Error Analysis of the TempEval Temporal Relation Identification Task

Chong Min Lee
Linguistics Department
Georgetown University
Washington, DC 20057, USA
cml54@georgetown.edu

Graham Katz
Linguistics Department
Georgetown University
Washington, DC 20057, USA
egk7@georgetown.edu

Abstract

The task to classify a temporal relation between temporal entities has proven to be difficult with unsatisfactory results of previous research. In TempEval07 that was a first attempt to standardize the task, six teams competed with each other for three simple relation-identification tasks and the results were comparably poor. In this paper we provide an analysis of the TempEval07 competition results, identifying aspects of the tasks which presented the systems with particular challenges and those that were accomplished with relative ease.

1 Introduction

The automatic temporal interpretation of a text has long been an important area of computational linguistics research (Bennett and Partee, 1972; Kamp and Reyle, 1993). In recent years, with the advent of the TimeML markup language (Pustejovsky et al., 2003) and the creation of the TimeBank resource (Pustejovsky et al., 2003) interest has focused on the application of a variety of automatic techniques to this task (Boguraev and Ando, 2005; Mani et al., 2006; Bramsen et al., 2006; Chambers et al., 2007; Lee and Katz, 2008). The task of identifying the events and times described in a text and classifying the relations that hold among them has proven to be difficult, however, with reported results for relation classification tasks ranging in F-score from 0.52 to 0.60.

Variation in the specifics has made comparison among research methods difficult, however. A first attempt to standardize this task was the 2007 TempEval competition (Verhagen et al., 2007). This competition provided a standardized training and evaluation scheme for automatic temporal interpretation systems. Systems were pitted against one another on three simple relation-identification tasks. The competing systems made use of a variety of techniques but their results were comparable, but poor, with average system performance on the tasks ranging in F-score from 0.74 on the easiest task to 0.51 on the most difficult. In this paper we provide an analysis of the TempEval07 competition, identifying aspects of the tasks which presented the systems with particular challenges and those that were accomplished with relative ease.

2 TempEval

The TempEval competition consisted of three tasks, each attempting to model an important subpart of the task of general temporal interpretation of texts. Each of these tasks involved identifying in running text the temporal relationships that hold among events and times referred to in the text.

- **Task A** was to identify the temporal relation holding between an event expression and a temporal expression occurring in the same sentence.
- **Task B** was to identify the temporal relations holding between an event expression and the Document Creation Time (DCT) for the text.
- **Task C** was to identify which temporal relation held between main events of described by sen-
For the competition, training and development data—newswire files from the TimeBank corpus (Pustejovsky et al., 2003)—was made available in which the events and temporal expressions of interest were identified, and the gold-standard temporal relation was specified (a simplified set of temporal relations was used: BEFORE, AFTER, OVERLAP, OVERLAP-OR-BEFORE, AFTER-OR-OVERLAP and VAGUE.). For evaluation, a set of newswire texts was provided in which the event and temporal expressions to be related were identified (with full and annotated in TimeML markup) but the temporal relations holding among them withheld. The task in was to identify these relations.

The text below allows illustrates the features of the TimeML markup that were made available as part of the training texts and which will serve as the basis for our analysis below:

```
<TIMEX3 tid="t13" type="DATE" value="1989-11-02" temporalFunction="false" functionInDocument="CREATION_TIME">11/02/89</TIMEX3> <s> Italian chemical giant Montedison S.p.A. <TIMEX3 tid="t19" type="DATE" value="1989-11-01" temporalFunction="true" functionInDocument="NONE" anchorTimeID="t13">yesterday</TIMEX3><EVENT eid="e2" class="OCCURRENCE" stem="offer" aspect="NONE" tense="PAST" polarity="POS" pos="NOUN">offered</EVENT> $37-a-share for all the common shares outstanding of Erbamont N.V.</s>
<pre>Montedison <TIMEX3 tid="t17" type="DATE" value="PRESENT_REF" temporalFunction="true" functionInDocument="NONE" anchorTimeID="t13">currently</TIMEX3><EVENT eid="e20" class="STATE" stem="own" aspect="NONE" tense="PRESENT" polarity="POS" pos="VERB">owns</EVENT> about 72% of Erbamont’s common shares outstanding.</pre>
```

