Classifying Temporal Relations Between Events

Nathanael Chambers and **Shan Wang** and **Dan Jurafsky**
Department of Computer Science
Stanford University
Stanford, CA 94305
{natec,shanwang,jurafsky}@stanford.edu

Abstract

This paper describes a fully automatic two-stage machine learning architecture that learns temporal relations between pairs of events. The first stage learns the temporal attributes of single event descriptions, such as tense, grammatical aspect, and aspectual class. These imperfect guesses, combined with other linguistic features, are then used in a second stage to classify the temporal relationship between two events. We present both an analysis of our new features and results on the TimeBank Corpus that is 3% higher than previous work that used perfect human tagged features.

1 Introduction

Temporal information encoded in textual descriptions of events has been of interest since the early days of natural language processing. Lately, it has seen renewed interest as Question Answering, Information Extraction and Summarization domains find it critical in order to proceed beyond surface understanding. With the recent creation of the Timebank Corpus (Pustejovsky et al., 2003), the utility of machine learning techniques can now be tested.

Recent work with the Timebank Corpus has revealed that the six-class classification of temporal relations is very difficult, even for human annotators. The highest score reported on Timebank achieved 62.5% accuracy when using gold-standard features as marked by humans (Mani et al., 2006). This paper describes an approach using features extracted automatically from raw text that not only duplicates this performance, but surpasses its accuracy by 3%. We do so through advanced linguistic features and a surprising finding that using automatic rather than hand-labeled tense and aspect knowledge causes only a slight performance degradation.

We briefly describe current work on temporal ordering in section 2. Section 4 describes the first stage of basic temporal extraction, followed by a full description of the second stage in 5. The evaluation and results on Timebank then follow in section 6.

2 Previous Work

Mani et. al (2006) built a MaxEnt classifier that assigns each pair of events one of 6 relations from an augmented Timebank corpus. Their classifier relies on perfect features that were hand-tagged in the corpus, including tense, aspect, modality, polarity and event class. Pairwise agreement on tense and aspect are also included. In a second study, they applied rules of temporal transitivity to greatly expand the corpus, providing different results on this enlarged dataset. We could not duplicate their reported performance on this enlarged data, and instead focus on performing well on the Timebank data itself.

Lapata and Lascarides (2006) trained an event classifier for inter-sentential events. They built a corpus by saving sentences that contained two events, one of which is triggered by a key time word (e.g. *after* and *before*). Their learner was based on syntax and clausal ordering features. Boguraev and Ando (2005) evaluated machine learning on related tasks, but not relevant to event-event classification.

Our work is most similar to Mani's in that we are

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 173–176,
Prague, June 2007.

learning relations given event pairs, but our work extends their results both with new features and by using fully automatic linguistic features from raw text that are not hand selected from a corpus.

3 Data

We used the Timebank Corpus (v1.1) for evaluation, 186 newswire documents with 3345 event pairs. Solely for comparison with Mani, we add the 73 document Opinion Corpus (Mani et al., 2006) to create a larger dataset called the OTC. We present both Timebank and OTC results so future work can compare against either. All results below are from 10-fold cross validation.

4 Stage One: Learning Event Attributes

The task in Stage One is to learn the five temporal attributes associated with events as tagged in the Timebank Corpus. (1) *Tense* and (2) grammatical *aspect* are necessary in any approach to temporal ordering as they define both temporal location and structure of the event. (3) *Modality* and (4) *polarity* indicate hypothetical or non-occuring situations, and finally, (5) *event class* is the type of event (e.g. process, state, etc.). The *event class* has 7 values in Timebank, but we believe this paper's approach is compatible with other class divisions as well. The range of values for each event attribute is as follows, also found in (Pustejovsky et al., 2003):

tense	none, present, past, future
aspect	none, prog, perfect, prog_perfect
class	report, aspectual, state, I_state I_action, perception, occurrence
modality	none, to, should, would, could can, might
polarity	positive, negative

4.1 Machine Learning Classification

We used a machine learning approach to learn each of the five event attributes. We implemented both Naive Bayes and Maximum Entropy classifiers, but found Naive Bayes to perform as well or better than Maximum Entropy. The results in this paper are from Naive Bayes with Laplace smoothing.

