A Multi-Axis Annotation Scheme for Event Temporal Relations

Qiang Ning,[1] **Hao Wu,**[2] **Dan Roth**[1,2]
Department of Computer Science
[1]University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]University of Pennsylvania, Philadelphia, PA 19104, USA
qning2@illinois.edu, {haowu4,danroth}@seas.upenn.edu

Abstract

Existing temporal relation (TempRel) annotation schemes often have low inter-annotator agreements (IAA) even between experts, suggesting that the current annotation task needs a better definition. This paper proposes a new multi-axis modeling to better capture the temporal structure of events. In addition, we identify that event end-points are a major source of confusion in annotation, so we also propose to annotate TempRels based on start-points only. A pilot expert annotation effort using the proposed scheme shows significant improvement in IAA from the conventional 60's to 80's (Cohen's Kappa). This better-defined annotation scheme further enables the use of crowdsourcing to alleviate the labor intensity for each annotator. We hope that this work can foster more interesting studies towards event understanding.[1]

1 Introduction

Temporal relation (TempRel) extraction is an important task for event understanding, and it has drawn much attention in the natural language processing (NLP) community recently (UzZaman et al., 2013; Chambers et al., 2014; Llorens et al., 2015; Minard et al., 2015; Bethard et al., 2015, 2016, 2017; Leeuwenberg and Moens, 2017; Ning et al., 2017, 2018a,b).

Initiated by TimeBank (TB) (Pustejovsky et al., 2003b), a number of TempRel datasets have been collected, including but not limited to the verb-clause augmentation to TB (Bethard et al., 2007), TempEval1-3 (Verhagen et al., 2007, 2010; UzZaman et al., 2013), TimeBank-Dense (TB-Dense) (Cassidy et al., 2014), EventTimeCorpus (Reimers et al., 2016), and datasets with both temporal and other types of relations (e.g., coreference and causality) such as CaTeRs (Mostafazadeh et al., 2016) and RED (O'Gorman et al., 2016). These datasets were annotated by experts, but most still suffered from low inter-annotator agreements (IAA). For instance, the IAAs of TB-Dense, RED and THYME-TimeML (Styler IV et al., 2014) were only below or near 60% (given that events are already annotated). Since a low IAA usually indicates that the task is difficult even for humans (see Examples 1-3), the community has been looking into ways to simplify the task, by reducing the label set, and by breaking up the overall, complex task into subtasks (e.g., getting agreement on which event pairs should have a relation, and then what that relation should be) (Mostafazadeh et al., 2016; O'Gorman et al., 2016). In contrast to other existing datasets, Bethard et al. (2007) achieved an agreement as high as 90%, but the scope of its annotation was narrowed down to a very special verb-clause structure.

(*e1*, *e2*), (*e3*, *e4*), and (*e5*, *e6*): TempRels that are difficult even for humans. Note that only relevant events are highlighted here.
Example 1: Serbian police tried to eliminate the pro-independence Kosovo Liberation Army and (***e1:restore***) order. At least 51 people were (***e2:killed***) in clashes between Serb police and ethnic Albanians in the troubled region.
Example 2: Service industries (***e3:showed***) solid job gains, as did manufacturers, two areas expected to be hardest (***e4:hit***) when the effects of the Asian crisis hit the American economy.
Example 3: We will act again if we have evidence he is (***e5:rebuilding***) his weapons of mass destruction capabilities, senior officials say. In a bit of television diplomacy, Iraq's deputy foreign minister (***e6:responded***) from Baghdad in less than one hour, saying that ...

This paper proposes a new approach to handling

[1]The dataset is publicly available at https://cogcomp.org/page/publication_view/834.

these issues in TempRel annotation. **First**, we introduce *multi-axis modeling* to represent the temporal structure of events, based on which we anchor events to different semantic axes; only events from the same axis will then be temporally compared (Sec. 2). As explained later, those event pairs in Examples 1-3 are difficult because they represent different semantic phenomena and belong to different axes. **Second**, while we represent an event pair using two time intervals (say, $[t^1_{start}, t^1_{end}]$ and $[t^2_{start}, t^2_{end}]$), we suggest that comparisons involving end-points (e.g., t^1_{end} vs. t^2_{end}) are typically more difficult than comparing start-points (i.e., t^1_{start} vs. t^2_{start}); we attribute this to the ambiguity of expressing and perceiving durations of events (Coll-Florit and Gennari, 2011). We believe that this is an important consideration, and we propose in Sec. 3 that TempRel annotation should focus on start-points. Using the proposed annotation scheme, a pilot study done by experts achieved a high IAA of .84 (Cohen's Kappa) on a subset of TB-Dense, in contrast to the conventional 60's.

In addition to the low IAA issue, TempRel annotation is also known to be labor intensive. Our **third contribution** is that we facilitate, for the first time, the use of crowdsourcing to collect a new, high quality (under multiple metrics explained later) TempRel dataset. We explain how the crowdsourcing quality was controlled and how *vague* relations were handled in Sec. 4, and present some statistics and the quality of the new dataset in Sec. 5. A baseline system is also shown to achieve much better performance on the new dataset, when compared with system performance in the literature (Sec. 6). The paper's results are very encouraging and hopefully, this work would significantly benefit research in this area.

2 Temporal Structure of Events

Given a set of events, one important question in designing the TempRel annotation task is: which pairs of events should have a relation? The answer to it depends on the modeling of the overall temporal structure of events.

2.1 Motivation

TimeBank (Pustejovsky et al., 2003b) laid the foundation for many later TempRel corpora, e.g., (Bethard et al., 2007; UzZaman et al., 2013; Cassidy et al., 2014).[2] In TimeBank, the annotators were allowed to label TempRels between any pairs of events. This setup models the overall structure of events using *a general graph*, which made annotators inadvertently overlook some pairs, resulting in low IAAs and many false negatives.