TimeML annotation associates with temporal expression and event expression identifiers (tid and eid, respectively). Task A was to identify the temporal relationships holding between time t19 and event e2 and between t17 and e20 (OVERLAP was the gold-standard answer for both). Task B was to identify the relationship between the events and the document creation time t13 (BEFORE for e2 and OVERLAP for e20). Task C was to identify the relationship between e2 and e20 (OVERLAP-OR-BEFORE). The TempEval07 training data consisted of a total of 162 document. This amounted to a total of 1490 total relations for Task A, 2556 for task B, and 1744 for Task C. The 20 documents of testing data had 169 Task A relations, 337 Task B relations, and 258 Task C relations. The distribution of items by relation type in the training and test data is given in Table 1.

Six teams participated in the TempEval competition. They made use of a variety of techniques, from the application of off-the-shelf machine learning tools to “deep” NLP. As indicated in Table 2, while the tasks varied in difficulty, within each task the results of the teams were, for the most part, comparable.

The systems (other than XRCE-T) did somewhat to quite a bit better than baseline on the tasks. Our focus here is on identifying features of the task that gave rise to difficult, using overall performance of the different systems as a metric. Of the 764 test items, a large portion were either ‘easy’—meaning that all the systems provided correct output—or ‘hard’—meaning none did.

|                | Task A | Task B | Task C |
|----------------|--------|--------|--------|
| CU-TMP         | 60.9   | 75.2   | 53.5   |
| LCC-TE         | 57.4   | 71.3   | 54.7   |
| NAIST          | 60.9   | 74.9   | 49.2   |
| TimeBandits    | 58.6   | 72.5   | 54.3   |
| WVALI          | 61.5   | 79.5   | 53.9   |
| XRCE-T         | 24.9   | 57.4   | 42.2   |
| average        | 54.0   | 71.8   | 51.3   |

Table 2: TempEval Accuracy (%)

In task A, the cases (24/14%) that all participants make correct prediction are when the target relation is overlap. And, the part-of-speeches of most events

---

1This contrasts with the 13 temporal relations supported by TimeML. The full TimeML markup of event and temporal expressions was maintained.

2TempEval was scored in a number of ways; we report accuracy of relation identification here as we will use this measure, and ones related to it below

3The XRCE-T team, which made use of the deep analysis engine XIP lightly modified for the competition, was a clear outlier.
in the cases are verbs (19 cases), and their tenses are past (13 cases). In task B, among 160 cases for that every participant predicts correct temporal relation, 159 cases are verbs, 122 cases have before as target relation, and 112 cases are simple past tenses. In task C, we find that 22 cases among 35 cases are reporting:reporting with overlap as target relation. In what follows we will identify aspects of the tasks that make some items difficult and some not so much so.

### 3 Analysis

In order to make fine-grained distinctions and to compare arbitrary classes of items, our analysis will be stated in terms of a summary statistic: the success measure (SM).

1. **Success measure**

   \[
   \sum_{k=0}^{6} \frac{kC_k}{6! \sum_{k=0}^{6} kC_k}
   \]

   where \( C_k \) is the number of items \( k \) systems got correct. This simply the proportion of total correct responses to items in a class (for all systems) divided by the total number of items in that class (a success measure of 1.0 is easy and of 0.0 is hard). For example, let’s suppose before relation have 10 instances. Among the instances, three cases are correct by all teams, four by three teams, two by two teams, and one by no teams. Then, SM of before relation is 0.567 (\((3\times6)+(4\times3)+(2\times2)+(1\times0)\)/(6\times(1+2+4+3))).