The features we used on this stage include part of speech tags (two before the event), lemmas of the event words, WordNet synsets, and the appearance of auxiliaries and modals before the event. This latter set included all derivations of *be* and *have* auxiliaries, modal words (e.g. may, might, etc.), and the presence/absence of *not*. We performed feature selection on this list of features, learning a different set of features for each of the five attributes. The list of selected features for each is shown in figure 1.

tense	POS-2-event, POS-1-event POS-of-event, have_word, be_word
aspect	POS-of-event, modal_word, be_word
class	synset
modality	none
polarity	none

Figure 1: Features selected for learning each temporal attribute. POS-2 is two tokens before the event.

Timebank Corpus			
	tense	aspect	class
Baseline	52.21	84.34	54.21
Accuracy	88.28	94.24	75.2
Baseline (OTC)	48.52	86.68	59.39
Accuracy (OTC)	87.46	88.15	76.1

Figure 2: Stage One results on classification.

Modality and *polarity* did not select any features because their majority class baselines were so high (98%) that learning these attributes does not provide much utility. A deeper analysis of event interaction would require a modal analysis, but it seems that a newswire domain does not provide great variation in modalities. Consequently, modality and polarity are not used in Stage Two. Tense, aspect and class are shown in figure 2 with majority class baselines. Tense classification achieves 36% absolute improvement, aspect 10% and class 21%. Performance on the OTC set is similar, although aspect is not as good. These guesses are then passed to Stage Two.

5 Stage Two: Event-Event Features

The task in this stage is to choose the temporal relation between two events, given the pair of events. We assume that the events have been extracted and that there exists some relation between them; the task is to choose the relation. The Timebank Corpus uses relations that are based on Allen's set of thirteen (Allen, 1984). Six of the relations are inverses of the other six, and so we condense the set to *before*, *ibefore*, *includes*, *begins*, *ends* and *simultaneous*. We map the thirteenth *identity* into *simultaneous*. One oddity is that Timebank includes both *during* and *included_by* relations, but *during* does not appear in Timebank documentation. While we don't know how previous work handles this, we condense *during* into *included_by* (invert to *includes*).

5.1 Features

Event Specific: The five temporal attributes from Stage One are used for each event in the pair, as well as the event strings, lemmas and WordNet synsets. Mani added two other features from these, indicators if the events agree on tense and aspect. We add a third, event class agreement. Further, to capture the dependency between events in a discourse, we create new bigram features of tense, aspect and class (e.g. "present past" if the first event is in the present, and the second past).

Part of Speech: For each event, we include the Penn Treebank POS tag of the event, the tags for the two tokens preceding, and one token following. We use the Stanford Parser[1] to extract them. We also extend previous work and create bigram POS features of the event and the token before it, as well as the bigram POS of the first event and the second event.

Event-Event Syntactic Properties: A phrase P is said to dominate another phrase Q if Q is a daughter node of P in the syntactic parse tree. We leverage the syntactic output of the parser to create the *dominance* feature for intra-sentential events. It is either on or off, depending on the two events' syntactic dominance. Lapata used a similar feature for subordinate phrases and an indicator *before* for textual event ordering. We adopt these features and also add a *same-sentence* indicator if the events appear in the same sentence.

Prepositional Phrase: Since preposition heads are often indicators of temporal class, we created a new feature indicating when an event is part of a prepositional phrase. The feature's values range over 34 English prepositions. Combined with event dominance (above), these two features capture direct intra-sentential relationships. To our knowledge, we are the first to use this feature in temporal ordering.

Temporal Discourse: Seeing tense as a type of anaphora, it is a natural conclusion that the relationship between two events becomes stronger as the textual distance draws closer. Because of this, we adopted the view that intra-sentential events are generated from a different distribution than inter-sentential events. We therefore train two models during learning, one for events in the same sentence, and the other for events crossing sentence boundaries. It essentially splits the data on the *same_sentence* feature. As we will see, this turned out to be a very useful feature. It is called the *split* approach in the next section.

Example (require, compromise):
*"Their solution **required** a **compromise**..."*
Features
(lemma1: require) (lemma2: compromise) (dominates: yes) (tense-bigram: past-none) (aspect-bigram: none-none) (tense-match: no) (aspect-match: yes) (before: yes) (same-sent: yes)

6 Evaluation and Results

All results are from a 10-fold cross validation using SVM (Chang and Lin, 2001). We also evaluated Naive Bayes and Maximum Entropy. Naive Bayes (NB) returned similar results to SVM and we present feature selection results from NB to compare the added value of our new features.

