Building a Corpus of Temporal-Causal Structure

Steven Bethard*, William Corvey†, Sara Klingenstein†, James H. Martin*

*Department of Computer Science
†Department of Linguistics
University of Colorado at Boulder
Boulder, Colorado 80309, USA
{steven.bethard,william.corvey}@colorado.edu
{sara.klingenstein,james.martin}@colorado.edu

Abstract

While recent corpus annotation efforts cover a wide variety of semantic structures, work on temporal and causal relations is still in its early stages. Annotation efforts have typically considered either temporal relations or causal relations, but not both, and no corpora currently exist that allow the relation between temporals and causals to be examined empirically. We have annotated a corpus of 1000 event pairs for both temporal and causal relations, focusing on a relatively frequent construction in which the events are conjoined by the word *and*. Temporal relations were annotated using an extension of the BEFORE and AFTER scheme used in the TempEval competition, and causal relations were annotated using a scheme based on connective phrases like *and as a result*. The annotators achieved 81.2% agreement on temporal relations and 77.8% agreement on causal relations. Analysis of the resulting corpus revealed some interesting findings, for example, that over 30% of CAUSAL relations do not have an underlying BEFORE relation. The corpus was also explored using machine learning methods, and while model performance exceeded all baselines, the results suggested that simple grammatical cues may be insufficient for identifying the more difficult temporal and causal relations.

1. Introduction

Recent corpus annotation efforts have made the semantic structure of text much more accessible. Projects like PropBank (Kingsbury and Palmer, 2002), TimeBank (Pustejovsky et al., 2003) and the Penn Discourse TreeBank (Miltsakaki et al., 2004) have linked words together with a wide variety of semantic relations. Still, many gaps exist. Consider the following text from the Penn TreeBank (Marcus et al., 1994):

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

It is clear to readers of this sentence that the food poisoning occurred BEFORE the shot in the shoulder, and that the CAUSE of the *food poisoning* was the *eating* of the sandwich. But this information is not annotated by any existing resource. In the TimeBank, no causal relations were annotated, and temporal relations were only annotated for pairs of events that the annotators deemed important. In PropBank, both temporal and causal relations were annotated, but `ARGM-TMP` did not distinguish between BEFORE and AFTER relations, and pairs of events could never be annotated as both `ARGM-TMP` and `ARGM-CAU`. Moreover, PropBank only annotated verbal arguments, so conjoined event constructions like the example above were out of the scope of the project. The Penn Discourse TreeBank annotated some conjoined event constructions, but only when full clauses were conjoined, and then only indicating the clause boundaries, not the type of temporal or causal relation between them.

Thus, work is needed to fill the gaps between these resources, in particular, to investigate parallel temporal and causal relations. This article describes the annotation of a corpus of such relations, with an initial focus on the conjoined event construction. This construction is frequently used to express both temporal and causal relations, and accounts for about 10% of all adjacent verbal events. Thus it was a good choice as a starting point to explore interactions between temporal and causal relations.

The remainder of this article is structured as follows. Section 3 and Section 4 describe how the annotation schemes for temporal and causal relations were developed. Section 5 and Section 6 give some details of the resulting corpus, and Section 7 describes some preliminary machine learning experiments. Section 8 summarizes the results and suggests some future directions.

2. Related Work

Research on temporal and causal relations has generally progressed as two separate fields, one focusing on linking events and times, and one focusing on causality. Recent work on temporal relations has mostly revolved around the TimeBank corpus (Pustejovsky et al., 2003), a small set of newswire documents annotated for events, times and the temporal relations between them. A variety of systems for identifying temporal relations were trained on this corpus (Boguraev and Ando, 2005; Mani et al., 2006) but systems had poor performance, in part due to the the low interannotator agreement and fine granularity of the TimeBank temporal relations.

In an attempt to improve on the TimeBank annotation scheme, Verhagen and colleagues organized the TempEval competition (Verhagen et al., 2007) which used a stricter annotation interface and a simplified set of temporal relations. Systems performed well on its tense identification task, but poorly on the other tasks which often required multiple stages of implicit temporal logic (Puşcaşu, 2007; Bethard and Martin, 2007). Building on the lessons of TimeBank and TempEval, Bethard and colleagues (Bethard et al., 2007) annotated some verb-clause constructions in the TimeBank, and showed that with a small amount of data, support vector machine models could be trained to find these temporal relations with accuracies of nearly 90%. Like work on temporal relations, early work in causal relations aimed to identify the relations in arbitrary text. Khoo and colleagues (Khoo et al., 2000; Khoo et al., 1998) tried to identify all causal relations in a section of the Wall Street Journal using hand-crafted patterns, but had inter-annotator agreement problems, and achieved only 24.9% precision and 67.7% recall with their patterns. Reitter (Reitter, 2003) trained support vector machine models on discourse relations like Attribution, Cause and Elaboration annotated on top of the Wall Street Journal, but while his system performed well for relations like Elaboration, for relations like Cause and Effect both precision and recall were under 25%.

