An Annotation Framework for Dense Event Ordering

Taylor Cassidy  
Army Research Lab  
IBM Research  
taylor.cassidy.ctr@mail.mil

Bill McDowell  
Carnegie Mellon University  
forkunited@gmail.com

Nathanael Chambers  
US Naval Academy  
nchamber@usna.edu

Steven Bethard  
Univ. of Alabama at Birmingham  
bethard@cis.uab.edu

Abstract

Today’s event ordering research is heavily dependent on annotated corpora. Current corpora influence shared evaluations and drive algorithm development. Partly due to this dependence, most research focuses on partial orderings of a document’s events. For instance, the TempEval competitions and the TimeBank only annotate small portions of the event graph, focusing on the most salient events or on specific types of event pairs (e.g., only events in the same sentence). Deeper temporal reasoners struggle with this sparsity because the entire temporal picture is not represented. This paper proposes a new annotation process with a mechanism to force annotators to label connected graphs. It generates 10 times more relations per document than the TimeBank, and our TimeBank-Dense corpus is larger than all current corpora. We hope this process and its dense corpus encourages research on new global models with deeper reasoning.

1 Introduction

The TimeBank Corpus (Pustejovsky et al., 2003) ushered in a wave of data-driven event ordering research. It provided a common dataset of relations between events and time expressions that allowed the community to compare approaches. Later corpora and competitions have based their tasks on the TimeBank setup. This paper addresses one of its shortcomings: sparse annotation. We describe a new annotation framework (and a TimeBank-Dense corpus) that we believe is needed to fulfill the data needs of deeper reasoners.
There were four or five people inside, and they just started firing.

Ms. Sanders was hit several times and was pronounced dead at the scene.

The other customers fled, and the police said it did not appear that anyone else was injured.

There were four or five people inside, and they just started firing.

Ms. Sanders was hit several times and was pronounced dead at the scene.

The other customers fled, and the police said it did not appear that anyone else was injured.

— Bethard, 2013

We describe the first annotation framework that forces annotators to annotate all pairs. With this new process, we created a dense ordering of document events that can properly evaluate both relation identification and relation annotation. Figure 1 illustrates one document before and after our new annotations.

### 2 Previous Annotation Work

The majority of corpora and competitions for event ordering contain sparse annotations. Annotators for the original TimeBank (Pustejovsky et al., 2003) only annotated relations judged to be salient by the annotator. Subsequent TempEval competitions (Verhagen et al., 2007; Verhagen et al., 2010; Uz-Zaman et al., 2013) mostly relied on the TimeBank, but also aimed to improve coverage by annotating relations between all events and times in the same sentence. However, event tokens that were mentioned fewer than 20 times were excluded and only one TempEval task considered relations between events in different sentences. In practical terms, the resulting evaluations remained sparse.

A major dilemma underlying these sparse tasks is that the unlabeled event/time pairs are ambiguous. Each unlabeled pair holds 3 possibilities:

1. The annotator looked at the pair of events and decided that no temporal relation exists.
2. The annotator did not look at the pair of events, so a relation may or may not exist.
3. The annotator failed to look at the pair of events, so a single relation may exist.

Training and evaluation of temporal reasoners is hampered by this ambiguity. To combat this, our annotation adopts the VAGUE relation introduced by TempEval 2007, and our approach forces annotators to use it. This is the only work that includes such a mechanism.

This paper is not the first to look into more dense annotations. Bramsen et al. (2006) annotated multi-sentence segments of text to build directed acyclic graphs. Kolomiyets et al. (2012) annotated “temporal dependency structures”, though they only focused on relations between pairs of events. Do et al. (2012) produced the densest annotation, but “the annotator was not required to annotate all pairs of event mentions, but as many as possible”. The current paper takes a different tack to annotation by requiring annotators to label every possible pair of events/times in a given window. Thus this work is the first annotation effort that can guarantee its event/time graph to be strongly connected.

Table 1 compares the size and density of our corpus to others. Ours is the densest and it contains the largest number of temporal relations.

---

Table 1: Events, times, relations and the ratio of relations to events + times (R) in various corpora.

