals C}

Scoring (1) at 0.33 precision misses the interdependence between the temporal relations. What we need to compare is not individual judgements but two partial orders.

details.

For TempEval however, the tasks are defined in a such a way that a simple pairwise comparison is possible since we do not aim to create a full temporal graph and judgements are made in isolation.

Recall that there are three basic temporal relations (BEFORE, OVERLAP, and AFTER) as well as three disjunctions over this set (BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE). The addition of these disjunctions raises the question of how to score a response of, for example, BEFORE given a key of BEFORE-OR-OVERLAP. We use two scoring schemes: strict and relaxed. The *strict scoring scheme* only counts exact matches as success. For example, if the key is OVERLAP and the response BEFORE-OR-OVERLAP than this is counted as failure. We can use standard definitions of precision and recall

$$Precision = R_c/R$$
$$Recall = R_c/K$$

where R_c is number of correct answers in the response, R the total number of answers in the response, and K the total number of answers in the key. For the *relaxed scoring scheme*, precision and recall are defined as

$$Precision = R_c w/R$$
$$Recall = R_c w/K$$

where $R_c w$ reflects the weighted number of correct answers. A response is not simply counted as 1 (correct) or 0 (incorrect), but is assigned one of the values in table 1.

	B	O	A	B-O	O-A	V
B	1	0	0	0.5	0	0.33
O	0	1	0	0.5	0.5	0.33
A	0	0	1	0	0.5	0.33
B-O	0.5	0.5	0	1	0.5	0.67
O-A	0	0.5	0.5	0.5	1	0.67
V	0.33	0.33	0.33	0.67	0.67	1

Table 1: Evaluation weights

This scheme gives partial credit for disjunctions, but not so much that non-commitment edges out precise assignments. For example, assigning VAGUE as the relation type for every temporal relation results in a precision of 0.33.

5 Participants

Six teams participated in the TempEval tasks. Three of the teams used statistics exclusively, one used a rule-based system and the other two employed a hybrid approach. This section gives a short description of the participating systems.

CU-TMP trained three support vector machine (SVM) models, one for each task. All models used the gold-standard TimeBank features for events and times as well as syntactic features derived from the text. Additionally, the relation types obtained by running the task B system on the training data for Task A and Task C, were added as a feature to the two latter systems. A subset of features was selected using cross-validations on the training data, discarding features whose removal improved the cross-validation F-score. When applied to the test data, the Task B system was run first in order to supply the necessary features to the Task A and Task C systems.

LCC-TE automatically identifies temporal referring expressions, events and temporal relations in text using a hybrid approach, leveraging various NLP tools and linguistic resources at LCC. For temporal expression labeling and normalization, they used a syntactic pattern matching tool that deploys a large set of hand-crafted finite state rules. For event detection, they used a small set of heuristics as well as a lexicon to determine whether or not a token is an event, based on the lemma, part of speech and WordNet senses. For temporal relation discovery, LCC-TE used a large set of syntactic and semantic features as input to a machine learning components.

NAIST-japan defined the temporal relation identification task as a sequence labeling model, in which the target pairs – a `TIMEX3` and an `EVENT` – are linearly ordered in the document. For analyzing the relative positions, they used features from dependency trees which are obtained from a dependency parser. The relative position between the target `EVENT` and a word in the target `TIMEX3` is used as a feature for a machine learning based relation identifier. The relative positions between a word in the target entities and another word are also introduced.

The **USFD** system uses an off-the-shelf Machine Learning suite(WEKA), treating the assignment of temporal relations as a simple classification task. The features used were the ones provided in the TempEval data annotation together with a few features straightforwardly computed from the document without any deeper NLP analysis.

WVALI's approach for discovering intra-sentence temporal relations relies on sentence-level syntactic tree generation, bottom-up propagation of the temporal relations between syntactic constituents, a temporal reasoning mechanism that relates the two targeted temporal entities to their closest ancestor and then to each other, and on conflict resolution heuristics. In establishing the temporal relation between an event and the Document Creation Time (DCT), the temporal expressions directly or indirectly linked to that event are first analyzed and, if no relation is detected, the temporal relation with the DCT is propagated top-down in the syntactic tree. Inter-sentence temporal relations are discovered by applying several heuristics and by using statistical data extracted from the training corpus.

XRCE-T used a rule-based system that relies on a deep syntactic analyzer that was extended to treat temporal expressions. Temporal processing is integrated into a more generic tool, a general purpose linguistic analyzer, and is thus a complement for a better general purpose text understanding system. Temporal analysis is intertwined with syntactico-semantic text processing like deep syntactic analysis and determination of thematic roles. TempEval-specific treatment is performed in a post-processing stage.

6 Results

The results for the six teams are presented in tables 2, 3, and 4.

team	strict			relaxed		
	P	R	F	P	R	F
CU-TMP	0.61	0.61	0.61	0.63	0.63	0.63
LCC-TE	0.59	0.57	0.58	0.61	0.60	0.60
NAIST	0.61	0.61	0.61	0.63	0.63	0.63
USFD*	0.59	0.59	0.59	0.60	0.60	0.60
WVALI	0.62	0.62	0.62	0.64	0.64	0.64
XRCE-T	0.53	0.25	0.34	0.63	0.30	0.41
average	0.59	0.54	0.56	0.62	0.57	0.59
stddev	0.03	0.13	0.10	0.01	0.12	0.08