Girju and colleagues took a step away from the whole-corpus style of annotation, and instead considered selected subsets of corpora. They identified verbs likely to indicate causal relations by finding nouns in WordNet linked by the word *cause* and searching the web for verbs between them. After annotating sentences for each of these verbs with CAUSAL and NON-CAUSAL relations, they were able to train decision tree models that achieved 73.9% precision and 88.7% recall. Inspired by the success of this approach, Girju and colleagues (Girju et al., 2007) organized a SemEval 2007 task in which pairs of nouns were selected by carefully constructed web search queries, and annotated for the presence or absence of relations like Cause-Effect. A system based on support vector machines was able to distinguish Cause-Effect noun pairs from other noun pairs with 77.5% accuracy (Beamer et al., 2007).

Thus, the prior work on both temporal and causal relations point to a similar conclusion: finding temporal and causal relations in arbitrary text is difficult, but in carefully selected subsets of corpora finding these relations can be much easier. Thus we follow this approach, and build our corpus by selecting a syntactically motivated subset of event pairs: event pairs conjoined by the word *and*. In preparation for the annotation of such a corpus, we designed two annotation schemes: one for temporal relations and one for causal relations.

3. Temporal Annotation Scheme

The TempEval (Verhagen et al., 2007) guidelines served as a starting point for the temporal annotation work here. TempEval tried to simplify the TimeBank annotation scheme, using the labels BEFORE, OVERLAP, AFTER, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE. We decided to focus only on the two basic BEFORE and AFTER relations, allowing our annotators to choose from the following labels:

BEFORE The first event fully precedes the second

AFTER The first event fully follows the second

NO-REL Neither event clearly precedes the other

To make these definitions a little more concrete, we provided the following additional guidelines.

Events were conceptualized separately from their tense and aspect markings. For example:

(2) The funding mechanism, which has [$_{\text{EVENT}}$ *received*] congressional approval and is [$_{\text{EVENT}}$ *expected*] to be signed by President Bush, would affect the antitrust operations of the Justice Department and the Federal Trade Commission. `wsj_0119`

Though the phrase *has received* may be conceived as a state, the *receiving* event itself is viewed as occurring strictly at the moment of reception, and so this instance was annotated as (*received* BEFORE *expected*).

Modal or conditional events were evaluated using a possible worlds analysis. Consider the sentence:

(3) Persons who examine the materials may [$_{\text{EVENT}}$ *make*] notes and no one will [$_{\text{EVENT}}$ *check*] to determine what notes a person has taken. `wsj_0108`

Here, though neither the *note-making* nor the *note-checking* have occurred at the time of the utterance, the instance was annotated as (*make* BEFORE *check*) because in the possible world where notes are *made* and *checked*, the *making* will have occurred before the *checking*.

Events that could be interpreted as overlapping on at least one endpoint were annotated with NO-REL. For example:

(4) NL shares [$_{\text{EVENT}}$ *closed*] unchanged at $22.75 and Valhi [$_{\text{EVENT}}$ *rose*] 62.5 cents to $15. `wsj_0080`

Since the *closing* event could either be interpreted as following the *rising* event or coinciding with the end of the *rising*, this instance was annotated as (*closed* NO-REL *rose*).

Events with a negative modifier or with a nonexistent subject (e.g. *nobody*) were annotated with NO-REL. For example:

(5) Mr. Black said he is "[$_{\text{EVENT}}$ *pleased*]" with the economy's recent performance, and doesn't [$_{\text{EVENT}}$ *see*] "a lot of excesses out there" `wsj_0072`

Trying to treat this as a regular *see* event is complicated because the *seeing* never occurred, and even in a possible worlds analysis, the *seeing* can not be placed at a particular time. Thus the instance was annotated as (*pleased* NO-REL *(doesn't) see*).