<table>
<thead>
<tr>
<th></th>
<th>Events</th>
<th>Times</th>
<th>Rel</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TimeBank</td>
<td>7935</td>
<td>1414</td>
<td>6418</td>
<td>0.7</td>
</tr>
<tr>
<td>Bramsen 2006</td>
<td>627</td>
<td>–</td>
<td>615</td>
<td>1.0</td>
</tr>
<tr>
<td>TempEval-07</td>
<td>6832</td>
<td>1249</td>
<td>5790</td>
<td>0.7</td>
</tr>
<tr>
<td>TempEval-10</td>
<td>5688</td>
<td>2117</td>
<td>4907</td>
<td>0.6</td>
</tr>
<tr>
<td>TempEval-13</td>
<td>11145</td>
<td>2078</td>
<td>11098</td>
<td>0.8</td>
</tr>
<tr>
<td>Kolomiyets-12</td>
<td>1233</td>
<td>–</td>
<td>1139</td>
<td>0.9</td>
</tr>
<tr>
<td>Do 2012²</td>
<td>324</td>
<td>232</td>
<td>3132</td>
<td>5.6</td>
</tr>
<tr>
<td><strong>This work</strong></td>
<td>1729</td>
<td>289</td>
<td><strong>12715</strong></td>
<td><strong>6.3</strong></td>
</tr>
</tbody>
</table>

---

¹As discussed below, all pairs in a given window size.

²Do et al. (2012) reports 6264 relations, but this includes both the relations and their inverses. We thus halve the count for accurate comparison to other corpora.
3 A Framework for Dense Annotation

Frameworks for annotating text typically have two independent facets: (1) the practical means of how to label the text, and (2) the higher-level rules about when something should be labeled. The first is often accomplished through a markup language, and we follow prior work in adopting TimeML here. The second facet is the focus of this paper: when should an annotator label an ordering relation?

Our proposal starts with documents that have already been annotated with events, time expressions, and document creation times (DCT). The following sentence serves as our motivating example:

*Police confirmed Friday that the body found along a highway in this municipality 15 miles south of San Juan belonged to Jorge Hernandez.*

This sentence is represented by a 4 node graph (3 events and 1 time). In a completely annotated graph it would have 6 edges (relations) connecting the nodes. In the TimeBank, from which this sentence is drawn, only 3 of the 6 edges are labeled.

The impact of these annotation decisions (i.e., when to annotate a relation) can be significant. In this example, a learner must somehow deal with the 3 unlabeled edges. One option is to assume that they are vague or ambiguous. However, all 6 edges have clear well-defined ordering relations:

- `belonged BEFORE confirmed`
- `found BEFORE confirmed`
- `belonged BEFORE Friday`
- `confirmed IS INCLUDED IN Friday`
- `found IS INCLUDED IN Friday`

Learning algorithms handle these unlabeled edges by making incorrect assumptions, or by ignoring large parts of the temporal graph. Several models with rich temporal reasoners have been published, but since they require more connected graphs, improvement over pairwise classifiers have been minimal (Chambers and Jurafsky, 2008; Yoshikawa et al., 2009). This paper thus proposes an annotation process that builds denser graphs with formal properties that learners can rely on, such as locally complete subgraphs.

3.1 Ensuring Dense Graphs

While the ideal goal is to create a complete graph, the time it would take to hand-label \( n(n-1)/2 \) edges is prohibitive. We approximate completeness by creating locally complete graphs over neighboring sentences. The resulting event graph for a document is strongly connected, but not complete. Specifically, the following edge types are included:

1. Event-Event, Event-Time, and Time-Time pairs in the same sentence
2. Event-Event, Event-Time, and Time-Time pairs between the current and next sentence
3. Event-DCT pairs for every event in the text
4. Time-DCT pairs for every time expression in the text

Our process requires annotators to annotate all of the above edge types, and we enforce this via an annotation tool. We describe the target relation set and this tool next.

3.1.1 Temporal Relations

The TimeBank corpus uses 14 relations based on the Allen interval relations. The TempEval contests have used a small set of relations (TempEval-1) and the larger set of 14 relations (TempEval-3). Published work has mirrored this trend, and different groups focus on different aspects of the semantics.

We chose a middle ground between coarse and fine-grained distinctions for annotation, settling on 6 relations: before, after, includes, is included, simultaneous, and vague. We do not adopt a more fine-grained set because we annotate pairs that are far more ambiguous than those considered in previous efforts. Decisions between relations like *before* and *immediately before* can complicate an already difficult task. The added benefit of a corpus (or working system) that makes fine-grained distinctions is also not clear. We lean toward higher annotator agreement with relations that have greater separation between their semantics.

3.1.2 Enforcing Annotation

Imposing the above rules on annotators requires automated assistance. We built a new tool that reads TimeML formatted text, and computes the set of required edges. Annotators are prompted to assign a label for each edge, and skipping edges is prohibited. The tool is unique in that it includes a transitive reasoner that infers relations based on

4 Note that annotators are presented with pairs in order from document start to finish, starting with the first two events.

