CU-TMP:
Temporal Relation Classification Using Syntactic and Semantic Features

Steven Bethard and **James H. Martin**
Department of Computer Science
University of Colorado at Boulder
430 UCB, Boulder, CO 80309, USA
{bethard,martin}@colorado.edu

Abstract

We approached the temporal relation identification tasks of TempEval 2007 as pair-wise classification tasks. We introduced a variety of syntactically and semantically motivated features, including temporal-logic-based features derived from running our Task B system on the Task A and C data. We trained support vector machine models and achieved the second highest accuracies on the tasks: 61% on Task A, 75% on Task B and 54% on Task C.

1 Introduction

In recent years, the temporal structure of text has become a popular area of natural language processing research. Consider a sentence like:

(1) The top commander of a Cambodian resistance force said Thursday he has sent a team to recover the remains of a British mine removal expert kidnapped and presumed killed by Khmer Rouge guerrillas almost two years ago.

English speakers immediately recognize that *kidnapping* came first, then *sending*, and finally *saying*, even though *before* and *after* never appeared in the text. How can machines learn to do the same?

The 2007 TempEval competition tries to address this question by establishing a common corpus on which research systems can compete to find temporal relations (Verhagen et al., 2007). TempEval considers the following types of event-time temporal relations:

Task A Events[1]and times within the same sentence
Task B Events[1] and document times
Task C Matrix verb events in adjacent sentences

In each of these tasks, systems attempt to annotate pairs with one of the following relations: BEFORE, BEFORE-OR-OVERLAP, OVERLAP, OVERLAP-OF-AFTER, AFTER or VAGUE. Competing systems are instructed to find all temporal relations of these types in a corpus of newswire documents.

We approach these tasks as pair-wise classification problems, where each event/time pair is assigned one of the TempEval relation classes (BEFORE, AFTER, etc.). Event/time pairs are encoded using syntactically and semantically motivated features, and then used to train support vector machine (SVM) classifiers.

The remainder of this paper is structured as follows. Section 2 describes the features used to characterize event/time relations. Section 3 explains how we used these features to train SVM models for each task. Section 4 discusses the performance of our models on the TempEval data, and Section 5 summarizes the lessons learned and future directions.

2 Features

We used a variety of lexical, syntactic and semantic features to characterize the different types of temporal relations. In each task, the events and times were characterized using the features:

word The text of the event or time words

[1] TempEval only considers events that occurred at least 20 times in the TimeBank (Pustejovsky et al., 2003) corpus for these tasks

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 129–132,
Prague, June 2007. ©2007 Association for Computational Linguistics

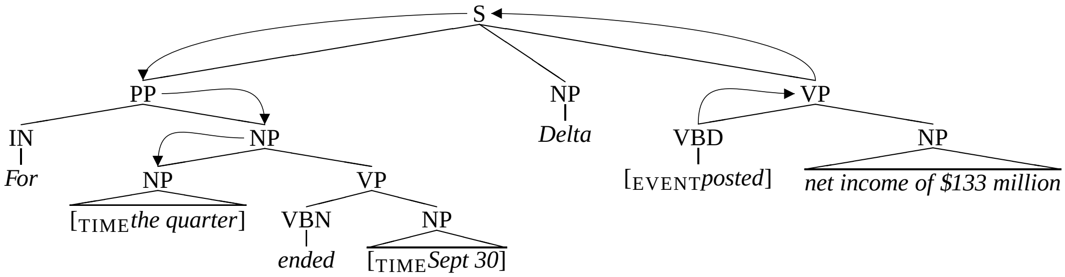

Figure 1: A syntactic tree. The path between *posted* and *the quarter* is VBD-VP-S-PP-NP-NP

pos The parts of speech[2] of the words, e.g. *this crucial moment* has the parts of speech DT-JJ-NN.

gov-prep Any prepositions governing the event or time, e.g. in *during the Iran-Iraq war*, the preposition *during* governs the event *war*, and in *after ten years*, the preposition *after* governs the time *ten years*.

gov-verb The verb that governs the event or time, e.g. in *rejected in peace talks*, the verb *rejected* governs the event *talks*, and in *withdrawing on Friday*, the verb *withdrawing* governs the time *Friday*. For events that are verbs, this feature is just the event itself.

gov-verb-pos The part of speech[2] of the governing verb, e.g. *withdrawing* has the part of speech VBG.

aux Any auxiliary verbs and adverbs modifying the governing verb, e.g. in *could not come*, the words *could* and *not* are considered auxiliaries for the event *come*, and in *will begin withdrawing on Friday*, the words *will* and *begin* are considered auxiliaries for the time *Friday*.