Example 4: Dense Annotation Scheme.
Serbian police (***e7:tried***) to (***e8:eliminate***) the pro-independence Kosovo Liberation Army and (***e1:restore***) order. At least 51 people were (***e2:killed***) in clashes between Serb police and ethnic Albanians in the troubled region.
Given 4 NON-GENERIC events above, the dense scheme presents 6 pairs to annotators one by one: (*e7*, *e8*), (*e7*, *e1*), (*e7*, *e2*), (*e8*, *e1*), (*e8*, *e2*), and (*e1*, *e2*). Apparently, not all pairs are well-defined, e.g., (*e8*, *e2*) and (*e1*, *e2*), but annotators are forced to label all of them.

To address this issue, Cassidy et al. (2014) proposed a dense annotation scheme, TB-Dense, which annotates all event pairs within a sliding, two-sentence window (see Example 4). It requires all TempRels between GENERIC[3] and NON-GENERIC events to be labeled as *vague*, which conceptually models the overall structure by *two disjoint time-axes*: one for the NON-GENERIC and the other one for the GENERIC.

However, as shown by Examples 1-3 in which the highlighted events are NON-GENERIC, the TempRels may still be ill-defined: In Example 1, Serbian police tried to restore order but ended up with conflicts. It is reasonable to argue that the attempt to ***e1:restore*** order happened *before* the conflict where 51 people were ***e2:killed***; or, 51 people had been ***killed*** but order had not been ***restored*** yet, so ***e1:restore*** is *after* ***e2:killed***. Similarly, in Example 2, service industries and manufacturers were originally expected to be hardest ***e4:hit*** but actually ***e3:showed*** gains, so ***e4:hit*** is *before* ***e3:showed***; however, one can also argue that the two areas had ***showed*** gains but had not been ***hit***, so ***e4:hit*** is *after* ***e3:showed***. Again, ***e5:rebuilding*** is a hypothetical event: "we will act if ***rebuilding*** is true". Readers do not know for sure if "he is already rebuilding weapons but we have no evidence", or "he will be building weapons in the future", so annotators may disagree on the relation between ***e5:rebuilding*** and ***e6:responded***. Despite, importantly, minimizing missing annota-

[2] EventTimeCorpus (Reimers et al., 2016) is based on TimeBank, but aims at anchoring events onto explicit time expressions in each document rather than annotating TempRels between events, which can be a good complementary to other TempRel datasets.

[3] For example, *lions eat meat* is GENERIC.

tions, the current dense scheme forces annotators to label many such ill-defined pairs, resulting in low IAA.

2.2 Multi-Axis Modeling

Arguably, an ideal annotator may figure out the above ambiguity by him/herself and mark them as *vague*, but it is not a feasible requirement for all annotators to stay clear-headed for hours; let alone crowdsourcers. What makes things worse is that, after annotators spend a long time figuring out these difficult cases, whether they disagree with each other or agree on the vagueness, the final decisions for such cases will still be *vague*.

As another way to handle this dilemma, TB-Dense resorted to a 80% confidence rule: annotators were allowed to choose a label if one is 80% sure that it was the writer's intent. However, as pointed out by TB-Dense, annotators are likely to have rather different understandings of 80% confidence and it will still end up with disagreements.

In contrast to these annotation difficulties, humans can easily grasp the meaning of news articles, implying a potential gap between the difficulty of the annotation task and the one of understanding the actual meaning of the text. In Examples 1-3, the writers did not intend to explain the TempRels between those pairs, and the original annotators of TimeBank[4] did not label relations between those pairs either, which indicates that both writers and readers did not think the TempRels between these pairs were crucial. Instead, what is crucial in these examples is that "Serbian police ***tried*** to restore order but ***killed*** 51 people", that "two areas were ***expected*** to be hit but ***showed*** gains", and that "***if*** he rebuilds weapons ***then*** we will act." To "***restore*** order", to be "hardest ***hit***", and "if he was ***rebuilding***" were only the intention of police, the opinion of economists, and the condition to ***act***, respectively, and whether or not they actually happen is not the focus of those writers.

This discussion suggests that a single axis is too restrictive to represent the complex structure of NON-GENERIC events. Instead, we need a modeling which is more restrictive than a general graph so that annotators can focus on relation annotation (rather than looking for pairs first), but also more flexible than a single axis so that ill-defined relations are not forcibly annotated. Specifically, we need axes for intentions, opinions, hypotheses, etc. in addition to the main axis of an article. We thus argue for *multi-axis modeling*, as defined in Table 1. Following the proposed modeling, Examples 1-3 can be represented as in Fig. 1. This modeling aims at capturing what the author has explicitly expressed and it only asks annotators to look at comparable pairs, rather than forcing them to make decisions on often vaguely defined pairs.

Event Type	Category
INTENTION, OPINION	On an orthogonal axis
HYPOTHESIS, GENERIC	On a parallel axis
NEGATION	Not on any axis
STATIC, RECURRENT	Other

Table 1: The interpretation of various event types that are not on the main axis in the proposed multi-axis modeling. The names are rather straightforward; see examples for each in Appendix A.

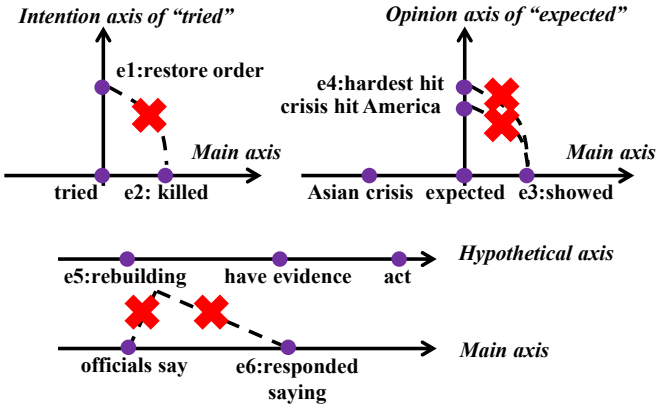

Figure 1: A multi-axis view of Examples 1-3. Only events on the same axis are compared.