5 For instance, a relation like *starts* is a special case of *includes* if events are viewed as open intervals, and *immediately before* is a special case of *before*. We avoid this overlap and only use *includes* and *before*
the annotator’s latest annotations. For example, if event \( e_1 \) is included in \( t_1 \), and \( t_1 \) before \( e_2 \), the tool automatically labels \( e_1 \) before \( e_2 \). The transitivity inference is run after each input label, and the human annotator cannot override the inferences. This prohibits the annotator from entering edges that break transitivity. As a result, several properties are ensured through this process: the graph (1) is a strongly connected graph, (2) is consistent with no contradictions, and (3) has all required edges labeled. These 3 properties are new to all current ordering corpora.

### 3.2 Annotation Guidelines

Since the annotation tool frees the annotators from the decision of when to label an edge, the focus is now what to label each edge. This section describes the guidelines for dense annotation.

**The 80% confidence rule:** The decision to label an edge as VAGUE instead of a defined temporal relation is critical. We adopted an 80% rule that instructed annotators to choose a specific non-vague relation if they are 80% confident that it was the writer’s intent that a reader infer that relation. By not requiring 100% confidence, we allow for alternative interpretations that conflict with the chosen edge label as long as that alternative is sufficiently unlikely. In practice, annotators had different interpretations of what constitutes 80% certainty, and this generated much discussion. We mitigated these disagreements with the following rule.

**Majority annotator agreement:** An edge’s final label is the relation that received a majority of annotator votes, otherwise it is marked VAGUE. If a document has 2 annotators, both have to agree on the relation or it is labeled VAGUE. A document with 3 annotators requires 2 to agree, and 4 annotators require 3 to agree. This agreement rule acts as a check to our 80% confidence rule, backing off to VAGUE when decisions are uncertain (arguably, this is the definition of VAGUE).

In addition to the above, we created other guidelines to encourage consistent labelings. Several of these are inspired by Bethard and Martin (2008).

**Modal and conditional events:** interpreted with a possible worlds analysis. The core event was treated as having occurred, whether or not the text implied that it had occurred. For example,

They \[EVENT expect\] him to \[EVENT cut\] costs throughout the organization.

This event pair is ordered (expect before cut) since the expectation occurs before the cutting (in the possible world where the cutting occurs). Negated events and hypotheticals are treated similarly. One assumes the event does occur, and all other events are ordered accordingly. Negated states like “is not anticipating” are interpreted as though the anticipation occurs, and surrounding events are ordered with regard to its presumed temporal span.

**Aspectual Events:** annotated as IS INCLUDED in their event arguments. For instance, events that describe the manner in which another event is performed are considered encompassed by the broader event. Consider the following example:

The move may \[EVENT help\] \[EVENT prevent\] Martin Ackerman from making a run at the computer-services concern.

This event pair is assigned the relation (help is included in prevent) because the help event is not meaningful on its own. It describes the proportion of the preventing accounted for by the move. In TimeBank, the intentional action class is used instead of the aspectual class in this case, but we still consider it covered by this guideline.

**Events that attribute a property:** to a person or event are interpreted to end when the entity ends. For instance, ‘the talk is nonsense’ evokes a nonsense event with an end point that coincides with the end of the talk.

**Time Expressions:** the words now and today were given “long now” interpretations if the words could be replaced with nowadays and not change the meaning of their sentences. The time’s duration starts sometime in the past and includes the DCT. If nowadays is not suitable, then the now was included in the DCT.

**Generic Events:** can be ordered with respect to each other, but must be VAGUE with respect to nearby non-generic events.

### 4 TimeBank-Dense: corpus statistics

We chose a subset of TimeBank documents for our new corpus: **TimeBank-Dense**. This provided an initial labeling of events and time expressions. Using the tool described above, we annotated 36 random documents with at least two annotators each. These 36 were annotated with 4 times as many relations as the entire 183 document TimeBank.

The four authors of this paper were the four annotators. All four annotated the same initial docu-
ment, conflicts and disagreements were discussed, and guidelines were updated accordingly. The rest of the documents were then annotated independently. Document annotation was not random, but we mixed pairs of authors where time constraints allowed. Table 2 shows the relation counts in the final corpus, and Table 3 gives the annotator agreement. We show precision (holding one annotation as gold) and kappa computed on the 4 types of pairs from section 3.1. Micro-averaged precision was 65.1%, compared to TimeBank’s 77%. Kappa ranged from .56-.64, a slight drop from TimeBank’s .71.