Events were further characterized using the features (the last six use gold-standard TempEval markup):

modal Whether or not the event has one of the auxiliaries, *can*, *will*, *shall*, *may*, or any of their variants (*could*, *would*, etc.).

gold-stem The stem, e.g. the stem of *fallen* is *fall*.

gold-pos The part-of-speech, e.g. NOUN or VERB.

gold-class The semantic class, e.g. REPORTING.

gold-tense The tense, e.g. PAST or PRESENT.

gold-aspect The aspect, e.g. PERFECTIVE.

gold-polarity The polarity, e.g. POS or NEG.

Times were further characterized using the following gold-standard TempEval features:

gold-type The type, e.g. DATE or TIME.

gold-value The value, e.g. PAST_REF or 1990-09.

gold-func The temporal function, e.g. TRUE.

These gold-standard event and time features are similar to those used by Mani and colleagues (2006).

The features above don't capture much of the differences between the tasks, so we introduced some task-specific features. Task A included the features:

inter-time The count of time expressions between the event and time, e.g. in Figure 1, there is one time expression, *Sept 30*, between the event *posted* and the time *the quarter*.

inter-path The syntactic path between the event and the time, e.g. in Figure 1 the path between *posted* and *the quarter* is VBD>VP>S<PP<NP<NP.

inter-path-parts The path, broken into three parts: the tags from the event to the lowest common ancestor (LCA), the LCA, and the tags from the LCA to the time, e.g. in Figure 1 the parts are VBD>VP, S and PP<NP<NP.

inter-clause The number of clause nodes along the syntactic path, e.g. in Figure 1 there is one clause node along the path, the top S node.

Our syntactic features were derived from a syntactic tree, though Boguraev and Ando (2005) suggest that some could be derived from finite state grammars.

For Task C we included the following feature:

tense-rules The relation predicted by a set of tense rules, where past tense events come BEFORE present tense events, present tense events come BEFORE future tense events, etc. In the text:

(2) Finally today, we [$_{\text{EVENT}}$ *learned*] that the space agency has taken a giant leap forward. Collins will be [$_{\text{EVENT}}$ *named*] commander of Space Shuttle Columbia.

[2] From MXPOST (ftp.cis.upenn.edu/pub/adwait/jmx/)

Since *learned* is in past tense and *named* is in future, the relation is (*learned* BEFORE *named*).

In preliminary experiments, the Task B system had the best performance, so we ran this system on the data for Tasks A and C, and used the output to add the following feature for both tasks:

task-b-rel The relation predicted by combining the output of the Task B system with temporal logic. For example, consider the text:

(3) [$_{\text{TIME}}$ *08-15-90 (=1990-08-15)*]
Iraq's Saddam Hussein
[$_{\text{TIME}}$ *today (=1990-08-15)*] sought
peace on another front by promising to
release soldiers captured during the
Iran-Iraq [$_{\text{EVENT}}$ *war*].

If Task B said (*war* BEFORE *08−15−90*) then since *08−15−90=1990−08−15=today*, the relation (*war* BEFORE *today*) must hold.

3 Models

Using the features described in the previous section, each temporal relation — an event paired with a time or another event — was translated into a set of feature values. Pairing those feature values with the TempEval labels (BEFORE, AFTER, etc.) we trained a statistical classifier for each task. We chose support vector machines[3](SVMs) for our classifiers as they have shown good performance on a variety of natural language processing tasks (Kudo and Matsumoto, 2001; Pradhan et al., 2005).

Using cross-validations on the training data, we performed a simple feature selection where any feature whose removal improved the cross-validation F-score was discarded. The resulting features for each task are listed in Table 1. After feature selection, we set the SVM free parameters, e.g. the kernel degree and cost of misclassification, by performing additional cross-validations on the training data, and selecting the model parameters which yielded the highest F-score for each task[4].

