Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations

Steven Bethard
Institute for Cognitive Science
Department of Computer Science
University of Colorado
Boulder, CO 80309, USA
steven.bethard@colorado.edu

James H. Martin
Institute for Cognitive Science
Department of Computer Science
University of Colorado
Boulder, CO 80309, USA
james.martin@colorado.edu

Abstract

Finding temporal and causal relations is crucial to understanding the semantic structure of a text. Since existing corpora provide no parallel temporal and causal annotations, we annotated 1000 conjoined event pairs, achieving inter-annotator agreement of 81.2% on temporal relations and 77.8% on causal relations. We trained machine learning models using features derived from WordNet and the Google N-gram corpus, and they outperformed a variety of baselines, achieving an F-measure of 49.0 for temporals and 52.4 for causals. Analysis of these models suggests that additional data will improve performance, and that temporal information is crucial to causal relation identification.

1 Introduction

Working out how events are tied together temporally and causally is a crucial component for successful natural language understanding. Consider the text:

(1) I ate a bad tuna sandwich, got food poisoning and had to have a shot in my shoulder. *wsj_0409*

To understand the semantic structure here, a system must order events along a timeline, recognizing that *getting food poisoning* occurred BEFORE *having a shot*. The system must also identify when an event is not independent of the surrounding events, e.g. *got food poisoning* was CAUSED by *eating a bad sandwich*. Recognizing these temporal and causal relations is crucial for applications like question answering which must face queries like *How did he get food poisoning?* or *What was the treatment?*

Currently, no existing resource has all the necessary pieces for investigating parallel temporal and causal phenomena. The TimeBank (Pustejovsky et al., 2003) links events with BEFORE and AFTER relations, but includes no causal links. PropBank (Kingsbury and Palmer, 2002) identifies ARGM-TMP and ARGM-CAU relations, but arguments may only be temporal or causal, never both. Thus existing corpora are missing some crucial pieces for studying temporal-causal interactions. Our research aims to fill these gaps by building a corpus of parallel temporal and causal relations and exploring machine learning approaches to extracting these relations.

2 Related Work

Much recent work on temporal relations revolved around the TimeBank and TempEval (Verhagen et al., 2007). These works annotated temporal relations between events and times, but low inter-annotator agreement made many TimeBank and TempEval tasks difficult (Boguraev and Ando, 2005; Verhagen et al., 2007). Still, TempEval showed that on a constrained tense identification task, systems could achieve accuracies in the 80s, and Bethard and colleagues (Bethard et al., 2007) showed that temporal relations between a verb and a complement clause could be identified with accuracies of nearly 90%.

Recent work on causal relations has also found that arbitrary relations in text are difficult to annotate and give poor system performance (Reitter, 2003). Girju and colleagues have made progress by selecting constrained pairs of events using web search patterns. Both manually generated Cause-Effect patterns (Girju et al., 2007) and patterns based on nouns

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 177–180,
Columbus, Ohio, USA, June 2008.

	Full	Train	Test
Documents	556	344	212
Event pairs	1000	697	303
BEFORE relations	313	232	81
AFTER relations	16	11	5
CAUSAL relations	271	207	64

Table 1: Contents of the corpus and its train/test sections

Task	Agreement	Kappa	F
Temporals	81.2	0.715	71.9
Causals	77.8	0.556	66.5

Table 2: Inter-annotator agreement by task.

linked causally in WordNet (Girju, 2003) were used to collect examples for annotation, with the resulting corpora allowing machine learning models to achieve performance in the 70s and 80s.

3 Conjoined Events Corpus

Prior work showed that finding temporal and causal relations is more tractable in carefully selected corpora. Thus we chose a simple construction that frequently expressed both temporal and causal relations, and accounted for 10% of all adjacent verbal events: events conjoined by the word *and*.

Our temporal annotation guidelines were based on the guidelines for TimeBank and TempEval, augmented with the guidelines of (Bethard et al., 2008). Annotators used the labels:

BEFORE The first event fully precedes the second
AFTER The second event fully precedes the first
NO-REL Neither event clearly precedes the other

Our causal annotation guidelines were based on paraphrasing rather than the intuitive notions of *cause* used in prior work (Girju, 2003; Girju et al., 2007). Annotators selected the best paraphrase of *"and"* from the following options:

CAUSAL *and as a result, and as a consequence, and enabled by that*
NO-REL *and independently, and for similar reasons*

To build the corpus, we first identified verbs that represented events by running the system of (Bethard and Martin, 2006) on the TreeBank. We then used a set of tree-walking rules to identify conjoined event pairs. 1000 pairs were annotated by two annotators and adjudicated by a third. Table 1

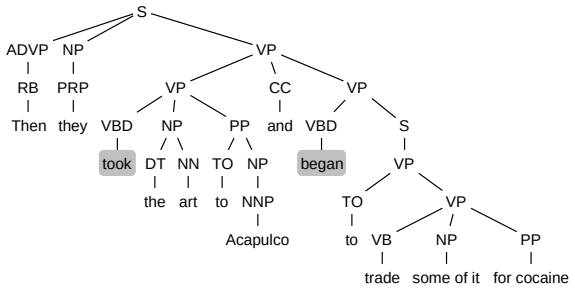

Figure 1: Syntactic tree from *wsj_0450* with events *took* and *began* highlighted.

and Table 2 give statistics for the resulting corpus[1]. The annotators had substantial agreement on temporals (81.2%) and moderate agreement on causals (77.8%). We also report F-measure agreement, since BEFORE, AFTER and CAUSAL relations are more interesting than NO-REL. Annotators had F-measure agreement of 71.9 on temporals and 66.5 causals.