In practice, we annotate one axis at a time: we first classify if an event is anchorable onto a given axis (this is also called the anchorability annotation step); then we annotate every pair of anchorable events (i.e., the relation annotation step); finally, we can move to another axis and repeat the two steps above. Note that ruling out cross-axis relations is only a strategy we adopt in this paper to separate well-defined relations from ill-defined relations. We do not claim that cross-axis relations are unimportant; instead, as shown in Fig. 2, we think that cross-axis relations are a different semantic phenomenon that requires additional investigation.

[4]Recall that they were given the entire article and only salient relations would be annotated.

2.3 Comparisons with Existing Work

There have been other proposals of temporal structure modelings (Bramsen et al., 2006; Bethard et al., 2012), but in general, the semantic phenomena handled in our work are very different and complementary to them. (Bramsen et al., 2006) introduces "temporal segments" (a fragment of text that does not exhibit abrupt changes) in the medical domain. Similarly, their temporal segments can also be considered as a special temporal structure modeling. But a key difference is that (Bramsen et al., 2006) only annotates inter-segment relations, ignoring intra-segment ones. Since those segments are usually large chunks of text, the semantics handled in (Bramsen et al., 2006) is in a very coarse granularity (as pointed out by (Bramsen et al., 2006)) and is thus different from ours.

(Bethard et al., 2012) proposes a tree structure for children's stories, which "typically have simpler temporal structures", as they pointed out. Moreover, in their annotation, an event can only be linked to a single nearby event, even if multiple nearby events may exist, whereas we do not have such restrictions.

In addition, some of the semantic phenomena in Table 1 have been discussed in existing work. Here we compare with them for a better positioning of the proposed scheme.

2.3.1 Axis Projection

TB-Dense handled the incomparability between main-axis events and HYPOTHESIS/NEGATION by *treating an event as having occurred* if the event is HYPOTHESIS/NEGATION.[5] In our multi-axis modeling, the strategy adopted by TB-Dense falls into a more general approach, "axis projection". That is, projecting events across different axes to handle the incomparability between any two axes (not limited to HYPOTHESIS/NEGATION). Axis projection works well for certain event pairs like ***Asian crisis*** and ***e4:hardest hit*** in Example 2: as in Fig. 1, ***Asian crisis*** is *before* ***expected***, which is again *before* ***e4:hardest hit***, so ***Asian crisis*** is *before* ***e4:hardest hit***.

Generally, however, since there is no direct evidence that can guide the projection, annotators may have different projections (imagine projecting ***e5:rebuilding*** onto the main axis: is it in the past or in the future?). As a result, axis projection requires many specially designed guidelines or strong external knowledge. Annotators have to rigidly follow the sometimes counter-intuitive guidelines or "guess" a label instead of looking for evidence in the text.

When strong external knowledge is involved in axis projection, it becomes a reasoning process and the resulting relations are a different type. For example, a reader may reason that in Example 3, it is well-known that they did "act again", implying his ***e5:rebuilding*** had happened and is *before* ***e6:responded***. Another example is in Fig. 2. It is obvious that relations based on these projections are not the same with and more challenging than those same-axis relations, so in the current stage, we should focus on same-axis relations only.

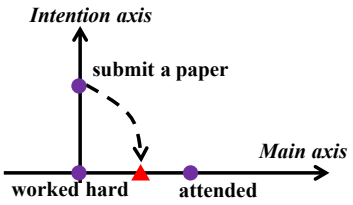

Figure 2: In *I worked hard to submit a paper . . . I attended the conference*, the projection of ***submit a paper*** onto the main axis is clearly *before* ***attended***. However, this projection requires strong external knowledge that a paper should be submitted before attending a conference. Again, this projection is only a guess based on our external knowledge and it is still open whether the paper is submitted or not.

2.3.2 Introduction of the Orthogonal Axes

Another prominent difference to earlier work is the introduction of orthogonal axes, which has not been used in any existing work as we know. A special property is that the intersection event of two axes can be compared to events from both, which can sometimes bridge events, e.g., in Fig. 1, ***Asian crisis*** is seemingly *before* ***hardest hit*** due to their connections to ***expected***. Since ***Asian crisis*** is on the main axis, it seems that ***e4:hardest hit*** is on the main axis as well. However, the "***hardest hit***" in "***Asian crisis*** *before* ***hardest hit***" is only a projection of the original ***e4:hardest hit*** onto the real axis and is valid only when this OPINION is true.

Nevertheless, OPINIONS are not always true and INTENTIONS are not always fulfilled. In Example 5, ***e9:sponsoring*** and ***e10:resolve*** are the opinions of the West and the speaker, respectively; whether or not they are true depends on the au-

[5] In the case of Example 3, it is to treat ***rebuilding*** as actually happened and then link it to ***responded***.

thors' implications or the readers' understandings, which is often beyond the scope of TempRel annotation.[6] Example 6 demonstrates a similar situation for INTENTIONS: when reading the sentence of ***e11:report***, people are inclined to believe that it is fulfilled. But if we read the sentence of ***e12:report***, we have reason to believe that it is not. When it comes to ***e13:tell***, it is unclear if everyone told the truth. The existence of such examples indicates that orthogonal axes are a better modeling for INTENTIONS and OPINIONS.

