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;

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 ~ annotation uncertainty ~ low agreement:
     e.g. now refers to this second, this minute, this hour...
   - Solution: characterizing TempV with a distribution of different annotations;
   - “distribution” ~ way more than 2 annotators ~ 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
   - Overall distribution
     - 65% of all tokens fall within 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;

6. Results

   Agreement with expert annotation
   - MTurk-internal agreement keeps pace with agreement with expert;
   - Both correlate with the concentration of intra-sentential links;

<table>
<thead>
<tr>
<th>Range</th>
<th>Agreement (%)</th>
<th>Concentration (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤0.2</td>
<td>48.2</td>
<td>20.5</td>
</tr>
<tr>
<td>0.2&lt;</td>
<td>50.5</td>
<td>23.4</td>
</tr>
<tr>
<td>0.5&lt;</td>
<td>71.7</td>
<td>41.3</td>
</tr>
<tr>
<td>0.7&lt;</td>
<td>74.9</td>
<td>67.9</td>
</tr>
<tr>
<td>0.8&lt;</td>
<td>83.2</td>
<td>67.2</td>
</tr>
<tr>
<td>0.9&lt;</td>
<td>91.5</td>
<td>93.7</td>
</tr>
<tr>
<td>Total</td>
<td>78.0</td>
<td>70.8</td>
</tr>
</tbody>
</table>

   Coverage: num. of events in a link/total num. of events;
   Note: The maximum value of coverage is not 100%. (Quiz: why?)

   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).

   MT annotation | Double-blind
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>Coverage</td>
</tr>
<tr>
<td>≥0.8</td>
<td>88.6</td>
</tr>
<tr>
<td>≥0.7</td>
<td>85.1</td>
</tr>
</tbody>
</table>

   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

Figure 1: Part of a HIT from the experiment