4 Machine Learning Methods

We used our corpus for machine learning experiments where relation identification was viewed as pair-wise classification. Consider the sentence:

(2) The man who had brought it in for an estimate had [EVENT *returned*] to collect it and was [EVENT *waiting*] in the hall. *wsj_0450*

A temporal classifier should label *returned-waiting* with BEFORE since *returned* occurred first, and a causal classifier should label it CAUSAL since this *and* can be paraphrased as *and as a result*.

We identified both syntactic and semantic features for our task. These will be described using the example event pair in Figure 1. Our syntactic features characterized surrounding surface structures:

- The event words, lemmas and part-of-speech tags, e.g. *took*, *take*, `VBD` and *began*, *begin*, `VBD`.
- All words, lemmas and part-of-speech tags in the verb phrases of each event, e.g. *took*, *take*, `VBD` and *began, to, trade*, *begin, trade*, `VBD,TO,VB`.
- The syntactic paths from the first event to the common ancestor to the second event, e.g. `VBD>VP`, `VP` and `VP<VBD`.

[1]Train: *wsj_0416-wsj_0759*. Test: *wsj_0760-wsj_0971*. verbs.colorado.edu/~bethard/treebank-verb-conj-anns.xml

- All words before, between and after the event pair, e.g. *Then, they* plus *the, art, to, Acapulco, and* plus *to, trade, some, of, it, for, cocaine*.

Our semantic features encoded surrounding word meanings. We used WordNet (Fellbaum, 1998) root synsets (*roots*) and lexicographer file names (*lexnames*) to derive the following features:

- All event roots and lexnames, e.g. *take#33, move#1 ... body, change ...* for *took* and *be#0, begin#1 ... change, communication ...* for *began*.
- All lexnames before, between and after the event pair, e.g. *all* plus *artifact, location*, etc. plus *possession, artifact*, etc.
- All roots and lexnames shared by both events, e.g. *took* and *began* were both *act#0, be#0* and *change, communication*, etc.
- The least common ancestor (LCA) senses shared by both events, e.g. *took* and *began* meet only at their roots, so the LCA senses are *act#0* and *be#0*.

We also extracted temporal and causal word associations from the Google N-gram corpus (Brants and Franz, 2006), using `<keyword> <pronoun> <word>` patterns, where *before* and *after* were the keywords for temporals, and *because* was the keyword for causals. Word scores were assigned as:

$$score(w) = log\left(\frac{N_{keyword}(w)}{N(w)}\right)$$

where $N_{keyword}(w)$ is the number of times the word appeared in the keyword's pattern, and $N(w)$ is the number of times the word was in the corpus. The following features were derived from these scores:

- Whether the event score was in at least the Nth percentile, e.g. *took*'s -6.1 *because* score placed it above 84% of the scores, so the feature was true for $N = 70$ and $N = 80$, but false for $N = 90$.
- Whether the first event score was greater than the second by at least N, e.g. *took* and *began* have *after* scores of -6.3 and -6.2 so the feature was true for $N = -1$, but false for $N = 0$ and $N = 1$.

5 Results

We trained SVM^{perf} classifiers (Joachims, 2005) for the temporal and causal relation tasks[2] using the train/test split from Table 1 and the feature sets:

	Temporals			Causals		
Model	P	R	F1	P	R	F1
BEFORE	26.7	94.2	41.6	-	-	-
CAUSAL	-	-	-	21.1	100.0	34.8
1st Event	35.0	24.4	28.8	31.0	20.3	24.5
2nd Event	36.1	30.2	32.9	22.4	17.2	19.5
POS Pair	46.7	8.1	13.9	30.0	4.7	8.1
Syntactic	36.5	53.5	43.4	24.4	79.7	37.4
Semantic	35.8	55.8	43.6	27.2	64.1	38.1
All	43.6	55.8	**49.0**	27.0	59.4	37.1
All+Tmp	-	-	-	46.9	59.4	**52.4**

Table 3: Performance of the temporal relation identification models: (A)ccuracy, (P)recision, (R)ecall and (F1)-measure. The null label is NO-REL.

Syntactic The syntactic features from Section 4.
Semantic The semantic features from Section 4.
All Both syntactic and semantic features.
All+Tmp (Causals Only) Syntactic and semantic features, plus the gold-standard temporal label.