   In addition, we would like to keep track of how important each class of errors is to the total evaluation. To indicate this, we compute the error proportion (ER) for each class: the proportion of total errors attributable to that class.

2. **Error proportion**

   \[
   \sum_{k=0}^{6} \frac{(6-k)C_k}{AllErrorsInTask \times NumberOfTeams}
   \]

   Table 3 provides the overall analysis by relation type. This shows that (as might be expected) the systems did best on the relations that were the majority class for each task: overlap in Task A, before in Task B, and overlap in Task C.

   Furthermore systems do poorly on all of the disjunctive classes, with this accounting for between 1% and 9% of the task error. In what follows we will ignore the disjunctive relations. Performance on the before relation is low for Task A but very good for Task B and moderate for Task C. For more detailed analysis we treat each task separately.

#### 3.1 Overall analysis

When a case shows high SM and high ER, we can guess that the case has lots of instances. With low SM and low ER, it says there is little instances. With high SM and low ER, we don’t need to focus on the case because the case show very good performance. Of particular interest are classes in which the SM is low and the ER is high because it has a room for the improvement.

#### 3.2 Task A

For Task A we analyze the results with respect to the attribute information of the EVENT and TIMEX3 TimeML tags. These are the event class (aspectual, i_action, i_state, occurrence, perception, reporting, and state)\textsuperscript{4} part-of-speech (basically noun and verb),

\textsuperscript{4}The detailed explanations on the event classes can be found in the TimeML annotation guideline at
Table 4: POS of EVENT in Task A

|          | NOUN | VERB |
|----------|------|------|
| BEFORE   | 0/5% | 0.324/15% |
| AFTER    | 0.119/8% | 0.507/15% |
| OVERLAP  | 0.771/7% | 0.747/24% |
| VAGUE    | 0/8% | 0/10% |

and tense&aspect marking for event expressions. Information about the temporal expression turned out not to be a relevant dimension of analysis.

As we seen in Table 4, verbal event expressions make for easier classification for before and after (there is a 75%/25% verb/noun split in the data). When the target relation is overlap, nouns and verbs have similar SMs.

One reason for this difference, of course, is that verbal event expressions have tense and aspect marking (the tense and aspect marking for nouns is simply none).

In Table 5 we show the detailed error analysis with respect to tense and aspect values of the event expression. The combination of tense and aspect values of verbs generates 10 possible values: future, infinitive, past, past-perfective, past-progressive (pastprog), past-participle (pastpart), present, present-perfective (presperf), present-progressive (presprog), and present-participle (prespart). Among them, only five cases (infinitive, past, present, presperf, and prespart) have more than 2 examples in test data. Past takes the biggest portions (40%) in test data and in errors (33%). Overlap seems less influenced with the values of tense and aspect than before and after when the five cases are considered. Before and after show 0.444 and 0.278 differences between infinitive and present and between infinitive and present. But, overlap scores 0.136 differences between present and past. And a problem case is before with past tense that shows 0.317 SM and 9% EP.

When we consider simultaneously SM and EP of the semantic class of events in Table 6, we can find three noticeable cases: occurrence and reporting of before, and occurrence of after. All of them have over 5% EP and under 0.4 SM. In case of reporting of after, its SM is over 0.5 but its EP shows some room for the improvement.

Table 5: Tense & Aspect of EVENT in Task A

|          | BEFORE | AFTER | OVERLAP | VAGUE |
|----------|--------|-------|---------|-------|
| FUTURE   | 0/0%   | 0.333/1% | 0.833/0% | 0/0%  |
| INFINITIVE | 0.333/3% | 0.667/2% | 0.119/8% | 0/1% |
| NONE     | 0/5%   | 0.507/15% | 0.765/7% | 0/8%  |
| PAST     | 0.317/9% | 0.544/9% | 0.782/10% | 0/5% |
| PASTPERF | 0/0%   | 0.333/1% | 0.833/0% | 0/0%  |
| PASTPROG | 0/0%   | 0.500/1% | 0.500/1% | 0/0%  |
| PRESENT  | 0.444/2% | 0.611/2% | 0.646/4% | 0/1%  |
| PRESPERF | 0.333/0% | 0/0%   | 0.690/3% | 0/0%  |
| PRESPROG | 0/0%   | 0/0%   | 0.833/0% | 0/0%  |
| PRESPART | 0/0%   | 0/0%   | 0.774/4% | 0/1%  |