The vague relation makes up 46% of the relations. This is the first empirical count of how many temporal relations in news articles are truly vague. Our lower agreement is likely due to the more difficult task. Table 5 breaks down the individual disagreements. The most common disagreements pertained to the VAGUE relation. Practically speaking, VAGUE was applied to the final graph if either annotator chose it. This seems appropriate since a disagreement between two annotators implies that the relation is vague.

The following example illustrates the difficulty of labeling edges with a VAGUE relation:

No one was hurt, but firefighters ordered the evacuation of nearby homes and said they’ll monitor the shifting ground.

Both annotators chose VAGUE to label ordered and said because the order is unclear. However, they disagreed on evacuation with monitor. One chose VAGUE, but the other chose IS INCLUDED. There is a valid interpretation where a monitoring process has already begun, and continues after the evacuation. This interpretation reached 80% confidence for one annotator, but not the other. In the face of such a disagreement, the pair is labeled VAGUE.

How often do these disagreements occur? Table 4 shows the 3 sources: (1) mutual vague: annotators agree it is vague, (2) partial vague: one annotator chooses vague, but the other does not, and (3) no vague: annotators choose conflicting non-vague relations. Only 17% of these disagreements are due to hard conflicts (no vague). The released corpus includes these 3 fine-grained VAGUE relations.

5 Conclusion

We described our annotation framework that produces corpora with formal guarantees about the annotated graph’s structure. Both the annotation tool and the new TimeBank-Dense corpus are publicly available. This is the first corpus with guarantees of connectedness, consistency, and a semantics for unlabeled edges. We hope to encourage a shift in the temporal ordering community to consider the entire document when making local decisions. Further work is needed to handle difficult pairs with the VAGUE relation. We look forward to evaluating new algorithms on this dense corpus.

Acknowledgments

This work was supported, in part, by the Johns Hopkins Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors. We also give thanks to Benjamin Van Durme for assistance and insight.

---

<table>
<thead>
<tr>
<th>Annotated Relation Count</th>
<th>BEFORE</th>
<th>INCLUDES</th>
<th>AFTER</th>
<th>INCLUDED IN</th>
<th>SIMULTAN.</th>
<th>VAGUE</th>
<th>Total Relations: 12715</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2590</td>
<td>836</td>
<td>2104</td>
<td>1060</td>
<td>215</td>
<td>5910</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The number of relations of each type in the TimeBank-Dense corpus.

<table>
<thead>
<tr>
<th>Annotators</th>
<th># Links</th>
<th>Prec</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>A and B</td>
<td>9282</td>
<td>.65</td>
<td>.56</td>
</tr>
<tr>
<td>A and D</td>
<td>1605</td>
<td>.72</td>
<td>.63</td>
</tr>
<tr>
<td>B and D</td>
<td>279</td>
<td>.70</td>
<td>.64</td>
</tr>
<tr>
<td>C and D</td>
<td>1549</td>
<td>.65</td>
<td>.57</td>
</tr>
</tbody>
</table>

Table 3: Agreement between different annotators.

<table>
<thead>
<tr>
<th># Vague</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual VAGUE</td>
<td>1657 (28%)</td>
<td></td>
</tr>
<tr>
<td>Partial VAGUE</td>
<td>3234 (55%)</td>
<td></td>
</tr>
<tr>
<td>No VAGUE</td>
<td>1019 (17%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: VAGUE relation origins. Partial vague: one annotator does not choose vague. No vague: neither annotator chooses vague.

<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
<th>i</th>
<th>ii</th>
<th>s</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>22</td>
<td>88</td>
<td>37</td>
<td>21</td>
<td>192</td>
</tr>
<tr>
<td>1444</td>
<td>52</td>
<td>102</td>
<td>9</td>
<td>135</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>34</td>
<td>642</td>
<td>45</td>
<td>23</td>
<td>191</td>
</tr>
<tr>
<td>81</td>
<td>76</td>
<td>40</td>
<td>826</td>
<td>31</td>
<td>230</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>25</td>
<td>28</td>
<td>147</td>
<td>29</td>
</tr>
<tr>
<td>500</td>
<td>441</td>
<td>289</td>
<td>336</td>
<td>64</td>
<td>1197</td>
</tr>
</tbody>
</table>

Table 5: Relation agreement between the two main annotators. Most disagreements involved whether or not a relation should be VAGUE.

---

6http://www.usna.edu/Users/cs/nchamber/caevo/
References