Example 5: Opinion events may not always be true.
He is ostracized by the West for (***e9:sponsoring***) terrorism.
We need to (***e10:resolve***) the deep-seated causes that have resulted in these problems.
Example 6: Intentions may not always be fulfilled.
A passerby called the police to (***e11:report***) the body.
A passerby called the police to (***e12:report***) the body. Unfortunately, the line was busy.
I asked everyone to (***e13:tell***) the truth.

2.3.3 Differences from Factuality

Event modality have been discussed in many existing event annotation schemes, e.g., Event Nugget (Mitamura et al., 2015), Rich ERE (Song et al., 2015), and RED. Generally, an event is classified as *Actual* or *Non-Actual*, a.k.a. factuality (Saurí and Pustejovsky, 2009; Lee et al., 2015).

The main-axis events defined in this paper seem to be very similar to *Actual* events, but with several important differences: **First**, future events are *Non-Actual* because they indeed have not happened, but they may be on the main axis. **Second**, events that are not on the main axis can also be *Actual* events, e.g., intentions that are fulfilled, or opinions that are true. **Third**, as demonstrated by Examples 5-6, identifying anchorability as defined in Table 1 is relatively easy, but judging if an event actually happened is often a high-level understanding task that requires an understanding of the entire document or external knowledge.

Interested readers are referred to Appendix B for a detailed analysis of the difference between *Anchorable* (onto the main axis) and *Actual* on a subset of RED.

3 Interval Splitting

All existing annotation schemes adopt the interval representation of events (Allen, 1984) and there are 13 relations between two intervals (for readers who are not familiar with it, please see Fig. 4 in the appendix). To reduce the burden of annotators, existing schemes often resort to a reduced set of the 13 relations. For instance, Verhagen et al. (2007) merged all the overlap relations into a single relation, *overlap*. Bethard et al. (2007); Do et al. (2012); O'Gorman et al. (2016) all adopted this strategy. In Cassidy et al. (2014), they further split *overlap* into *includes*, *included* and *equal*.

Let $[t^1_{start}, t^1_{end}]$ and $[t^2_{start}, t^2_{end}]$ be the time intervals of two events (with the implicit assumption that $t_{start} \leq t_{end}$). Instead of reducing the relations between two intervals, we try to explicitly compare the time points (see Fig. 3). In this way, the label set is simply *before, after* and *equal*,[7] while the expressivity remains the same. This interval splitting technique has also been used in (Raghavan et al., 2012).

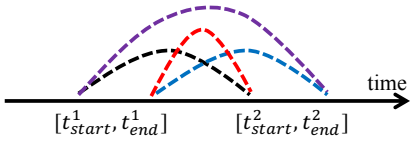

Figure 3: The comparison of two event time intervals, $[t^1_{start}, t^1_{end}]$ and $[t^2_{start}, t^2_{end}]$, can be decomposed into four comparisons t^1_{start} vs. t^2_{start}, t^1_{start} vs. t^2_{end}, t^1_{end} vs. t^2_{start}, and t^1_{end} vs. t^2_{end}, without loss of generality.

In addition to same expressivity, interval splitting can provide even more information when the relation between two events is *vague*. In the conventional setting, imagine that the annotators find that the relation between two events can be either *before* or *before and overlap*. Then the resulting annotation will have to be *vague*, although the annotators actually agree on the relation between t^1_{start} and t^2_{start}. Using interval splitting, however, such information can be preserved.

An obvious downside of interval splitting is the increased number of annotations needed (4 point comparisons vs. 1 interval comparison). In practice, however, it is usually much fewer than 4 comparisons. For example, when we see $t^1_{end} < t^2_{start}$ (as in Fig. 3), the other three can be skipped because they can all be inferred. Moreover, although the number of annotations is increased, the work load for human annotators may still be the same, because even in the conventional scheme, they still need to think of the relations between start- and

[6] For instance, there is undoubtedly a *causal* link between ***e9:sponsoring*** and ***ostracized***.

[7] We will discuss *vague* in Sec. 4.

end-points before they can make a decision.

3.1 Ambiguity of End-Points

During our pilot annotation, the annotation quality dropped significantly when the annotators needed to reason about relations involving end-points of events. Table 2 shows four metrics of task difficulty when only t^1_{start} vs. t^2_{start} or t^1_{end} vs. t^2_{end} are annotated. Non-anchorable events were removed for both jobs. The first two metrics, qualifying pass rate and survival rate are related to the two quality control protocols (see Sec. 4.1 for details). We can see that when annotating the relations between end-points, only one out of ten crowdsourcers (11%) could successfully pass our qualifying test; and even if they had passed it, half of them (56%) would have been kicked out in the middle of the task. The third line is the overall accuracy on gold set from all crowdsourcers (excluding those who did not pass the qualifying test), which drops from 67% to 37% when annotating end-end relations. The last line is the average response time per annotation and we can see that it takes much longer to label an end-end TempRel (52s) than a start-start TempRel (33s). This important discovery indicates that the TempRels between end-points is probably governed by a different linguistic phenomenon.

Metric	t^1_{start} vs. t^2_{start}	t^1_{end} vs. t^2_{end}
Qualification pass rate	50%	11%
Survival rate	74%	56%
Accuracy on gold	67%	37%
Avg. response time	33s	52s

Table 2: Annotations involving the end-points of events are found to be much harder than only comparing the start-points.

We hypothesize that the difficulty is a mixture of how durative events are expressed (by authors) and perceived (by readers) in natural language. In cognitive psychology, Coll-Florit and Gennari (2011) discovered that human readers take longer to perceive durative events than punctual events, e.g., *owe 50 bucks* vs. *lost 50 bucks*. From the writer's standpoint, durations are usually fuzzy (Schockaert and De Cock, 2008), or assumed to be a prior knowledge of readers (e.g., college takes 4 years and watching an NBA game takes a few hours), and thus not always written explicitly. Given all these reasons, we ignore the comparison of end-points in this work, although event duration is indeed, another important task.