Table 7: Distance in Task A

Boguraev and Ando (2005) report a slight increase in performance in relation identification based on proximity of the event expression to the temporal expression. We investigated this in Table 7, looking at the distance in word tokens.

We can see noticeable cases in before and after of ≤ 16 row. Both cases show over 13% EP and under 0.5 SM. The participants show good SM in overlap of ≤ 4. Overlap of ≤ 16 has the biggest EP (17%). When its less satisfactory SM (0.654) is considered, it seems to have a room for the improvement. One of the cases that have 13% EP is vague of ≥ 16. It says that it is difficult even for humans to make a decision on a temporal relation when the distance between an event and a temporal expression is greater than and equal to 16 words.

3.3 Task B

Task B is to identify a temporal relation between an EVENT and DCT. We analyze the participants performance with part-of-speech. This analysis shows how poor the participants are on after and overlap of nouns (0.167 and 0.115 SM). And the EM of overlap of verbs (26%) shows that the improvement is needed on it.

In test data, occurrence and reporting have similar number of examples: 135 (41%) and 106 (32%) in 330 examples. In spite of the similar distribution, their error rates show difference. It suggests that reporting is easier than occurrence. Moreover,
Table 6: EVENT Class in Task A

|        | ASPECTUAL | I_ACTION | I_STATE | OCCURRENCE | PERCEPTION | REPORTING | STATE |
|--------|-----------|----------|---------|------------|------------|-----------|-------|
| BEFORE | 0.167/1%  | 0/0%     | 0.333/5%| 0.067/6%   | 0/0%       | 0.364/9%  | 0/1%  |
| AFTER  | 0.111/3%  | 0/0%     | 0.000/0%| 0.317/9%   | 0/0%       | 0.578/8%  | 0.167/2%|
| OVERLAP| 0.917/0%  | 0.778/1% | 0.583/5%| 0.787/15%  | 0.750/1%   | 0.667/9%  | 0.815/2%|
| VAGUE  | 0/1%      | 0/1%     | 0/0%    | 0/0%       | 0/0%       | 0/0%      | 0/0%  |

Table 7: EVENT Class in Task B

|        | NOUN | VERB |
|--------|------|------|
| BEFORE | 0.735/6% | 0.908/16% |
| AFTER  | 0.167/8% | 0.667/14% |
| OVERLAP| 0.115/13% | 0.645/26% |
| VAGUE  | 0/4% | 0/1% |

Table 9: EVENT Class in Task B

Table 8: POS of EVENT in Task B

Table 9 shows most errors in after occur with occurrence class 65% (15%/23%) when we consider 23% EP in Table 3. Occurrence and reporting of before show noticeably good performance (0.818 and 0.949). And occurrence of overlap has the biggest error rate (17%) with 0.367 of SM.

In case of state, it has 22 examples (7%) but takes 10% of errors. And it is interesting that the most errors are concentrated in state. In our intuition, it is not a difficult task to identify overlap relation of state class.

Table 9 does not clearly show what causes the poor performance of nouns in after and overlap. In the additional analysis of nouns with class information, occurrence shows poor performance in after and overlap: 0.111/6% and 0.083/8%. And other noticeable case in nouns is state of overlap: 0.125/4%. We can see the low performance of nouns in overlap is due to the poor performance of state and occurrence, but only occurrence is a cause of the poor performance in after.