Ambiguous cases were annotated with NO-REL. For example:

(6) Nashua immediately responded by [$_{\text{EVENT}}$ *strengthening*] a poison-pill plan and [$_{\text{EVENT}}$ *saying*] it will buy back up to one million of its shares `wsj_0520`

Since the *strengthening* is not clearly before the *saying* nor is the *saying* clearly before the *strengthening*, this instance was annotated as (*strengthening* NO-REL *saying*).

4. Causal Annotation Scheme

Many earlier efforts at annotating causality relied on only intuitive notions of the term *cause* (Khoo et al., 2000; Girju, 2003; Girju et al., 2007). In an attempt to make these notions more explicit, a couple different causal annotation schemes were explored in the current work.

Event Pairs	Agreement	Kappa
50	78	0.67
100	64	0.46
200	80	0.70
200	74	0.61

Table 1: Agreement for necessary-sufficient annotations

One scheme was based on the classic formulation of causality in terms of *necessary* and *sufficient* conditions. So for example:

(7) The agency said it [EVENT *monitored*] Newmark & Lewis's advertised prices before and after the ad campaign, and [EVENT *found*] that the prices of at least 50 different items either increased or stayed the same. `wsj_0358`

The event *monitored* was annotated as being NECESSARY for the event *found* since the *finding* could not have occurred if the *monitoring* had not.

Analysis of annotator agreement showed some difficulties with this annotation scheme. Table 1 shows several samples of data annotated using the NECESSARY and SUFFICIENT labels. Agreement was lower than hoped, varied quite a bit between data sets, and did not seem to improve with training. In examining the disagreements, we found that annotators had trouble agreeing both on the direction of the relation (NECESSARY vs. SUFFICIENT), and on the boundaries of the two events. For an example of the latter problem, consider:

(8) A Japanese company might [EVENT *make*] television picture tubes in Japan, [EVENT *assemble*] the sets in Malaysia and [EVENT *export*] them to Indonesia. `wsj_0043`

While *making-picture-tubes* is clearly NECESSARY for *assembling-the-sets*, it is not true that *a-Japanese-company-making-picture-tubes* is NECESSARY for *assembling-the-sets-in-Malaysia*. Thus, a different sort of annotation scheme was needed.

To try to establish a closer link between the annotation labels and natural language, annotators were instead asked to judge the quality of several paraphrases of each sentence. The paraphrases were generated using both CAUSAL and NO-REL substitutions for the word *and*. The substitutions we considered were:

CAUSAL *and as a result, and as a consequence, and enabled by that*

NO-REL *and independently, and for similar reasons*

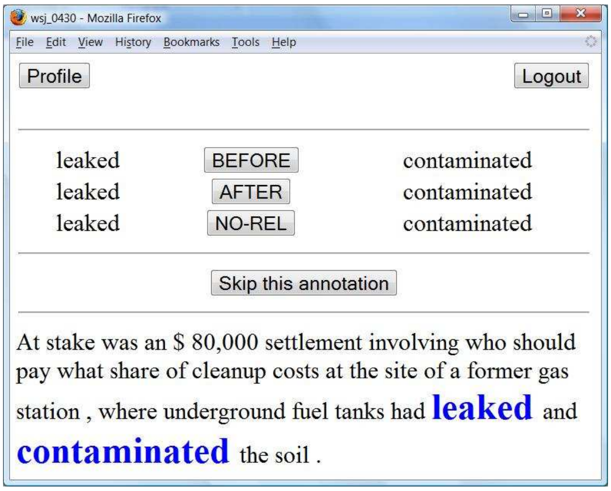
So given a sentence like:

(9) Fuel tanks had [EVENT *leaked*] and [EVENT *contaminated*] the soil. `wsj_0430`

Annotators determined that the best paraphrase was a CAUSAL one, and in particular, one that replaced *and* with *and as a result*. Note that under this scheme, the annotators were not required to determine the extent of an event, only to find the connective phrase that best matched the sentence semantics. Table 2 shows that agreement under this scheme was more stable and seemed to improve with training. Therefore this approach was used to annotate the corpus.

Event Pairs	Agreement	Kappa
100	76	0.52
100	78	0.56
100	82	0.64

Table 2: Agreement for paraphrase annotations


wsj_0430 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Profile
Logout
leaked BEFORE contaminated
leaked AFTER contaminated
leaked NO-REL contaminated
Skip this annotation

At stake was an $ 80,000 settlement involving who should pay what share of cleanup costs at the site of a former gas station , where underground fuel tanks had **leaked** and **contaminated** the soil .