4 Annotation Scheme Design

To summarize, with the proposed multi-axis modeling (Sec. 2) and interval splitting (Sec. 3), our annotation scheme is two-step. First, we mark every event candidate as being temporally *Anchorable* or not (based on the time axis we are working on). Second, we adopt the dense annotation scheme to label TempRels only between *Anchorable* events. Note that we only work on verb events in this paper, so non-verb event candidates are also deleted in a preprocessing step. We design crowdsourcing tasks for both steps and as we show later, high crowdsourcing quality was achieved on both tasks. In this section, we will discuss some practical issues.

4.1 Quality Control for Crowdsourcing

We take advantage of the quality control feature in CrowdFlower in our crowdsourcing jobs. For any job, a set of examples are annotated by experts beforehand, which is considered gold and will serve two purposes. (i) Qualifying test: Any crowdsourcer who wants to work on this job has to pass with 70% accuracy on 10 questions randomly selected from the gold set. (ii) Surviving test: During the annotation process, questions from the gold set will be randomly given to crowdsourcers without notice, and one has to maintain 70% accuracy on the gold set till the end of the annotation; otherwise, he or she will be forbidden from working on this job anymore and all his/her annotations will be discarded. At least 5 different annotators are required for every judgement and by default, the majority vote will be the final decision.

4.2 Vague Relations

How to handle *vague* relations is another issue in temporal annotation. In non-dense schemes, annotators usually skip the annotation of a *vague* pair. In dense schemes, a majority agreement rule is applied as a postprocessing step to back off a decision to *vague* when annotators cannot pass a majority vote (Cassidy et al., 2014), which reminds us that annotators often label a *vague* relation as non-vague due to lack of thinking.

We decide to proactively reduce the possibility of such situations. As mentioned earlier, our label set for t^1_{start} vs. t^2_{start} is *before*, *after*, *equal* and *vague*. We ask two questions: Q1=Is it possible that t^1_{start} is before t^2_{start}? Q2=Is it possible that t^2_{start} is before t^1_{start}? Let the an-

swers be A1 and A2. Then we have a one-to-one mapping as follows: A1=A2=yes↦*vague*, A1=A2=no↦*equal*, A1=yes, A2=no↦*before*, and A1=no, A2=yes↦*after*. An advantage is that one will be prompted to think about all possibilities, thus reducing the chance of overlook.

Finally, the annotation interface we used is shown in Appendix C.

5 Corpus Statistics and Quality

In this section, we first focus on annotations on the main axis, which is usually the primary story-line and thus has most events. Before launching the crowdsourcing tasks, we checked the IAA between two experts on a subset of TB-Dense (about 100 events and 400 relations). A Cohen's Kappa of .85 was achieved in the first step: anchorability annotation. Only those events that both experts labeled *Anchorable* were kept before they moved onto the second step: relation annotation, for which the Cohen's Kappa was .90 for Q1 and .87 for Q2. Table 3 furthermore shows the distribution, Cohen's Kappa, and F_1 of each label. We can see the Kappa and F_1 of *vague* (κ=.75, F_1=.81) are generally lower than those of the other labels, confirming that temporal *vagueness* is a more difficult semantic phenomenon. Nevertheless, the overall IAA shown in Table 3 is a significant improvement compared to existing datasets.

	b	a	e	v	Overall
Distribution	.49	.23	.02	.26	1
IAA: Cohen's κ	.90	.87	1	.75	.84
IAA: F_1	.92	.93	1	.81	.90

Table 3: IAA of two experts' annotations in a pilot study on the main axis. Notations: **b**efore, **a**fter, **e**qual, and **v**ague.

With the improved IAA confirmed by experts, we sequentially launched the two-step crowdsourcing tasks through CrowdFlower on top of the same 36 documents of TB-Dense. To evaluate how well the crowdsourcers performed on our task, we calculate two quality metrics: accuracy on the gold set and the Worker Agreement with Aggregate (WAWA). WAWA indicates the average number of crowdsourcers' responses agreed with the aggregate answer (we used majority aggregation for each question). For example, if N individual responses were obtained in total, and n of them were correct when compared to the aggregate answer, then WAWA is simply n/N. In the first step, crowdsourcers labeled 28% of the events as *Non-Anchorable* to the main axis, with an accuracy on the gold of .86 and a WAWA of .79.

With *Non-Anchorable* events filtered, the relation annotation step was launched as another crowdsourcing task. The label distribution is b=.50, a=.28, e=.03, and v=.19 (consistent with Table 3). In Table 4, we show the annotation quality of this step using accuracy on the gold set and WAWA. We can see that the crowdsourcers achieved a very good performance on the gold set, indicating that they are consistent with the authors who created the gold set; these crowdsourcers also achieved a high-level agreement under the WAWA metric, indicating that they are consistent among themselves. These two metrics indicate that the annotation task is now well-defined and easy to understand even by non-experts.

No.	Metric	Q1	Q2	All
1	Accuracy on Gold	.89	.88	.88
2	WAWA	.82	.81	.81

Table 4: Quality analysis of the relation annotation step of MATRES. "Q1" and "Q2" refer to the two questions crowdsourcers were asked (see Sec. 4.2 for details). Line 1 measures the level of consistency between crowdsourcers and the authors and line 2 measures the level of consistency among the crowdsourcers themselves.

We continued to annotate INTENTION and OPINION which create orthogonal branches on the main axis. In the first step, crowdsourcers achieved an accuracy on gold of .82 and a WAWA of .89. Since only 16% of the events are in this category and these axes are usually very short (e.g., ***allocate*** *funds to* ***build*** *a museum*.), the annotation task is relatively small and two experts took the second step and achieved an agreement of .86 (F_1).

We name our new dataset *MATRES* for Multi-Axis Temporal RElations for Start-points. Each individual judgement cost us $0.01 and MATRES in total cost about $400 for 36 documents.