DCT can be considered as speech time. Then, tense and aspect of verb events can be a cue in predicting temporal relations between verb events and DCT. The better performance of the participants in verbs can be an indirect evidence. The analysis with tense & aspect can tell us which tense & aspect information is more useful. A problem with the information is sparsity. Most cases appear less than 3 times. The cases that have more than or equal to three instances are 13 cases among the possible combinations of 7 tenses and 4 aspects in TimeML. Moreover, only two cases are over 5% of the whole data: past with before (45%) and present with overlap (15%). In Table 10, tense and aspect information seems valuable in judging a relation between a verb event and DCT. The participants show good performances in the cases that seem easy intuitively: past with before, future with after, and present with overlap. Among intuitively obvious cases that are past, present, or future tense, present tense makes large errors (20% of verb errors). And present shows 7% EP in before.

When events has no cue to infer a relation like infinitive, none, pastpart, and prespart, their SMs are lower than 0.500 except infinitive and none of after. infinitive of overlap shows poor performance with the biggest error rate (0.125/12%).

3.4 Task C

The task is to identify the relation between consecutive main events. There are four part-of-speeches in Task C: adjective, noun, other, and verb. Among eight possible pairs of part-of-speeches, only three pairs have over 1% in 258 TLINKs: noun and verb (4%), verb and noun (4%), and verb and verb (85%). When we see the distribution of verb and verb by three relations (before, after, and overlap), the relations show 19%, 14%, and 41% distribution each. In Table 11, the best SM is verb:verb of overlap (0.690). And verb:verb shows around 0.5 SM in before and after.

Tense & aspect pairs of main event pairs show
skewed distribution, too. The cases that have over 1% data are eight: past: none, past: past, past: present, present: past, present: present, presperf: present, and presperf: presperf. Among them, past tense pairs show the biggest portion (40%). The performance of the eight cases is reported in Table 12. As we can guess with the distribution of tense & aspect, most errors are from past: past (40%). When the target relation of past: past is overlap, the participants show reasonable SM (0.723). But, their performances are unsatisfactory in before and after.

When we consider cases over 1% of test data in main event class pairs, we can see eleven cases as Table 13. Among the eleven cases, four pairs have over 5% data: occurrence: occurrence (13%), occurrence: reporting (14%), reporting: occurrence (9%), and reporting: reporting (17%). Reporting: reporting shows the best performance (0.934/2%) in overlap. Two class pairs have over 10% EP: occurrence: occurrence (15%), and occurrence: reporting (14%). In addition, occurrence pairs seem difficult tasks when target relations are before and after because they show low SMs (0.317 and 0.200) with 5% and 3% error rates.

4 Discussion and Conclusion

Our analysis shows that the participants have the difficulty in predicting a relation of a noun event when its target relation is before and after in Task A, and after and overlap in Task B. When the distance is in the range from 5 to 16 in Task A, more effort seems to be needed.

In Task B, tense and aspect information seems valuable. Six teams show good performance when simple tenses such as past, present, and future appear with intuitively relevant target relations such as before, overlap, and after. Their poor performance with none and infinitive tenses, and nouns can be another indirect evidence.

A difficulty in analyzing Task C is sparsity. So, this analysis is focused on verb: verb pair. When we see in (12), past pairs still show the margin for the improvement. But, a lot of reporting events are used as main events. When we consider that important events in newspaper are cited, the current TempEval task can miss useful information.

Six participants make very little correct predictions on before-or-overlap, overlap-or-after, and vague. A reason on the poor prediction can be small distribution in the training data as we can see in Table 1. Data sparsity problem is a bottleneck in natural language processing. The addition of the disjunctive relations and vague to the target labels can make the sparsity problem worse. When we consider the participants’ poor performance on the labels, we suggest to use three labels (before, overlap, and after) as the target labels.