Figure 1: The annotation interface for temporal relations. The interface for causal relations looked almost identical but with CAUSAL and NO-REL labels instead.

5. Corpus Annotation

The first step of the annotation process was to select sets of conjoined event pairs from the Penn TreeBank. Because gold standard events were not available for the entire TreeBank, events were first identified automatically, using the event identification system of (Bethard and Martin, 2006). Conjoined event pairs were identified using a simple set of tree-walking rules, resulting in 5,013 event pairs[1]. These conjoined event pairs then served as the basis for the annotation.

For both temporal and causal annotation, annotators used a browser-based interface that showed a single sentence with the event pair highlighted, and asked them to select an appropriate label, as shown in Figure 1. Annotators were trained on the interface and the guidelines using several hundred event pairs from the beginning of the corpus. Once training was complete, annotators moved on to the main section of the corpus, 1000 event pairs from the Wall Street Journal documents 0416-0971. Annotation on this data was performed in parallel by two annotators, and then adjudicated afterward by a third[2].

[1] Verbs not identified as events by the system but conjoined to identified events were also assumed to be events.

[2] Annotation for temporal relations took roughly 30 seconds per instance, while annotation for causal relations took closer to one minute per instance.

	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 3: Number of documents, event pairs and different relation types in the corpus. These statistics are shown for the full corpus, the training section (`wsj_0416-wsj_0759`) and the test section (`wsj_0760-wsj_0971`).

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

Table 4: Inter-annotator agreement for temporal and causal relations.

The result of this annotation was a corpus of 1000 event pairs, annotated both for temporal and causal relations[3]. Table 3 gives some basic statistics for the corpus. On the average, there was about one BEFORE and one CAUSAL relation for every two documents in the corpus. For comparison, in the much more extensive PropBank project, `ARGM-TMP` roles average about nine times per document, while `ARGM-CAU` roles average a little less than once a document. Table 4 shows the inter-annotator agreement for our corpus. Since NO-REL labels indicated the lack of temporal or causal relations, in addition to simple agreement and the kappa statistic we also reported F-measure agreement between the annotators. F-measure agreement gives more importance to the labels BEFORE, AFTER and CAUSAL, and is calculated as twice the number of BEFORE, AFTER and CAUSAL labels that both annotators agreed on, divided by the total such labels that were annotated by all annotators[4]. The annotators had substantial agreement (81.2%, 0.715 kappa, 71.9 F) on temporal relations and moderate agreement (77.8%, 0.556 kappa, 66.5 F) on causal relations.

6. Corpus Analysis

This corpus offered the chance to explore some of the ties between the temporal and causal annotations. Initially we expected that almost every CAUSAL relation would be accompanied by an underlying BEFORE relation, since *causes* are generally expected to precede *effects*. In fact, 32% of CAUSAL relations in the corpus did not have an underlying BEFORE relation. For example:

(10) IBM established its standard to try to stop falling behind upstart Apple Computer, but NEC [$_{\text{EVENT}}$ *was*] ahead from the start and didn't [$_{\text{EVENT}}$ *need*] to invite in competitive allies.

Paraphrasing this sentence to say *NEC was ahead from the start and as a consequence didn't need to invite in competitive allies* sounds quite reasonable and maintains the same sentence semantics. Yet, on the temporal side, the annotators did not assign the relation (*was* BEFORE *need*) because neither of these events clearly preceded the other.

There seemed to be two major categories of event pairs like this that were causally related yet lacked a BEFORE relation. In about 55% of such event pairs, the first event was stative and overlapping with the second event, but the start of the first event preceded the start of the second event. For example:

(11) Japanese local governments are [$_{\text{EVENT}}$ *expected*] to invest heavily in computer systems over the next few years, and many companies [$_{\text{EVENT}}$ *expect*] that field to provide substantial revenue.

Both *expecting* events are occurring simultaneously, yet *and as a result* is a good paraphrase here. This seems to be a due to the fact that the *expected to invest* event began before the *expected to provide revenue* event, allowing the beginning of the *expected to invest* event to serve as the cause for the other event. This suggests that it may be useful to introduce more fine-grained relation labels than simply BEFORE and AFTER.

