Language experiment results

Study on dialogue segmentation with large numbers of anonymous volunteer annotators from the Internet

Daniel Midgley, University of Western Australia

June 2008
Hello.

- Thanks for taking part in and/or asking about the language experiment.
- This presentation tells a bit about it, and how it’s going.
- The data presented here are from a study that was conducted in May 2008.
What is this experiment?

- If you’ve done the experiment, you’ll remember some of this:
- Volunteers read a brief description of how to segment dialogues
  - and can choose to annotate up to eleven medium-size dialogues
- In particular, the instructions tell them to look for
  - nuclei: utterances where one speaker says something and another responds
  - satellites: shorter utterances that give feedback on the nucleus

Excerpt from the annotation guide

The experiment can be found at http://tinyurl.com/ynwmx9
Experimental design

- Annotation consisted of clicking spaces in between utterances, where respondents thought a segment boundary should go.
- In the pilot phase of this experiment,
  - 389 responses were received
  - from about 100 different volunteers

Excerpt from dialogue corpus:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>CND</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I think a break should go here.</td>
<td>I think a break should go here.</td>
<td>CND</td>
<td>good afternoon</td>
</tr>
<tr>
<td></td>
<td>I think a break should go here.</td>
<td>ARE</td>
<td>hey what is going on</td>
</tr>
<tr>
<td></td>
<td>I think a break should go here.</td>
<td>CND</td>
<td>not much</td>
</tr>
<tr>
<td></td>
<td>I think a break should go here.</td>
<td>CND</td>
<td>going to plan our next business meeting</td>
</tr>
<tr>
<td></td>
<td>I think a break should go here.</td>
<td>ARE</td>
<td>I guess so</td>
</tr>
<tr>
<td></td>
<td>I think a break should go here.</td>
<td>CND</td>
<td>I always look forward to them</td>
</tr>
<tr>
<td></td>
<td>I think a break should go here.</td>
<td>ARE</td>
<td>break</td>
</tr>
</tbody>
</table>
Individual responses were mostly quite good

Many respondents did well according to an algorithm called ‘WindowDiff’, which is how we measured performance

- Lower numbers = agreed with our judgments more
- Most people scored around .20
  - mean = 0.229
- Notice the long tail of seemingly random responses

But something happens when all responses are overlaid on each other...
Each click a vote

- When responses are aggregated, broad patterns of agreement show up:
  - High peaks, where many annotators agreed on a segment boundary
  - Some confusion where segmentation may be ambiguous
  - Some background noise from near-random responders
How do we determine the group’s answers?

- We noticed ‘peaks’ in the data.
- A ‘peak’ is any utterance $u$ where
  - $\text{votes}_{u-1} < \text{votes}_u > \text{votes}_{u+1}$
  - $\text{votes}_u$ is higher than background noise
    - somewhat arbitrarily defined as anything over the mean
- Utterance ‘a’ is a peak
  - Both its neighbours are lower numbers
- Utterance ‘b’ is not a peak
  - because adjacent utterance ‘a’ is higher
- Utterance ‘c’, though a local maximum, is not a peak
  - No higher than background noise (the mean)
When we grouped the various responses and treated the data like that of a single annotator, we found extremely close agreement between our judgments and those of the group.

The next slides show the experimental data:

- Numbers inside a dot show how many respondents voted for that spot to be a boundary.
- A yellow dot marks our judgments as to where a boundary goes.
Sample dialogue result
Sample dialogue result

mean = 8.29

n = 39
Sample dialogue result

- **WindowDiff = 0.093**
Sample dialogue result
Sample dialogue result

mean = 7.89
n = 36
Sample dialogue result

- WindowDiff = 0.044

(mean = 7.89  n = 36)
Overall results

- Responses of the group are always better than the average
  - and in almost half the cases, equal to or better than the best annotator
- These are extremely encouraging results! The agreement we’re seeing here is unusually high for a dialogue task.

