An Annotation Framework for Dense Event Ordering

Taylor Cassidy
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 for 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.

The TimeBank includes a small subset of all possible relations in its documents. The annotators were instructed to label relations critical to the document's understanding. The result is a sparse labeling that leaves much of the document unlabeled. The TempEval contests have largely followed suit and focused on specific types of event pairs. For instance, TempEval (Verhagen et al., 2007) only labeled relations between events that syntactically dominated each other. This paper is the first attempt to annotate a document's entire temporal graph.

A consequence of focusing on all relations is a shift from the traditional *classification* task, where the system is given a pair of events and asked only to label the type of relation, to an *identification* task, where the system must determine for itself which events in the document to pair up. For example, in TempEval-1 and 2 (Verhagen et al., 2007; Verhagen et al., 2010), systems were given event pairs in specific syntactic positions: events and times in the same noun phrase, main events in consecutive sentences, etc. We now aim for a shift in the community wherein all pairs are considered candidates for temporal ordering, allowing researchers to ask questions such as: how must algorithms adapt to label the complete graph of pairs, and if the more difficult and ambiguous event pairs are included, how must feature-based learners change?

We are not the first to propose these questions, but this paper is the first to directly propose the means by which they can be addressed. The stated goal of TempEval-3 (UzZaman et al., 2013) was to focus on relation identification instead of classification, but the training and evaluation data followed the TimeBank approach where only a subset of event pairs were labeled. As a result, many systems focused on classification, with the top system classifying pairs in only three syntactic constructions

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 501–506,
Baltimore, Maryland, USA, June 23-25 2014.

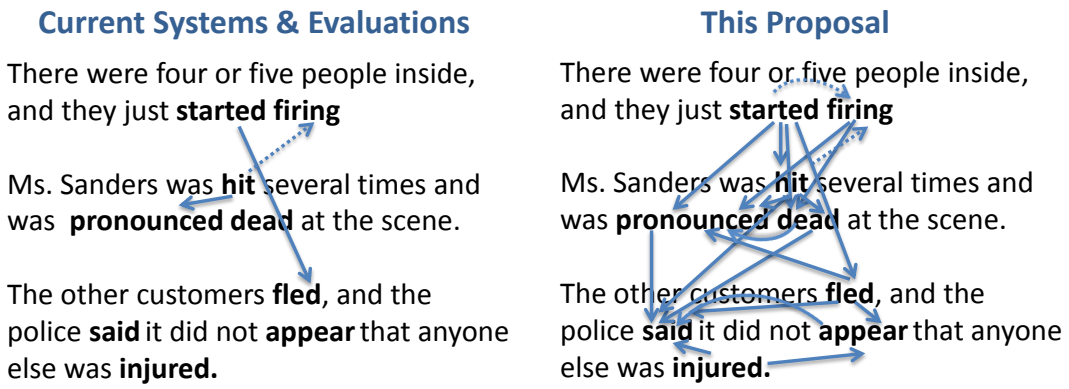

Figure 1: A TimeBank annotated document is on the left, and this paper's TimeBank-Dense annotation is on the right. Solid arrows indicate BEFORE relations and dotted arrows indicate INCLUDED_IN relations.

(Bethard, 2013). We describe the first annotation framework that forces annotators to annotate all pairs[1]. 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; UzZaman 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.

	Events	Times	Rels	R
TimeBank	7935	1414	6418	0.7
Bramsen 2006	627	–	615	1.0
TempEval-07	6832	1249	5790	0.7
TempEval-10	5688	2117	4907	0.6
TempEval-13	11145	2078	11098	0.8
Kolomiyets-12	1233	–	1139	0.9
Do 2012[2]	324	232	3132	5.6
This work	1729	289	**12715**	**6.3**

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

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.

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

[2] Do et al. (2012) reports 6264 relations, but this includes both the relations and their inverses. We thus halve the count

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 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*
> *belonged* BEFORE *found*
> *found* BEFORE *confirmed*
> *belonged* BEFORE *Friday*
> *confirmed* IS INCLUDED IN *Friday*
> *found* IS INCLUDED IN *Friday*[3]

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 the above edge types, enforced via an annotation tool. We describe the 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[4].

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.[5] The tool is unique in that it includes a transitive reasoner that infers relations based on the annotator's latest annotations. For example,

for accurate comparison to other corpora.

[3] Revealed by the previous sentence (not shown here).

[4] 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*

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

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 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. 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).

We also encouraged consistent labelings with guidelines 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 document, conflicts and disagreements were discussed,

Annotated Relation Count

BEFORE	2590	INCLUDES	836
AFTER	2104	INCLUDED IN	1060
SIMULTAN.	215	VAGUE	5910
Total Relations: 12715			

Table 2: Relation counts in TimeBank-Dense.

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 frequent 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 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 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.

Annotators	# Links	Prec	Kappa
A and B	9282	.65	.56
A and D	1605	.72	.63
B and D	279	.70	.64
C and D	1549	.65	.57

Table 3: Agreement between different annotators.

	# Vague
Mutual VAGUE	1657 (28%)
Partial VAGUE	3234 (55%)
No VAGUE	1019 (17%)

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

	b	**a**	**i**	**ii**	**s**	**v**
b	1776	22	88	37	21	192
a	17	1444	32	102	9	155
i	71	34	642	45	23	191
ii	81	76	40	826	31	230
s	12	8	25	28	147	29
v	500	441	289	356	64	1197

Table 5: Relation agreement between the two main annotators. Most disagreements involved VAGUE.

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.[6] 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.

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

References

Steven Bethard, William J Corvey, Sara Klingenstein, and James H Martin. 2008. Building a corpus of temporal-causal structure. In *LREC*.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

P. Bramsen, P. Deshpande, Y.K. Lee, and R. Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198. ACL.

N. Chambers and D. Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 698–706. ACL.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687, Jeju Island, Korea, July. Association for Computational Linguistics.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea, July. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.

K. Yoshikawa, S. Riedel, M. Asahara, and Y. Matsumoto. 2009. Jointly identifying temporal relations with Markov Logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 405–413. ACL.