Another 30% of CAUSAL-but-not-BEFORE event pairs was accounted for by events that were so closely related that they appeared as two different views of the same event. Example 10 is of this type, as was the following example:

(12) Abbie [$_{\text{EVENT}}$ *lies*] back and [$_{\text{EVENT}}$ *leaves*] the frame empty. `wsj_0633`

Here, *lying back* and *leaving the frame empty* are really part of the same event, and therefore occur simultaneously. Still, *and as a result* was a good paraphrase for this sentence and so it was annotated CAUSAL. The interpretation here seems to be that the less agentive view of the event, *leaving the frame empty* is the result of the more agentive view, *lying back*. This suggests that it may be useful to include some sort of event identity relation in the annotation schema.

In addition to our explorations of the annotation schemas, we also explored how predictive some surface-level features were of the presence of a temporal or causal relation. A natural first place to look would be a difference in tenses, e.g. a past tense event would likely occur before a present tense event. There were no gold standard tense annotations in our data, but there were gold standard part of speech annotations from the Penn TreeBank which included tags like `VBD` (past tense verb) and `VBZ` (present tense, third person singular verb). Thus we explored part of speech tags as a proxy for tense. However, it turned out that in over 75% of event pairs, both events shared the same part of speech tag. This matches the common linguistic belief that coordinated structures, like the conjunction construction considered here, prefer parallel structures, e.g. the same tense

[3] This data is available at `http://verbs.colorado.edu/~bethard/treebank-verb-conj-anns.xml`

[4] This formula is derived by simplifying the standard formula for F-measure which depends on precision and recall. For a pair of annotators A and B, precision is the number of causal labels they agreed on, L_{AB}, divided by the number of causal labels annotator A identified, L_A. Recall is the number agreed on divided by annotator B's number of causal labels, L_B. F-measure is the harmonic mean of precision and recall, thus: $F = \frac{2*P*R}{P+R} = \frac{2*\frac{L_{AB}}{L_A}*\frac{L_{AB}}{L_B}}{\frac{L_{AB}}{L_A}+\frac{L_{AB}}{L_B}} = \frac{2*L_{AB}*\frac{L_{AB}}{L_A*L_B}}{L_{AB}*(\frac{1}{L_A}+\frac{1}{L_B})} = \frac{2*\frac{L_{AB}}{L_A*L_B}}{\frac{L_A+L_B}{L_A*L_B}} = \frac{2*L_{AB}}{L_A+L_B}$

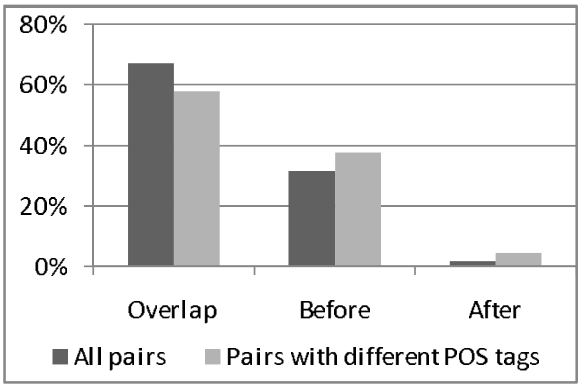


Figure 2: Distribution of BEFORE, AFTER and CAUSAL relations for all event pairs in the corpus and for the 25% of event pairs that had different part of speech tags.

in both branches of the coordination. Of the 25% of event pairs that did differ in their part of speech tags, the distribution of BEFORE, AFTER and CAUSAL relations was much like that of the overall corpus, as shown in Figure 2. Thus, part of speech, and therefore tense, seemed to be a poor predictor of temporal or causal relations in our coordinate constructions.

Finally, we looked at the distribution of events in the corpus. There were 1124 unique event words across all event pairs, with 708 unique first events, and 665 unique second events. Since there were only a total of 1000 event pairs in the corpus, this means that only about 30% of events in either position were observed more than once in the corpus. Even more striking was that, in the 1000 total event pairs, there were 975 unique word pairs, meaning that only 25 event pairs (2.5%) were observed more than once. There were only four word pairs observed more than twice – *buy-sell*, *rose-was*, *called-said* and *said-said* – and the last of these was observed with all possible labels (BEFORE, AFTER, CAUSAL and NO-REL). Thus not only is the data quite sparse in terms of event pairs, but observing an event pair with one label may be a poor predictor of the label for that event pair in a new context. This suggested that the task of automatically learning such temporal and causal relations would be quite challenging.