5.1 Comparison to TB-Dense

To get another checkpoint of the quality of the new dataset, we compare with the annotations of TB-Dense. TB-Dense has 1.1K verb events, between which 3.4K event-event (EE) relations are annotated. In the new dataset, 72% of the events (0.8K) are anchored onto the main axis, resulting in 1.6K EE relations, and 16% (0.2K) are anchored onto orthogonal axes, resulting in 0.2K EE relations.

The following comparison is based on the 1.8K EE relations in common. Moreover, since TB-Dense annotations are for intervals instead of start-points only, we converted TB-Dense's interval relations to start-point relations (e.g., if ***A*** includes ***B***, then t^A_{start} is before t^B_{start}).

	b	a	e	v	All
b	455	11	5	42	513
a	45	309	16	68	438
e	13	7	2	10	32
v	450	138	20	192	800
All	963	465	43	312	1783

Table 5: An evaluation of MATRES against TB-Dense. Horizontal: MATRES. Vertical: TB-Dense (with interval relations mapped to start-point relations). Please see explanation of these numbers in text.

The confusion matrix is shown in Table 5. A few remarks about how to understand it: **First**, when TB-Dense labels *before* or *after*, MATRES also has a high-probability of having the same label (b=455/513=.89, a=309/438=.71); when MATRES labels *vague*, TB-Dense is also very likely to label *vague* (v=192/312=.62). This indicates the *high agreement level* between the two datasets if the interval- or point-based annotation difference is ruled out. **Second**, many *vague* relations in TB-Dense are labeled as *before*, *after* or *equal* in MATRES. This is expected because TB-Dense annotates relations between *intervals*, while MATRES annotates *start-points*. When durative events are involved, the problem usually becomes more difficult and interval-based annotation is more likely to label *vague* (see earlier discussions in Sec. 3). Example 7 shows three typical cases, where ***e14:became***, ***e17:backed***, ***e18:rose*** and ***e19:extending*** can be considered durative. If only their start-points are considered, the crowdsourcers were correct in labeling ***e14*** before ***e15***, ***e16*** after ***e17***, and ***e18*** equal to ***e19***, although TB-Dense says *vague* for all of them. **Third**, *equal* seems to be the relation that the two dataset mostly disagree on, which is probably due to crowdsourcers' lack of understanding in time granularity and event coreference. Although *equal* relations only constitutes a small portion in all relations, it needs further investigation.

Example 7: Typical cases that TB-Dense annotated *vague* but MATRES annotated *before*, *after*, and *equal*, respectively.
At one point , when it (***e14:became***) clear controllers could not contact the plane, someone (***e15:said***) a prayer. *TB-Dense:* vague*; MATRES:* before
The US is bolstering its military presence in the gulf, as President Clinton (***e16:discussed***) the Iraq crisis with the one ally who has (***e17:backed***) his threat of force, British prime minister Tony Blair. *TB-Dense:* vague*; MATRES:* after
Average hourly earnings of nonsupervisory employees (***e18:rose***) to $12.51. The gain left wages 3.8 percent higher than a year earlier, (***e19:extending***) a trend that has given back to workers some of the earning power they lost to inflation in the last decade. *TB-Dense:* vague*; MATRES:* equal

6 Baseline System

We develop a baseline system for TempRel extraction on MATRES, assuming that all the events and axes are given. The following commonly-used features for each event pair are used: (i) The part-of-speech (POS) tags of each individual event and of its neighboring three words. (ii) The sentence and token distance between the two events. (iii) The appearance of any modal verb between the two event mentions in text (i.e., will, would, can, could, may and might). (iv) The appearance of any temporal connectives between the two event mentions (e.g., before, after and since). (v) Whether the two verbs have a common synonym from their synsets in WordNet (Fellbaum, 1998). (vi) Whether the input event mentions have a common derivational form derived from WordNet. (vii) The head words of the preposition phrases that cover each event, respectively. And (viii) event properties such as Aspect, Modality, and Polarity that come with the TimeBank dataset and are commonly used as features.

The proposed baseline system uses the averaged perceptron algorithm to classify the relation between each event pair into one of the four relation types. We adopted the same train/dev/test split of TB-Dense, where there are 22 documents in train, 5 in dev, and 9 in test. Parameters were tuned on the train-set to maximize its F_1 on the dev-set, after which the classifier was retrained on the union of train and dev. A detailed analysis of the baseline system is provided in Table 6. The performance on *equal* and *vague* is lower than on *before* and *after*, probably due to shortage in these labels in the training data and the inherent difficulty in event coreference and temporal vagueness. We can see, though, that the overall performance on MATRES is much better than those in the literature for TempRel extraction, which used to be in the low 50's (Chambers et al., 2014; Ning et al., 2017). The same system was also retrained

and tested on the original annotations of TB-Dense (Line "Original"), which confirms the significant improvement if the proposed annotation scheme is used. Note that we *do not* mean to say that the proposed baseline system itself is better than other existing algorithms, but rather that the proposed annotation scheme and the resulting dataset lead to better defined machine learning tasks. In the future, more data can be collected and used with advanced techniques such as ILP (Do et al., 2012), structured learning (Ning et al., 2017) or multi-sieve (Chambers et al., 2014).

	Training			Testing		
	P	R	F_1	P	R	F_1
Before	.74	.91	.82	.71	.80	.75
After	.73	.77	.75	.55	.64	.59
Equal	1	.05	.09	-	-	-
Vague	.75	.28	.41	.29	.13	.18
Overall	**.73**	**.81**	**.77**	**.66**	**.72**	**.69**
Original	.44	.67	.53	.40	.60	.48

Table 6: Performance of the proposed baseline system on MATRES. Line "Original" is the same system retrained on the original TB-Dense and tested on the same subset of event pairs. Due to the limited number of *equal* examples, the system did not make any *equal* predictions on the testset.