Feature	Task A	Task B	Task C
event-word			
event-pos	X		X
event-gov-prep	X		X
event-gov-verb	X		X
event-gov-verb-pos	X	X	2
event-aux	X	X	X
modal	X		X
gold-stem	X	X	1
gold-pos	X		X
gold-class	X	X	X
gold-tense	X	X	X
gold-aspect	X		X
gold-polarity	X		X
time-word	X		
time-pos	X		
time-gov-prep	X		
time-gov-verb	X		
time-gov-verb-pos	X		
time-aux	X		
gold-type			
gold-value	X	X	
gold-func	X		
inter-time	X		
inter-path	X		
inter-path-parts	X		
inter-clause	X		
tense-rules			X
task-b-rel	X		X

Table 1: Features used in each task. An X indicates that the feature was used for that task. For Task C, 1 indicates that the feature was used only for the first event and not the second, and 2 indicates the reverse.

	Strict			Relaxed		
Task	P	R	F	P	R	F
A	0.61	0.61	0.61	0.63	0.63	0.63
B	0.75	0.75	0.75	0.76	0.76	0.76
C	0.54	0.54	0.54	0.60	0.60	0.60

Table 2: (P)recision, (R)ecall and (F)-measure of the models on each task. Precision, recall and F-measure are all equivalent to classification accuracy.

4 Results

We evaluated our classifers on the TempEval test data. Because the Task A and C models derived features from the Task B temporal relations, we first ran the Task B classifer over all the data, and then ran the Task A and Task C classifiers over their individual data. The resulting temporal relation classifications were evalutated using the standard TempEval scoring script. Table 2 summarizes these results.

Our models achieved an accuracy of 61% on Task A, 75% on Task B and 54% on Task C, the second highest scores on all these tasks. The Temp-

[3] We used the TinySVM implementation from http://chasen.org/%7Etaku/software/TinySVM/ and trained one-vs-rest classifiers.

[4] We only experimented with polynomial kernels.

Task	Feature Removed	Model Accuracy
A	-	0.663
	time-gov-prep	0.650
	gold-value	0.652
	polarity	0.655
	task-b-rel	0.656
B	-	0.809
	event-aux	0.780
	gold-stem	0.784
	gold-class	0.794
C	-	0.534
	event-gov-verb-2	0.522
	event-aux-2	0.525
	gold-class-1	0.526
	gold-class-2	0.527
	event-pos-2, task-b-rel	0.529

Table 3: Feature analysis. The '-' lines show the accuracy of the model with all features.

Eval scoring script also reported a relaxed measure where for example, systems could get partial credit for matching a gold standard label like OVERLAP-OR-AFTER with OVERLAP or AFTER. Under this measure, our models achieved an accuracy of 63% on Task A, 76% on Task B and 60% on Task C, again the second highest scores in the competition.

We performed a basic feature analysis where, for each feature in a task, a model was trained with that feature removed and all other features retained. We evaluated the performance of the resulting models using cross-validations on the training data[5]. Features whose removal resulted in the largest drops in model performance are listed in Table 3.

For Task A, the most important features were the preposition governing the time and the time's normalized value. For Task B, the most important features were the auxiliaries governing the event, and the event's stem. For Task C, the most important features were the verb and auxiliaries governing the second event. For both Tasks A and C, the features based on the Task B relations were one of the top six features. In general however, no single feature dominated any one task — the greatest drop in performance from removing a feature was only 2.9%.

5 Conclusions

TempEval 2007 introduced a common dataset for work on identifying temporal relations. We framed the TempEval tasks as pair-wise classification problems where pairs of events and times were assigned a temporal relation class. We introduced a variety of syntactic and semantic features, including paths between constituents in a syntactic tree, and temporal relations deduced by running our Task B system on the Task A and C data. Our models achieved an accuracy of 61% on Task A, 75% on Task B and 54% on Task C. Analysis of these models indicated that no single feature dominated any given task, and suggested that future work should focus on new features to better characterize temporal relations.

6 Acknowledgments

This research was performed under an appointment of the first author to the DHS Scholarship and Fellowship Program, administered by the ORISE through an interagency agreement between DOE and DHS. ORISE is managed by ORAU under DOE contract number DE-AC05-06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.

References

B. Boguraev and R. K. Ando. 2005. Timebank-driven timeml analysis. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, Dagstuhl Seminars. German Research Foundation.

T. Kudo and Y. Matsumoto. 2001. Chunking with support vector machines. In *NAACL*.

I. Mani, M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky. 2006. Machine learning of temporal relations. In *COLING/ACL*.

S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

J. Pustejovsky, P. Hanks, R. Saur, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The timebank corpus. In *Corpus Linguistics*, pages 647–656.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, and J. Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *SemEval-2007: 4th International Workshop on Semantic Evaluations*.

[5] We used cross-validations on the training data to preserve the validity of the TempEval test data for future research