7. Machine Learning Experiments

We treated the automatic identification of temporal and causal relations as pair-wise classification problems, i.e. given a pair of events, we asked a classifier to label the pair with an appropriate relation type. For example, consider the sentence:

(13) The man who had brought it in for an estimate had [$_{\text{EVENT}}$ *returned*] to collect it and was [$_{\text{EVENT}}$ *waiting*] in the hall. `wsj_0450`

The temporal relation classifier should examine the events *returned* and *waiting* and assign them the label BEFORE since *returned* occurred first. Similarly, the causal relation classifier should examine the pair and assign them the label CAUSAL since this *and* can be paraphrased as *and as a result*. This approach treats temporal relation identification as a three-way classification task between BEFORE, AFTER and NO-REL, and causal relation identification as a two-way classification task between CAUSAL and NO-REL.

We chose support vector machine (SVM) classifiers for our machine learning experiments because they have been successful in a variety of related NLP tasks (Reitter, 2003; Pradhan et al., 2005; Bethard et al., 2007). In particular, we used the SVM^{perf} implementation because it has dramatically reduced training times and can optimize against the F1-measure and other loss functions directly. SVMs are binary classifiers, so to produce multiclass classifiers (for the temporal relations task), we applied the standard *one-vs-rest* formulation in which one binary SVM is trained for each possible label, and labels are assigned by finding the binary SVM which assigns the highest value to its label.

Like all machine learning algorithms, SVMs require that we characterize each pair of events with a set of *features* which identify the clues we'd like the learning algorithm to consider. We used a set of lexical and syntactic features based on the work of (Bethard et al., 2007). We refer to the following sentence and its syntactic tree as shown in Figure 3 to illustrate these features:

(14) Then they [$_{\text{EVENT}}$ *took*] the art to Acapulco and [$_{\text{EVENT}}$ *began*] to trade some of it for cocaine `wsj_0450`

The features were:

- The text of the events, e.g. *took* and *began*
- The event lemmas, e.g. *take* and *begin*
- The event part-of-speech tags, e.g. `VBD` and `VBD`
- All words in the verb phrases of each event, e.g. *took* and *began, to, trade*.
- The lemmas of all content words in the verb phrases of each event, e.g. *take* and *begin, trade*.
- The part-of-speech tags for all words in the verb phrases of each event, e.g. `VBD` and `VBD,TO,VB`.
- The syntactic category of the events' common ancestor in the syntactic tree, e.g. `VP`.
- The sequence of syntactic tags from the first event to the common ancestor, e.g. `VBD>VP`.
- The sequence of syntactic tags from the common ancestor to the second event, e.g. `VP<VBD`.
- All words preceding the first event, e.g. *Then, they*.
- All words between the two events, e.g. *the, art, to, Acapulco, and*.
- All words following the second event, e.g. *to, trade, some, of, it, for, cocaine*.

Using these features, we trained our SVM classifiers for the temporal and causal relation identification tasks. The corpus was split into a train section of 697 event pairs, and a test section of 303 event pairs as shown in Table 3. SVM^{perf}

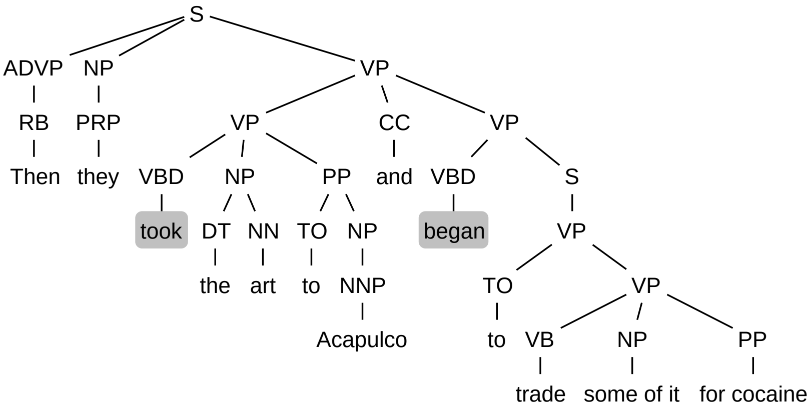


Figure 3: Syntactic tree for Example 14 with events *took* and *began* highlighted.

models have a number of free parameters, which we set by exploring a variety of different settings and evaluating their performance using five-fold cross-validations on the training data[5].

We compared our models against several baselines:

All <label> Classifies all instances with the same label. **All BEFORE** was the majority class baseline for temporal relations, and **All NO-REL** was the majority class baseline for causal relations.