7 Conclusion

This paper proposes a new scheme for TempRel annotation between events, simplifying the task by focusing on a single time axis at a time. We have also identified that end-points of events is a major source of confusion during annotation due to reasons beyond the scope of TempRel annotation, and proposed to focus on start-points only and handle the end-points issue in further investigation (e.g., in event duration annotation tasks). Pilot study by expert annotators shows significant IAA improvements compared to literature values, indicating a better task definition under the proposed scheme. This further enables the usage of crowdsourcing to collect a new dataset, MATRES, at a lower time cost. Analysis shows that MATRES, albeit crowdsourced, has achieved a reasonably good agreement level, as confirmed by its performance on the gold set (agreement with the authors), the WAWA metric (agreement with the crowdsourcers themselves), and consistency with TB-Dense (agreement with an existing dataset). Given the fact that existing schemes suffer from low IAAs and lack of data, we hope that the findings in this work would provide a good start towards understanding more sophisticated semantic phenomena in this area.

Acknowledgements

We thank Martha Palmer, Tim O'Gorman, Mark Sammons and all the anonymous reviewers for providing insightful comments and critique in earlier stages of this work. This research is supported in part by a grant from the Allen Institute for Artificial Intelligence (allenai.org); the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network; by DARPA under agreement number FA8750-13-2-0008; and by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053 (the ARL Network Science CTA).

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, of the Army Research Laboratory or the U.S. Government. Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the view of the ARL.

References

James F Allen. 1984. Towards a general theory of action and time. *Artificial intelligence* 23(2):123–154.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 806–814.

Steven Bethard, Oleksandr Kolomiyets, and Marie-Francine Moens. 2012. Annotating story timelines as temporal dependency structures. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)*. ELRA, pages 2721–2726.

Steven Bethard, James H Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *IEEE International Conference on Semantic Computing (ICSC)*. pages 11–18.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1052–1062.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, pages 565–572.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. pages 189–198.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 501–506.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics* 2:273–284.

Marta Coll-Florit and Silvia P Gennari. 2011. Time in language: Event duration in language comprehension. *Cognitive psychology* 62(1):41–79.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1643–1648.

Tuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 Task 5: QA TEMPEVAL - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pages 792–800.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, Ruben Urizar, and Fondazione Bruno Kessler. 2015. SemEval-2015 Task 4: TimeLine: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pages 778–786.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event nugget annotation: Processes and issues. In *Proceedings of the Workshop on Events at NAACL-HLT*.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation*. pages 51–61.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1038–1048.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018a. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 841–851.

Qiang Ning, Zhongzhi Yu, Chuchu Fan, and Dan Roth. 2018b. Exploiting partially annotated data for temporal relation extraction. In *The Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 148–153.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*. Association for Computational Linguistics, Austin, Texas, pages 47–56.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering* 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The TIMEBANK corpus. In *Corpus linguistics*. volume 2003, pages 647–656.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M Lai. 2012. Learning to temporally order medical events in clinical text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 70–74.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2195–2204.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation* 43(3):227.

Steven Schockaert and Martine De Cock. 2008. Temporal reasoning about fuzzy intervals. *Artificial Intelligence* 172(8-9):1158–1193.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Association for Computational Linguistics, Denver, Colorado, pages 89–98.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2:143.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, 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*. volume 2, pages 1–9.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval temporal relation identification. In *SemEval*. pages 75–80.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *SemEval*. pages 57–62.

A Examples of Table 1

The names of those categories in Table 1 are straightforward. Here we further provide examples for each of them in Example 8. Note that most of them are consistent with the definitions in the literature, with one exception for INTENTION. In TimeML (Pustejovsky et al., 2003a), there are two types of intentions, I-Action (e.g., *attempt, try* and *promise*) and I-State (e.g., *believe, intend* and *want*). But our definition of intention is the actual intent of these verbs. For example, in Example 8, ***e20*** and ***e21*** are INTENTION. This definition is more general so that verbs that are not I-Action or I-State can still create orthogonal axis of intention, e.g., the verb "allocated" in the sentence of ***e21***.

Example 8
[Orthogonal axis] INTENTION/OPINION I plan/want to (***e20:leave***) tomorrow. The mayor has allocated funds to (***e21:build***) a museum. I think he will (***e22:win***) the race.
[Parallel axis] HYPOTHESIS/GENERIC If I'm (***e23:elected***), I will cut income tax. If I'm elected, I will (***e24:cut***) income tax. Fruit (***e25:contains***) water. Lions (***e26:hunt***) zebras.
[Not on any axis] NEGATION The financial assistance from the Wolrd Bank is not (***e27:helping***). They don't (***e28:want***) to play with us. He failed to (***e29:find***) buyers.
[Other] STATIC/RECURRENT He (***e30:is***) brave. New York (***e31:is***) on the east coast. The shuttle will be (***e32:departing***) at 6:30am every day.

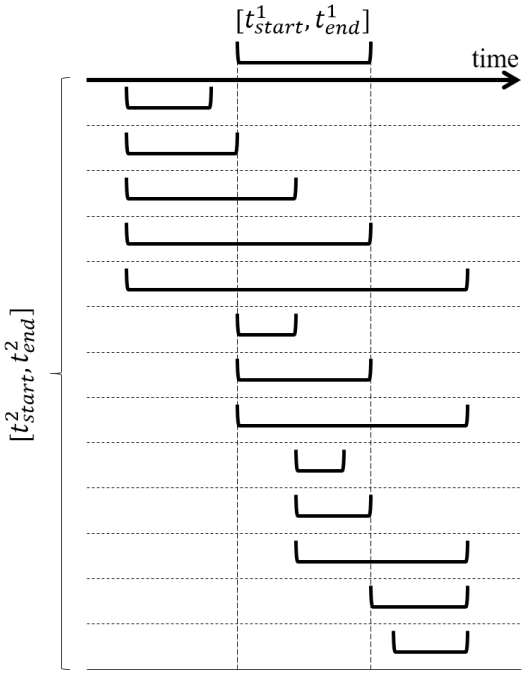

Figure 4: Thirteen possible relations between two events whose timescopes are $[t^1_{start}, t^1_{end}]$ and $[t^2_{start}, t^2_{end}]$ (from top to bottom): *after, immediately after, after and overlap, ends, included, started by, equal, starts, includes, ended by, before and overlap, immediately before* and *before*.