The input to Stage Two is a list of pairs of events; the task is to classify each according to one of six temporal relations. Four sets of results are shown in figure 3. *Mani*, *Mani+Lapata* and *All+New* correspond to performance on features as listed in the figure. The three table columns indicate how a gold-standard Stage One (*Gold*) compares against imperfect guesses (*Auto*) and the guesses with split distributions (*Auto-Split*).

A clear improvement is seen in each row, indicating that our new features provide significant improvement over previous work. A decrease in performance is seen between columns *gold* and *auto*, as expected, because imperfect data is introduced, however, the drop is manageable. The *auto-split* distributions make significant gains for the Mani and Lapata features, but less when all new features are

[1] http://nlp.stanford.edu/software/lex-parser.shtml

Timebank Corpus	Gold	Auto	Auto-Split
Baseline	37.22	37.22	46.58
Mani	50.97	50.19	53.42
Mani+Lapata	52.29	51.57	55.10
All+New	60.45	59.13	**59.43**

Mani stage one attributes, tense/aspect-match, event strings
Lapata dominance, before, lemma, synset
New prep-phrases, same-sent, class-match, POS uni/bigrams, tense/aspect/class-bigrams

Figure 3: Incremental accuracy by adding features.

Same Sentence		Diff Sentence	
POS-1 Ev1	2.5%	Tense Pair	1.6%
POS Bigram Ev1	3.5%	Aspect Ev1	0.5%
Preposition Ev1	2.0%	POS Bigram	0.2%
Tense Ev2	0.7%	POS-1 Ev2	0.3%
Preposition Ev2	0.6%	Word EV2	0.2%

Figure 4: Top 5 features as added in feature selection w/ Naive Bayes, with their percentage improvement.

involved. The highest fully-automatic accuracy on Timebank is 59.43%, a 4.3% gain from our new features. We also report 67.57% *gold* and 65.48% *auto-split* on the OTC dataset to compare against Mani's reported hand-tagged features of 62.5%, a gain of 3% with our automatic features.

7 Discussion

Previous work on OTC achieved classification accuracy of 62.5%, but this result was based on "perfect data" from human annotators. A low number from good data is at first disappointing, however, we show that performance can be improved through more linguistic features and by isolating the distinct tasks of ordering inter-sentential and intra-sentential events.

Our new features show a clear improvement over previous work. The features that capture dependencies between the events, rather than isolated features provide the greatest utility. Also, the impact of imperfect temporal data is surprisingly minimal. Using Stage One's results instead of gold values hurts performance by less than 1.4%. This suggests that much of the value of the hand-coded information can be achieved via automatic approaches. Stage One's *event class* shows room for improvement, yet the negative impact on Event-Event relationships is manageable. It is conceivable that more advanced features would better classify the *event class*, but improvement on the event-event task would be slight.

Finally, it is important to note the difference in classifying events in the same sentence vs. cross-boundary. Splitting the 3345 pairs of corpus events into two separate training sets makes our data more sparse, but we still see a performance improvement when using Mani/Lapata features. Figure 4 gives a hint to the difference in distributions as the best features of each task are very different. Intra-sentence events rely on syntax cues (e.g. preposition phrases and POS), while inter-sentence events use tense and aspect. However, the differences are minimized as more advanced features are added. The final row in figure 3 shows minimal split improvement.

8 Conclusion

We have described a two-stage machine learning approach to event-event temporal relation classification. We have shown that imperfect event attributes can be used effectively, that a range of event-event dependency features provide added utility to a classifier, and that events within the same sentence have distinct characteristics from those across sentence boundaries. This fully automatic raw text approach achieves a 3% improvement over previous work based on perfect human tagged features.

Acknowledgement: This work was supported in part by the DARPA GALE Program and the DTO AQUAINT Program.

References

James Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

Branimir Boguraev and Rie Kubota Ando. 2005. Timeml-compliant text analysis for temporal reasoning. In *IJCA-05*.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Mirella Lapata and Alex Lascarides. 2006. Learning sentence-internal temporal relations. In *Journal of AI Research*, volume 27, pages 85–117.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *ACL-06*, July.

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. *Corpus Linguistics*, pages 647–656.