Table 10: Tense & Aspect of EVENT in Task B

|              | BEFORE | AFTER | OVERLAP | VAGUE |
|--------------|--------|-------|---------|-------|
| FUTURE       | 0/0%   | 0.963/1% | 0.333/2% | 0/0% |
| FUTURE-PROGRESSIVE | 0/0% | 0/0% | 0.167/1% | 0/0% |
| INFINITIVE    | 0.367/5% | 0.621/7% | 0.125/12% | 0/2% |
| NONE          | 0/0%   | 0.653/7% | 0/2% | 0/0% |
| PAST          | 0.984/1% | 0.333/1% | 0.083/2% | 0/0% |
| PASTPERF      | 1.000/0% | 0/0% | 0/0% | 0/0% |
| PASTPROG      | 1.000/0% | 0/0% | 0/0% | 0/0% |
| PASTPART      | 0.583/1% | 0/0% | 0/0% | 0/0% |
| PRESENT       | 0.429/7% | 0.167/3% | 0.850/10% | 0/0% |
| PRESERP       | 0.861/3% | 0/0% | 0/2% | 0/0% |
| PRESENT-PROGRESSIVE | 0/0% | 0/0% | 0.967/0% | 0/0% |
| PRESPART      | 0/0% | 0.444/3% | 0.310/8% | 0/0% |

Table 11: POS pairs in Task C

|             | BEFORE | AFTER | OVERLAP | VAGUE |
|-------------|--------|-------|---------|-------|
| NOUN:VERB   | 0.250/2% | 0/0% | 0.625/1% | 0/0% |
| VERB:NOUN   | 0.583/1% | 0.500/2% | 0.333/1% | 0/1% |
| VERB:VERB   | 0.500/20% | 0.491/15% | 0.690/26% | 0.220/12% |
Our analysis can be used as a cue in adding an additional module for weak points. When a pair of a noun event and a temporal expression appears in a sentence, a module can be added based on our study.

## References

Branimir Boguraev and Rie Kubota Ando. 2005. TimeML-Compliant Text Analysis for Temporal Reasoning. *Proceedings of IJCAI-05*, 997–1003.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth Sundheim. 2003. The TIMEBANK corpus. *Proceedings of Corpus Linguistics 2003*, 647–656.

Michael Bennett and Barbara Partee. 1972. Toward the logic of tense and aspect in English. *Technical report, System Development Corporation*. Santa Monica, CA

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilai. 2006. Inducing Temporal Graphs *Proceedings of EMNLP 2006*, 189–198.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying Temporal Relations Between Events *Proceedings of ACL 2007*, 173–176.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine Learning of Temporal Relations. *Proceedings of ACL-2006*, 753–760.

Chong Min Lee and Graham Katz. 2008. Toward an Automated Time-Event Anchoring System. *The Fifth Midwest Computational Linguistics Colloquium*.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to modeltheoretic semantics of natural language*. Kluwer Academic, Boston.

James Pustejovsky, José Castaño, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *IWCS-5, Fifth International Workshop on Computational Semantics*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, James Pustejovsky. 2007. *SemEval-2007 Task 15: TempEval Temporal Relation Identification*. *Proceedings of SemEval-2007*, 75–80.

Caroline Hagège and Xavier Tannier. 2007 *XRCE-T: XIP Temporal Module for TempEval campaign*. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 492–495.

Steven Bethard and James H. Martin. 2007. *CU-TMP: Temporal Relation Classification Using Syntactic and Semantic Features*. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 129–132.

Congmin Min, Munirathnam Srikanth, and Abraham Fowler. 2007. *LCC-TE: A Hybrid Approach to Temporal Relation Identification in News Text*. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 219–222.

Yuchang Cheng, Masayuki Asahara, and Yuji Matsumoto. 2007. *NAIST.Japan: Temporal Relation...*
Identification Using Dependency Parsed Tree. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 245–248.

Georgiana Pușcașu. 2007. WVALI: Temporal Relation Identification by Syntactico-Semantic Analysis Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 484–487.

Mark Hepple, Andrea Setzer, and Robert Gaizauskas. 2007. USFD: Preliminary Exploration of Features and Classifiers for the TempEval-2007 Task. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 438–441.