Table 2: Results for Task A

team	strict			relaxed		
	P	R	F	P	R	F
CU-TMP	0.75	0.75	0.75	0.76	0.76	0.76
LCC-TE	0.75	0.71	0.73	0.76	0.72	0.74
NAIST	0.75	0.75	0.75	0.76	0.76	0.76
USFD*	0.73	0.73	0.73	0.74	0.74	0.74
WVALI	0.80	0.80	0.80	0.81	0.81	0.81
XRCE-T	0.78	0.57	0.66	0.84	0.62	0.71
average	0.76	0.72	0.74	0.78	0.74	0.75
stddev	0.03	0.08	0.05	0.03	0.06	0.03

Table 3: Results for Task B

team	strict			relaxed		
	P	R	F	P	R	F
CU-TMP	0.54	0.54	0.54	0.58	0.58	0.58
LCC-TE	0.55	0.55	0.55	0.58	0.58	0.58
NAIST	0.49	0.49	0.49	0.53	0.53	0.53
USFD*	0.54	0.54	0.54	0.57	0.57	0.57
WVALI	0.54	0.54	0.54	0.64	0.64	0.64
XRCE-T	0.42	0.42	0.42	0.58	0.58	0.58
average	0.51	0.51	0.51	0.58	0.58	0.58
stddev	0.05	0.05	0.05	0.04	0.04	0.04

Table 4: Results for Task C

All tables give precision, recall and f-measure for both the strict and the relaxed scoring scheme, as well as averages and standard deviation on the precision, recall and f-measure numbers. The entry for USFD is starred because the system developers are co-organizers of the TempEval task.[3]

For task A, the f-measure scores range from 0.34 to 0.62 for the strict scheme and from 0.41 to 0.63 for the relaxed scheme. For task B, the scores range from 0.66 to 0.80 (strict) and 0.71 to 0.81 (relaxed). Finally, task C scores range from 0.42 to 0.55 (strict) and from 0.56 to 0.66 (relaxed).

The differences between the systems is not spectacular. WVALI's hybrid approach outperforms the other systems in task B and, using relaxed scoring, in task C as well. But for task A, the winners barely edge out the rest of the field. Similarly, for task C using strict scoring, there is no system that clearly separates itself from the field.

It should be noted that for task A, and in lesser extent for task B, the XRCE-T system has recall scores that are far below all other systems. This seems mostly due to a choice by the developers to not assign a temporal relation if the syntactic analyzer did not find a clear syntactic relation between the two elements that needed to be linked for the TempEval task.

[3]There was a strict separation between people assisting in the annotation of the evaluation corpus and people involved in system development.

7 Conclusion: the Future of Temporal Evaluation

The evaluation approach of TempEval avoids the interdependencies that are inherent to a network of temporal relations, where relations in one part of the network may constrain relations in any other part of the network. To accomplish that, TempEval deliberately focused on subtasks of the larger problem of automatic temporal annotation.

One thing we may want to change to the present TempEval is the definition of task A. Currently, it instructs to temporally link all events in a sentence to all time expressions in the same sentence. In the future we may consider splitting this into two tasks, where one subtask focuses on those anchorings that are very local, like *"...White House spokesman Marlin Fitzwater [said] [late yesterday] that..."*. We expect both inter-annotator agreement and system performance to be higher on this subtask.

There are two research avenues that loom beyond the current TempEval: (1) definition of other subtasks with the ultimate goal of establishing a hierarchy of subtasks ranked on performance of automatic taggers, and (2) an approach to evaluate entire timelines.

Some other temporal linking tasks that can be considered are ordering of consecutive events in a sentence, ordering of events that occur in syntactic subordination relations, ordering events in coordinations, and temporal linking of reporting events to the document creation time. Once enough temporal links from all these subtasks are added to the entire temporal graph, it becomes possible to let confidence scores from the separate subtasks drive a constraint propagation algorithm as proposed in (Allen, 1983), in effect using high-precision relations to constrain lower-precision relations elsewhere in the graph.

With this more complete temporal annotation it is no longer possible to simply evaluate the entire graph by scoring pairwise comparisons. Instead the entire timeline must be evaluated. Initial ideas regarding this focus on transforming the temporal graph of a document into a set of partial orders built

around precedence and inclusion relations and then evaluating each of these partial orders using some kind of edit distance measure.[4]

We hope to have taken the first baby steps with the three TempEval tasks.

8 Acknowledgements

We would like to thank all the people who helped prepare the data for TempEval, listed here in no particular order: Amber Stubbs, Jessica Littman, Hongyuan Qiu, Emin Mimaroglu, Emma Barker, Catherine Havasi, Yonit Boussany, Roser Saurí, and Anna Rumshisky.

Thanks also to all participants to this new task: Steven Bethard and James Martin (University of Colorado at Boulder), Congmin Min, Munirathnam Srikanth and Abraham Fowler (Language Computer Corporation), Yuchang Cheng, Masayuki Asahara and Yuji Matsumoto (Nara Institute of Science and Technology), Mark Hepple, Andrea Setzer and Rob Gaizauskas (University of Sheffield), Caroline Hagegè and Xavier Tannier (XEROX Research Centre Europe), and Georgiana Puşcaşu (University of Wolverhampton and University of Alicante).

Part of the work in this paper was funded by the DTO/AQUAINT program under grant number N61339-06-C-0140 and part funded by the EU VIKEF project (IST- 2002-507173).

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. forthcoming. Timebank evolution as a community resource for timeml parsing. *Language Resources and Evaluation*.

Inderjeet Mani, Ben Wellner, Marc Verhagen, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia. ACL.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, January.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, March.

[4] Edit distance was proposed by Ben Wellner as a way to evaluate partial orders of precedence relations (personal communication).