We compared our models against several baselines, using precision, recall and F-measure since the NO-REL labels were uninteresting. Two simple baselines had 0% recall: a lookup table of event word pairs[3], and the majority class (NO-REL) label for causals. We therefore considered the following baselines:

BEFORE Classify all instances as BEFORE, the majority class label for temporals.
CAUSAL Classify all instances as CAUSAL.
1st Event Use a lookup table of 1st words and the labels they were assigned in the training data.
2nd Event As **1st Event**, but using 2nd words.
POS Pair As **1st Event**, but using part of speech tag pairs. POS tags encode tense, so this suggests the performance of a tense-based classifier.

The results on our test data are shown in Table 3. For temporal relations, the F-measures of all SVM models exceeded all baselines, with the combination of syntactic and semantic features performing 5 points better (43.6% precision and 55.8% recall) than either feature set individually. This suggests that our syntactic and semantic features encoded complementary information for the temporal relation task. For

[2]We built multi-class SVMs using the *one-vs-rest* approach and used 5-fold cross-validation on the training data to set parameters. For temporals, C=0.1 (for syntactic-only models), C=1.0 (for all other models), and loss-function=F1 (for all models). For causals, C=0.1 and loss-function=precision/recall break even point (for all models).

[3]Only 3 word pairs from training were seen during testing.

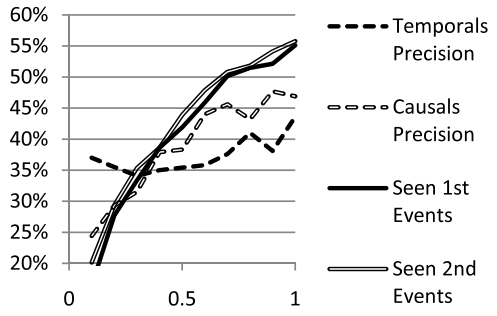

Figure 2: Model precisions (dotted lines) and percent of events in the test data seen during training (solid lines), given increasing fractions of the training data.

causal relations, all SVM models again exceeded all baselines, but combining syntactic features with semantic ones gained little. However, knowing about underlying temporal relations boosted performance to 46.9% precision and 59.4% recall. This shows that progress in causal relation identification will require knowledge of temporal relations.

We examined the effect of corpus size on our models by training them on increasing fractions of the training data and evaluating them on the test data. The precisions of the resulting models are shown as dotted lines in Figure 2. The models improve steadily, and the causals precision can be seen to follow the solid curves which show how event coverage increases with increased training data. A logarithmic trendline fit to these seen-event curves suggests that annotating all 5,013 event pairs in the Penn TreeBank could move event coverage up from the mid 50s to the mid 80s. Thus annotating additional data should provide a substantial benefit to our temporal and causal relation identification systems.

6 Conclusions

Our research fills a gap in existing corpora and NLP systems, examining parallel temporal and causal relations. We annotated 1000 event pairs conjoined by the word *and*, assigning each pair both a temporal and causal relation. Annotators achieved 81.2% agreement on temporal relations and 77.8% agreement on causal relations. Using features based on WordNet and the Google N-gram corpus, we trained support vector machine models that achieved 49.0 F on temporal relations, and 37.1 F on causal relations. Providing temporal information to the causal relations classifier boosted its results to 52.4 F. Future work will investigate increasing the size of the corpus and developing more statistical approaches like the Google N-gram scores to take advantage of large-scale resources to characterize word meaning.

Acknowledgments

This research was performed in part under an appointment to the U.S. Department of Homeland Security (DHS) Scholarship and Fellowship Program.

References

S. Bethard and J. H. Martin. 2006. Identification of event mentions and their semantic class. In *EMNLP-2006*.

S. Bethard, J. H. Martin, and S. Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *ICSC-2007*.

S. Bethard, W. Corvey, S. Klingenstein, and J. H. Martin. 2008. Building a corpus of temporal-causal structure. In *LREC-2008*.

B. Boguraev and R. K. Ando. 2005. Timebank-driven timeml analysis. In *Annotating, Extracting and Reasoning about Time and Events*. IBFI, Schloss Dagstuhl, Germany.

T. Brants and A. Franz. 2006. Web 1t 5-gram version 1. Linguistic Data Consortium, Philadelphia.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press.

R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *SemEval-2007*.

R. Girju. 2003. Automatic detection of causal relations for question answering. In *ACL Workshop on Multilingual Summarization and Question Answering*.

T. Joachims. 2005. A support vector method for multivariate performance measures. In *ICML-2005*.

P. Kingsbury and M. Palmer. 2002. From Treebank to PropBank. In *LREC-2002*.

J. Pustejovsky, P. Hanks, R. Saurí, 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.

D. Reitter. 2003. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *LDV-Forum, GLDV-Journal for Computational Linguistics and Language Technology*, 18(1/2):38–52.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *SemEval-2007*.