Memorize Event Pair Looks at the pair of event words, classifying new pairs with the most common relation seen for that pair of event words in the training data. Uses the majority class label for unseen event word pairs.

Memorize 1st Event Similar to **Memorize Event Pair**, but it only looks at the first event word in the pair.

Memorize 2nd Event Similar to **Memorize Event Pair**, but it only looks at the second event word in the pair.

Memorize POS Pair Similar to **Memorize Event Pair**, but it looks at the part of speech tags for the words instead of the words themselves. This serves as a proxy for a tense based analysis, since the part of speech tags encode some tense information, e.g. VBD is a past tense verb, and VBZ is a present tense (3rd person singular) verb.

The results on our test data for these baselines and the SVM models are shown in Table 5 and Table 6. Note that we report model performance in terms of precision, recall and F-measure instead of simple accuracy since NO-REL labels simply indicate the lack of a BEFORE, AFTER or CAUSAL relation. Thus, under this evaluation, the **All NO-REL** baselines identify no relations of interest, and so they receive 0% recall.

Model	P	R	F1
All NO-REL	-	0.0	0.0
All BEFORE	26.7	94.2	41.6
Memorize Event Pair	0.0	0.0	0.0
Memorize 1st Event	35.0	24.4	28.8
Memorize 2nd Event	36.1	30.2	32.9
Memorize POS Pair	46.7	8.1	13.9
SVM	36.5	53.5	43.4

Table 5: Performance of the temporal relation identification models: (A)ccuracy, (P)recision, (R)ecall and (F1)-measure.

Model	P	R	F1
All NO-REL	-	0.0	0.0
All CAUSAL	21.1	100.0	34.8
Memorize Event Pair	0.0	0.0	0.0
Memorize 1st Event	31.0	20.3	24.5
Memorize 2nd Event	22.4	17.2	19.5
Memorize POS Pair	30.0	4.7	8.1
SVM	24.4	79.7	37.4

Table 6: Performance of the causal relation identification models: (P)recision, (R)ecall and (F1)-measure.

The **Memorize Event Pair** models were poor baselines, identifying no BEFORE, AFTER or CAUSAL relations at all. This is mainly due to the sparsity of event pairs – only 3 event pairs seen in the training data were also seen in the test data. The **Memorize 1st Event** and **Memorize 2nd Event** baselines address these sparsity problems to some degree, but there is clearly not enough information in a single event word to guess an appropriate temporal or causal relation – F-measures for these models reach only as high as 32.9 for temporals and 24.5 for causals.

Looking only at part of speech tag pairs also avoids the data sparsity problem, giving some of the highest precisions in both tasks (46.7% for temporals and 30.0% for causals). Still, recalls for these models are extremely low, under 10% for both tasks. Since the parts of speech encode much of the tense information, this is a clear indicator that simple

[5] A C of 0.1 was selected for all models. The F1 loss function was selected for temporal classification, while the precision/recall break-even point loss function was selected for causal classification.

tense analysis is not sufficient for the difficult tasks here. This confirms the hypothesis from Section 6 where it was noted that most events in pairs had exactly the same part of speech tag, and that the few differences in tense did not seem to usefully predict temporal or causal relations.

The SVM models outperform all baselines in F-measure, scoring 43.4 for temporal relations and 37.4 for causal relations. It is promising to see that machine learning models can combine the variety of surface and syntactic features they were given to outperform the baselines here. Yet the performance of these models is still quite low, most likely because the features encode only lexical and syntactic information, not the deep semantic information that is necessary for these tasks. We performed a basic error analysis of the models, and found that around 50% of the errors require some sort of world knowledge. Here are some examples of such errors, where the system labeled it NO-REL, but should have labeled it BEFORE:

(15) A former U.S. Marine, Mr. Dinkins got off to a quick start in politics, joining a local Democratic political club in the 1950s, [$_{\text{EVENT}}$ *linking*] up with black urban leaders such as Charles Rangel, Basil Paterson and Mr. Sutton, and [$_{\text{EVENT}}$ *getting*] himself elected to the state assembly in 1965. wsj_0765

(16) "I will [$_{\text{EVENT}}$ *sit*] down and [$_{\text{EVENT}}$ *talk*] some of the problems out, but take on the political system? Uh-uh," he says with a shake of the head. wsj_0765

(17) Some of the funds will be used to [$_{\text{EVENT}}$ *demolish*] unstable buildings and [$_{\text{EVENT}}$ *clear*] sites for future construction. wsj_0766

(18) Last summer, he [$_{\text{EVENT}}$ *chucked*] his 10-year career as a London stockbroker and [$_{\text{EVENT}}$ *headed*] for the mountains. wsj_0776

So, for example, getting Example 15 correct requires knowing that linking up with leaders usually precedes getting elected to an office. Likewise, getting Example 17 correct requires knowing that building sites are only cleared after the buildings are demolished. All of these examples introduce the same difficulty – surface level features like tense give no clue as to the relation. To be able to learn such relations, the models need access to some sort of information about the typical ordering of events.

