Exploring Temporal Vagueness with Mechanical Turk

Yuping Zhou
Computer Science Department
Brandeis University
Waltham, MA 02452
yzhou@brandeis.edu

Nianwen Xue
Computer Science Department
Brandeis University
Waltham, MA 02452
xuen@brandeis.edu

Abstract

This paper proposes schematic changes to the TempEval framework that target the temporal vagueness problem. Specifically, two elements of vagueness are singled out for special treatment: vague time expressions, and explicit/implicit temporal modification of events. As proof of concept, an annotation experiment on explicit/implicit modification is conducted on Amazon's Mechanical Turk. Results show that the quality of a considerable segment of the annotation is comparable to annotation obtained in the traditional double-blind setting, only with higher coverage. This approach offers additional flexibility in how the temporal annotation data can be used.

1 Introduction

Event-based temporal inference aims at determining temporal anchoring and relative ordering of events in text. It is a fundamental natural language technology that supports a wide range of natural language applications, such as Information Extraction (Ji, 2010), Question Answering (Harabagiu and Bejan, 2005; Harabagiu and Bejan, 2006) and Text Summarization (Lin and Hovy, 2001; Barzilay et al., 2002). Crucial to developing this technology is consistently annotated, domain-independent data sufficient to train automatic systems, but this has proven to be challenging.

The difficulty has mainly been attributed to rampant temporal vagueness in natural language, affecting all high-level annotation tasks (Verhagen et al., 2009). Focusing on one of the tasks, Zhou and Xue (2011) show that by pairing up discourse-related events and by making the classification scheme paying more attention to vagueness in natural language, inter-annotator agreement increases from 65% to the low 80%. Despite the significant improvement, problems identified by Zhou and Xue (2011) towards the end of their paper suggest that how temporal modification is handled in the TempEval annotation scheme needs to be revised to further keep vagueness in line. This paper is an attempt in that direction.

The rest of the paper is organized as follows: In Section 2, we first offer arguments for changing the way temporal modification is handled in temporal annotation, then lay out an outline for the change and motivate the experiment being carried out on Amazon's Mechanical Turk. We then describe the design of the experiment in detail in Section 3, and present experiment results in Section 4. And finally in Section 5, we conclude the paper.

2 Motivation

2.1 Treatment of temporal modification in the TempEval framework

In the TempEval framework (Verhagen et al., 2009; Verhagen et al., 2010), the part of temporal modification to be annotated is time expressions, i.e. those bearing the <TIMEX3> tag following the definition in the TimeML (Pustejovsky et al., 2003). Simply put, they are elements that express time, date, duration etc., for example, *7 o'clock*, *June 19, 2008*, and *ten years*. In this framework, time expressions in text are identified and subjected to the following

Proceedings of the 6th Linguistic Annotation Workshop, pages 124–128,
Jeju, Republic of Korea, 12-13 July 2012.

kinds of annotation:

- their type is classified: {*time, date, duration, set*};
- their value is specified in a normalized form (e.g. "2008/6/19" for *June 19, 2008*);
- their temporal relation to some selected events is classified: {*before, overlap, after, before-or-overlap, overlap-or-after, vague*}.

2.2 Problems concerning "temporal vagueness"

2.2.1 Do all time expressions fit into the same mold?

In the current scheme, all time expressions have a VALUE attribute and the TimeML specifies how to standardize it (Pustejovsky et al., 2003). However, a subgroup of time expressions are noticeably ignored by the specifications: those whose value is hard to pinpoint, for example, *now*, *soon*, *several years* etc. These vague expressions constitute a large part of the vagueness problem in temporal annotation. Although their values are hard to pinpoint, they are an important part of temporal specification in natural language, and can provide information useful in temporal inference if they are adequately characterized in a way communicable with those having a definite value.

2.2.2 Should time expressions participate in temporal relation with events?

How useful a temporal relation classification is between an event and a time expression in certain types of temporal modifier is highly questionable. Let us take *from June 6 to August 14* as an example. According to the TimeML, there are two time expressions in this phrase: *June 6* and *August 14*, but suppose it is used to specify the temporal location of an event *e1* in a sentence, to specify that *e1* OVERLAPS *June 6* and that *e1* OVERLAPS *August 14* does not capture the exact relation between *from June 6 to August 14* and *e1*.[1] In other words, temporal vagueness is artificially introduced into annotation by the scheme when the text itself is perfectly clear in this respect. Other types of temporal modifiers that share this problem include *since [1990]*, *[three years] ago*, *until [now]* etc. (square brackets delimit time expressions).

[1] It is possible to capture this temporal relation with the full-blown TimeML apparatus, however, there is a reason why the TempEval framework is a simplified version of the TimeML (Verhagen et al., 2009).

2.2.3 How to choose time∼event pairs for annotation?

How to find annotation targets for different types of temporal relation has been a long-standing problem in temporal annotation, and the normal solution is to annotate all pairs that satisfy some technical constraints specified in syntactic, semantic and/or discourse terms (Verhagen et al., 2009; Xue and Zhou, 2010; Zhou and Xue, 2011). In the case of temporal relation between time and event, Xue and Zhou (2010) proposed to let annotators judge which event(s) a given time expression is intended to modify. There are at least three problems with this proposal as it stood.