B *Anchorable* vs. *Actual*

As discussed in the paper, when we check if an event is *Anchorable* onto the main axis, it seems very similar to annotating whether an event is *Actual* in REALIS labeling. We have discussed the differences

in Sec. 2.3.3. To better understand them, we randomly selected 5 documents from RED (O'Gorman et al., 2016), where there are 314 events, 166 of which are verbs (we only handle verb events). Two experts annotated the anchorability of these 166 verb events independently without looking at the original REALIS annotation from RED, and they achieved a Cohen's Kappa of .88 in anchorability annotation, consistent with their Cohen's Kappa achieved on MATRES. To aggregate the result from two experts, we mark an event as *Anchorable* only when both experts labeled *Anchorable*. As for REALIS labeling in RED, we group GENERIC, HYPOTHETICAL, and HEDGED into a single label of *Non-Actual*.

		Anchorable	
		Yes	No
Actual	Yes	108	25
	No	0	33

Table 7: Comparison between anchrability and factuality on a subset of verb events randomly selected from RED.

The comparison between *Anchorable* and *Actual* is shown in Table 7. On this subset of 166 events, we did not see *Anchorable* events that are *Non-Actual* because such cases are indeed less frequent in practice; the only difference is that we annotated 25 events as *Non-Anchorable*, while RED annotated them as *Actual*. Among the 25 different cases, 11 are INTENTION, 4 are OPINION, 6 are STATIC, and 4 are NEGATION. Typical examples from each category are shown in Example 9. Note that if we calculate the McNemar's statistics based on Table 7, *Anchorable* and *Actual* are statistically different with $p \ll 0.001$.

Example 9: Typical cases that RED annotated *Actual* and we annotated *Non-Anchorable*.
Libya has since agreed to (***e33:pay***) compensation to the families of the Berlin disco victims as well as the families of the victims of the 1988 Pan Am 103 bombing over Lockerbie, Scotland, which killed 270 people, including 189 Americans. [**We think it is INTENTION**]
Gadhafi had long been ostracized by the West for (***e34:sponsoring***) terrorism, but in recent years sought to emerge from his pariah status by abandoning weapons of mass destruction and renouncing terrorism in 2003. [**We think it is OPINION**]
We need to resolve the deep-seated causes that have resulted in these problems, Premier Wen said in an interview with Hong Kong-(***e35:based***) Phoenix Television. [**We think it is STATIC**]
Fuel prices had been frozen for six years, but the government said it could no longer afford to (***e36:subsidize***) them. [**We think it is NEGATION**]

C Annotation Interface

The annotation interface was designed based on the web interface of CrowdFlower. In the anchorability annotation step (i.e., the first step), we show each crowdsourcer one event at a time, along with the full context of this event. Crowdsourcers only need to make a binary decision of Yes/No, as shown in Fig. 5.

The interface design for the relation annotation step (i.e., the second step) is tricky. As explained in Sec. 4.2, we need to ask two questions for each pair of events to figure out the actual TempRel: Q1=Is it possible that t^1_{start} is before t^2_{start}? Q2=Is it possible that t^2_{start} is before t^1_{start}? We notice in practice that asking Q1 and Q2 simultaneously (as shown in Fig. 6) gives annotators the wrong impression that there has to be one "yes" and one "no". Therefore, we decide to ask Q1 and Q2 separately. Specifically, we launch two separate tasks. One task only has Q1 (Task A), and the other only has Q2 (Task B), so that a same annotator is guaranteed not to see Q1 and Q2 simultaneously (as shown in Fig. 7).

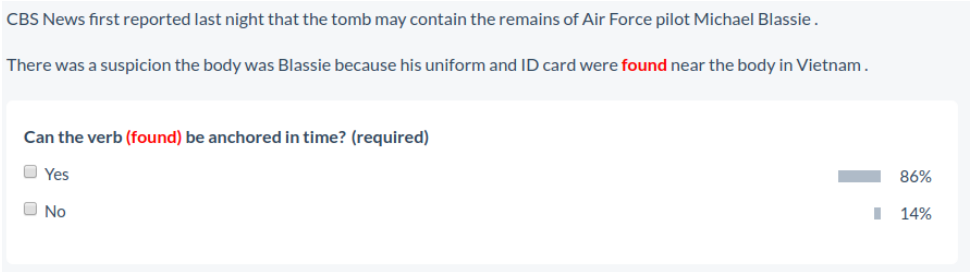

Figure 5: Annotation interface for the first step: temporal anchorability. The owner of the task can see the crowdsourcers' distribution of each answer (e.g., 86% and 14%), which is of course not available to crowdsourcers.

Figure 6: Tentative annotation interface for the second step: relation annotation. This design gives crowdsourcers the wrong impression to select one "yes" and one "no" for Q1 and Q2, leading to strong correlation between answers of Q1 and answers of Q2.

(a) Task A: Only ask Q1

(b) Task B: Only ask Q2

Figure 7: The final annotation interface, where Q1 and Q2 are posed in separate tasks so that a single annotator will not see both two questions simultaneously, forcing them to think the temporal relation carefully instead of simply putting the opposite answer to the other question.