Some of this information may become available simply by additional exposure to the various event words. Figure 4 shows the percent of events in the test data seen in the training data for varying amounts of training data. Logarithmic trendlines fit to these curves suggest that annotating the full 5,013 event pairs in the Penn TreeBank could move individual event coverage up to the mid 80s, meaning that most events encountered by the system would have been seen in the same position in the training data. So, additional annotation would at least partially remove the data sparsity problem, giving the system a better understanding of the individual events. However, there is still a clear need for measures that can suggest, for example, that *buy* typically precedes *sell*. Informative statistical measures of this kind will be crucial for providing the world knowledge necessary to identify temporal and causal relations.

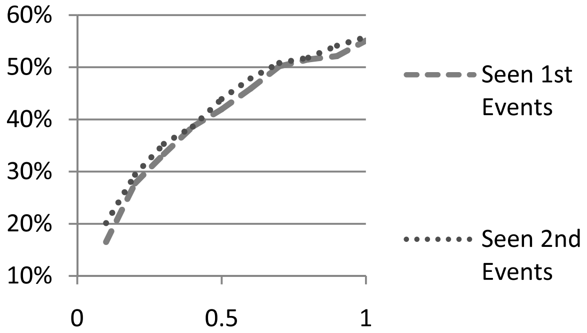


Figure 4: Percent of events in the test data seen during training given increasing fractions of the training data.

8. Conclusions

We designed a corpus of parallel temporal and causal relations to fill a gap in the temporal-causal structure annotated by existing resources like PropBank, TimeBank and the Penn Discourse TreeBank. We selected 1000 event pairs conjoined by the word *and*, and annotated them for temporal and causal relations. BEFORE and AFTER temporal relations were annotated using an extension of the TempEval guidelines, and CAUSAL relations were annotated using a set of causal and non-causal paraphrases for the word *and*. Annotators were able to achieve substantial agreement, 81.2%, for temporal relations, and moderate agreement, 77.8%, for causal relations.

Analysis of the corpus revealed some interesting interactions between temporal and causal relations. Over 30% of causal relations were not accompanied by an underlying BEFORE relation, even though causes are expected to precede effects. This suggests that additional work on temporal and causal annotation schemes may be helpful to design a single cohesive theory about how temporal and causal relations interact. Study of the corpus also revealed that simple surface features like tense help little in identifying temporal and causal relations for conjoined events. Machine learning experiments confirmed this finding, though the support vector machine models trained on the surface features were able to outperform all baselines they were compared against.

Future work will consider a more in-depth analysis of the corpus and the relation between temporal and causal structures. The results of this analysis should identify useful semantic clues to the presence of a temporal or causal relation, and thus offer the opportunity to improve the performance of machine learning models.

Acknowledgments

This research was performed under an appointment of the first author to the DHS Scholarship and Fellowship Program, administered by ORISE through an interagency agreement between the U.S. 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.

9. References

Brandon Beamer, Suma Bhat, Brant Chee, Andrew Fister, Alla Rozovskaya, and Roxana Girju. 2007. Uiuc: A knowledge-rich approach to identifying semantic relations between nominals. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

Steven Bethard and James H. Martin. 2006. Identification of event mentions and their semantic class. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Steven Bethard and James H. Martin. 2007. CU-TMP: Temporal relation classification using syntactic and semantic features. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *International Conference on Semantic Computing (ICSC)*.

Branimir Boguraev and Rie Kubota 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.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*.

Christopher S. G. Khoo, Jaklin Kornfilt, Robert N. Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186.

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Language Resources and Evaluation*.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Human Language Technologies and North American Chapter of the Assocation of Computational Linguistics (HLT-NAACL) Workshop on Frontiers in Corpus Annotation*.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Georgiana Puşcaşu. 2007. Wvali: Temporal relation identification by syntactico-semantic analysis. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

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

David 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.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.