Exploring Temporal Vagueness with Mechanical Turk

Yuping Zhou (yzhou@brandeis.edu), Nianwen Xue (xuen@brandeis.edu)
Computer Science Department, Brandeis University, Waltham, MA 02452, USA

1. Why Temporal Vagueness (TempV)?

- TempV is a major obstacle to consistent temporal annotation;
- Temporal annotation's crucial for training of temporal inference-capable systems;
- Temporal inference supports applications like Information Extraction [Ji2010], Question Answering [Harabagiu and Bejan2005, Harabagiu and Bejan2006] and Text Summarization [Lin and Hovy2001, Barzilay et al.2002].

2. What kind of TempV?

- Vague time expressions: *now, soon, a long time* etc.;
- Implicit modification of multiple events by one temporal modifier, esp. across sentence/paragraph boundaries.

3. Why Mechanical Turk (MTurk)?

- TempV $\rightsquigarrow$ annotation uncertainty $\rightsquigarrow$ low agreement ;
  e.g. *now* refers to this second, this minute, this hour...?
- Solution: characterizing TempV with a distribution of different annotations;
- "distribution" $\rightsquigarrow$ way more than 2 annotators $\rightsquigarrow$ MTurk.

4. The experiment

- Task: linking temporal modifiers (TMod) with modified events;
- 10 annotators per item;
- Data source: Chinese data from the TempEval-2 campaign [Verhagen et al.2010], and the Chinese TreeBank [Xue et al.2005].

5. HIT design

- <20 events (plus 1 non-event) per HIT;
- One sentence per line;
- Event: in boldface
  TMod: underlined;
- Drop-down list next to event:
  - <temporal modifiers in quotes>
  - *not in the list*
  - *not the main element of a predicate*

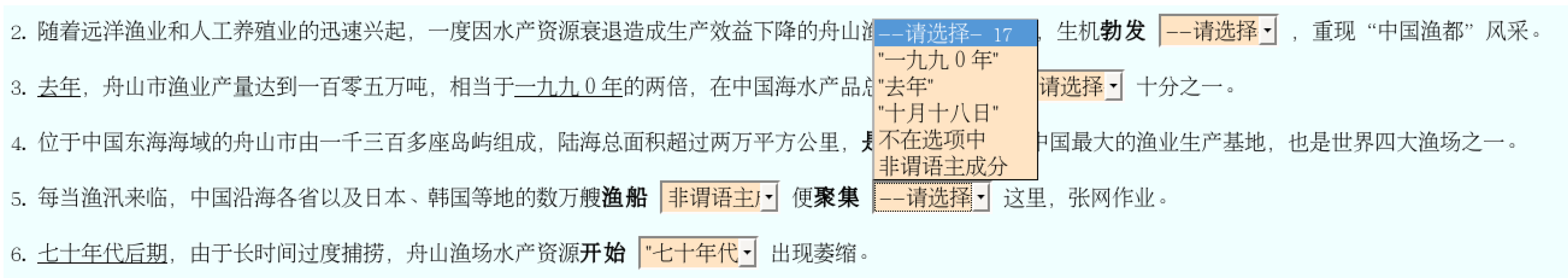

Figure 1: Part of a HIT from the experiment

6. Results

Overall distribution

- 65% of all tokens fall whithin the 0.7-1 Mturk-internal agreement;
- 70.7% of all majority annotations produce a TMod∼event link;
- 72.5% of links created have an MTurk-internal agreement of 0.7 or higher.
- Intra-sentential links: very concentrated in the top MTurk-internal agreement range;

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.

Agreement with expert annotation

- MTurk-internal agreement keeps pace with agreement with expert;
- Both correlate with the concentration 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

Comparison with double-blind annotation

- Within the high-agreement range (≥0.7), the quality of MT annotation is comparable to that produced in a double-blind setting [Xue and Zhou2010];
- At comparable levels of agreement, MT annotation achieves higher coverage (11-15 percentage points).
  Coverage: num. of events in a link/total num. of events;
  Note: The maximum value of coverage is not 100%. (Quiz: why?)

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.
*: this number is directly based on the TempEval-2 Chinese data.

7. Conclusions

- To tackle the vagueness problem, elements of vagueness need to be identified and treated with care;
- Vagueness can be characterized with a distribution of different annotations and MT makes it feasible;
- 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.

References

[Barzilay et al.2002] 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.

[Harabagiu and Bejan2005] 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.

[Harabagiu and Bejan2006] Sanda Harabagiu and Cosmin Adrian Bejan. 2006. An Answer Bank for Temporal Inference. In *Proceedings of LREC 2006*, Genoa, Italy.

[Ji2010] Heng Ji. 2010. Challenges from information extraction to information fusion. In *Proceedings of COLING 2010*, pages 507–515, Beijing, China, August.

[Lin and Hovy2001] Chin-Yew Lin and Eduard Hovy. 2001. Neats: A multidocument summarizer. In *Proceedings of the Document Understanding Workshop*.

[Verhagen et al.2010] 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.

[Xue and Zhou2010] 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.

[Xue et al.2005] 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.