First, as alluded to in Section 2.2.2, time expressions usually do not modify predicates by themselves, unless they can stand alone as a temporal modifier (e.g. *now*, *tomorrow*, *this week*). To use the temporal modifier *from June 6 to August 14* as an example again, neither *June 6* nor *August 14*, but the whole prepositional phrase, has an intended modification target.

Second, the modification relation is construed in terms of syntactic scope, hence the range of choice is restricted to the same sentence. This is of course understandable: Given the double-blind setup and inherently greater uncertainty associated with modification relation across sentence boundaries, it makes sense to minimize uncertainty for higher agreement. On the other hand though, this restriction can potentially result in significant information loss since a temporal expression can have (semantic/discourse) scope over several sentences or even paragraphs. So who should decide precision or recall should take precedence? And at what point?

The third problem is the directionality of it: to find events given a time or the other way around? This may seem a trivial point–and it is with the "same sentence" restriction in place–but operationally it makes quite a difference if the restriction is abandoned. Suppose we are to find all time∼event pairs in an article containing 10 temporal modifiers and 60 events. In a simplified version, to find events

given a temporal modifier amounts to 10 searches to find an uncertain number of hits out of 60 candidates, whereas to find the temporal modifier for a given event amounts to 60 searches to find 1 hit out of 10 candidates. Clearly the latter way presents an easier individual task than the former, but presents it more times, so the overall quality of the results is probably better. Furthermore, if we consider the problem in a more realistic scenario where temporal modification only happens to events in the same sentence and below, to find the temporal modifier of a given event can be done in the (relatively) normal flow of one careful reading because the candidates for selection are already in the familiar territory. To find events being modified by a given temporal modifier means doing the search and paying attention to new material at the same time, which can be highly distracting.

2.3 Outline of a solution

Two levels should be distinguished in annotation with respect to temporal modification: The first level is time expressions (as defined in the TimeML) and the second is temporal modifiers, the predicate-modifying units, usually (but not always) time expressions along with prepositions/postpositions associated with them.

These two levels are obviously related, but play different roles in temporal annotation. Time expressions should be divided into two subgroups: *definite* and *indefinite*, each associated with a different value-characterizing scheme. Annotation of time expressions serves as a building block to interpretation of temporal modifiers, and temporal modifiers are linked directly to events that they modify, explicitly or implicitly. In other words, it is temporal modifiers, not time expressions, that have a relation with events; furthermore, it is a modification relation that should be identified according to speakers' interpretation of the text.

Two parts of this solution are challenging, if not impossible, for the traditional double-blind annotation: characterization of indefinite time expressions, and linking events with modifying temporal expressions without distance restrictions. Both would involve a healthy amount of variability and would rely on a distribution for usable data. This leads us to Amazon's Mechanical Turk (MT). In this paper, we only describe the experiment that deals with linking temporal modifiers with events.

3 HIT design

We make use of data from two sources. The first source is Chinese annotation data prepared for the TempEval-2 campaign (Verhagen et al., 2010), from which we use the time expressions and verbal events. The second source is the Chinese TreeBank (CTB) (Xue et al., 2005), in which temporal-related nodes (close to our notion of "temporal modifier") are suffixed with the "-TMP" function tag, so we use it to expand time expressions (taken directly from TempEval-2 data) into temporal modifiers as follows: Without passing an S node, find the nearest ancestor of the time expression that bears the "-TMP" suffix and then use all the terminal nodes within as the corresponding temporal modifier.

Verbal events (taken directly from TempEval-2 data) are split into groups so that each HIT deals with fewer than 20 events. A non-event is chosen randomly as a decoy to help weed out irresponsible Tukers. In each HIT, the article is presented in the one-sentence-per-line format, with temporal expressions underlined and events in boldface (see Figure 1 for a screenshot). Next to each event is a drop-down list, presenting three types of choice:

1. <temporal modifiers in quotes>
2. *not in the list*
3. *not the main element of a predicate*

The *not the main element of a predicate* option is for the decoys and the *not in the list* option is for atemporal events, events that do not have a temporal modifier, or events that have a temporal modifier outside the given list. Temporal expressions appearing in the text up to the event anchor are presented in quotation marks in the reverse order of their occurrence, with the newer instance of the same lexical item replacing the old one as it emerges. In Figure 1, each type of choice has a representative.

4 Results

The distribution of all annotations and those representing a time∼event link with respect to the majority MT-internal agreement is shown in Table 1.

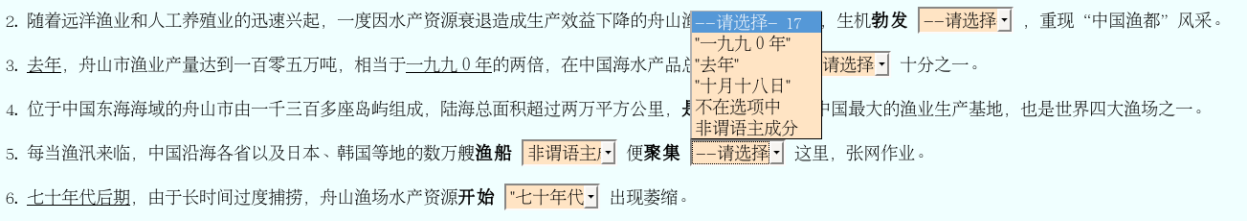

Figure 1: Part of a HIT from the experiment

Range	***No. tkn (percent)***	***Links***	
		Total (percent)	***No. intraS***
0.2-0.5	153(6.3)	83(3.4)	17
0.5-0.6	449(18.6)	244(10.1)	57
0.6-0.7	245(10.1)	143(5.9)	59
0.7-0.8	138(5.7)	84(3.5)	57
0.8-0.9	353(14.6)	235(9.7)	158
0.9-1.0	1082(44.7)	922(38.1)	864
Total:	2420(100)	1711(70.7)	1212

Table 1: Distribution of all annotations and time∼event links. *No. intraS*: number of intra-sentential links.

Range	***Agreement*** (%)	***Concentration intraS (%)***
0.2≤ A <0.5	48.2	20.5
0.5≤ A <0.6	59.5	23.4
0.6≤ A <0.7	71.7	41.3
0.7≤ A <0.8	74.9	67.9
0.8≤ A <0.9	83.2	67.2
0.9≤ A ≤1.0	91.5	93.7
Total:	78.0	70.8

Table 2: Agreement with expert annotation

65% of all tokens fall within the 0.7-1 MT-internal agreement range, 70.7% of all majority annotations produce a link between a temporal modifier and an event, and 72.5% of links created have an MT-internal agreement of 0.7 or higher. Intra-sentential links are very concentrated in the top MT-internal agreement range, and their concentration for the most part correlates with both the MT-internal agreement and agreement with expert annotation, as shown in Table 2 below. Also, the decline of agreement with expert annotation by and large keeps pace with the MT-internal agreement. These trends are consistent with what one expects from annotation of this sort and the assumption that the uncertainty level increases as annotation goes across sentence boundaries.

Within the high-agreement range, the quality of the MT annotation is comparable to that produced in a double-blind setting with trained annotators (Xue and Zhou, 2010), as shown in Table 3. With comparable levels of agreement, the MT annotation has a coverage 11-15 percentage points greater than the previously reported double-blind annotation of the same data, presumably because the “same sentence” restriction is lifted. It should be noted that the maximum value of coverage is not 100% (i.e. not all events have a temporal modifier), and with the problem of vagueness, is probably unknowable.

MT annotation			***Double-blind***	
Range	*Agr*	*Coverage*	*Agr*	*Coverage*
≥0.8	88.6	47.8%	86	36.4%*
≥0.7	86.1	51.3%		

Table 3: Comparison with double-blind annotation of the same data. *Coverage*: no. of events in a link/total no. of events; *: this number is directly based on the TempEval-2 Chinese data.

With this distribution of data, the MT annotation offers greater flexibility in using the annotation: Depending on demands on different levels of data reliability, one can take a section of the data by choosing different cutoffs. So this choice is left to the user of the annotation data, not the creator.

5 Conclusions

Three takeaways: i) To tackle the vagueness problem, elements of vagueness need to be identified and treated with care; ii) vagueness can be characterized with a distribution of different annotations and MT

makes it feasible; iii) this approach, when implemented successfully, not only provides high-quality data, but also offers additional flexibility in data use with respect to information quantity vs. certainty.

Acknowledgments

This work is supported by the National Science Foundation via Grant No. 0855184 entitled "Building a community resource for temporal inference in Chinese". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

References

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Sanda Harabagiu and Cosmin Adrian Bejan. 2005. Question Answering Based on Temporal Inference. In *Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering*, Pittsburgh, Pennsylvania.

Sanda Harabagiu and Cosmin Adrian Bejan. 2006. An Answer Bank for Temporal Inference. In *Proceedings of LREC 2006*, Genoa, Italy.

Heng Ji. 2010. Challenges from information extraction to information fusion. In *Proceedings of COLING 2010*, pages 507–515, Beijing, China, August.

Chin-Yew Lin and Eduard Hovy. 2001. Neats: A multidocument summarizer. In *Proceedings of the Document Understanding Workshop*.

James Pustejovsky, Jose Castano, Roser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, Tilburg, July.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval Challenge: Identifying Temporal Relation in Text. *Language Resources and Evaluation*, 43(1):161–179.

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, Uppsala, Sweden, July. Association for Computational Linguistics.

Nianwen Xue and Yuping Zhou. 2010. Applying Syntactic, Semantic and Discourse Constraints to Chinese Temporal Annotation. In *Proceedings of COLING 2010*, pages 1363–1372, Beijing, China, August.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.

Yuping Zhou and Nianwen Xue. 2011. Discourse-constrained temporal annotation. In *Linguistic Annotation Workshop 2011*, pages 161–169.