<table>
<thead>
<tr>
<th>Dialogue</th>
<th>WD average as marked by volunteers</th>
<th>WD single annotator best</th>
<th>WD single annotator worst</th>
<th>WD for group opinion</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>e041a</td>
<td>0.210</td>
<td>0.094</td>
<td>0.766</td>
<td>0.094</td>
<td>Equal to best annotator</td>
</tr>
<tr>
<td>e041b</td>
<td>0.276</td>
<td>0.127</td>
<td>0.794</td>
<td>0.095</td>
<td>Better than best annotator</td>
</tr>
<tr>
<td>e059</td>
<td>0.236</td>
<td>0.080</td>
<td>0.920</td>
<td>0.107</td>
<td>Better than the average</td>
</tr>
<tr>
<td>e081a</td>
<td>0.244</td>
<td>0.037</td>
<td>0.611</td>
<td>0.148</td>
<td>Better than the average</td>
</tr>
<tr>
<td>e081b</td>
<td>0.267</td>
<td>0.093</td>
<td>0.537</td>
<td>0.148</td>
<td>Better than the average</td>
</tr>
<tr>
<td>e096a</td>
<td>0.219</td>
<td>0.083</td>
<td>0.604</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e096b</td>
<td>0.161</td>
<td>0.000</td>
<td>0.689</td>
<td>0.044</td>
<td>Better than the average</td>
</tr>
<tr>
<td>e115</td>
<td>0.214</td>
<td>0.079</td>
<td>0.750</td>
<td>0.079</td>
<td>Equal to best annotator</td>
</tr>
<tr>
<td>e119</td>
<td>0.238</td>
<td>0.102</td>
<td>0.610</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e123a</td>
<td>0.259</td>
<td>0.043</td>
<td>1.000</td>
<td>0.174</td>
<td>Better than the average</td>
</tr>
<tr>
<td>e123b</td>
<td>0.193</td>
<td>0.093</td>
<td>0.581</td>
<td>0.047</td>
<td>Better than best annotator</td>
</tr>
</tbody>
</table>
Comparison to trivial baselines

- We wanted to see if any baseline methods outperformed the aggregated annotators.

- These are just some ‘dumb’ methods, but we need to know if they perform better than the group.

1. Random
   - Segments are placed randomly, as many as there are in the reference

2. Majority class
   - Each utterance is marked with the most common class; in this case, ‘not a boundary’

3. Segment by last utterance in turn
   - A segment is the last utterance in one speaker’s turn, plus all but the last utterance in the next speaker’s turn
   - Motivated from Sacks, Schegloff, and Jefferson (1974), who noted that the last utterance in a turn was likely to be a ‘first pair part’.
Comparison to trivial baselines

4. The Trigger segments according to the following hand-built rules:

- In order to start a new segment,
  - both speakers must have spoken, and
  - there has to be an utterance of 4 words or less

- When both these conditions are met, the trigger is set.

- The next utterance over 4 words is the start of a new segment.

<table>
<thead>
<tr>
<th>BAT</th>
<th>whatever you prefer whether you prefer like I said a movie or a play or we could go sight seeing somewhere</th>
<th>Start segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRA</td>
<td>actually sight seeing would probably be good</td>
<td>Other speaker has spoken</td>
</tr>
<tr>
<td>BAT</td>
<td>okay that sounds like fun</td>
<td></td>
</tr>
<tr>
<td>BAT</td>
<td>we can check with the tourist bureau as soon as we get there find out what there is to do</td>
<td>An utterance of four words or less has appeared: trigger is set</td>
</tr>
<tr>
<td>KRA</td>
<td>okay</td>
<td>More than four words; this is a new segment</td>
</tr>
<tr>
<td>BAT</td>
<td>okay that looks like we have our trip set then</td>
<td></td>
</tr>
<tr>
<td>BAT</td>
<td>so I will see you when it gets closer to time</td>
<td></td>
</tr>
<tr>
<td>BAT</td>
<td>bye</td>
<td></td>
</tr>
</tbody>
</table>
It turns out that the aggregate volunteer scores were more accurate than baseline methods.
Why does this method work?

One reason might be that it is fault tolerant.

- A single annotator may perform poorly, even when trained.
- Using many annotators, no one annotator needs to perform particularly well.
- No need to strike out near-random or malicious respondents (e.g. number 5).

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Expert</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance 1</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Utterance 2</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>2</td>
</tr>
<tr>
<td>Utterance 3</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>3</td>
</tr>
<tr>
<td>Utterance 4</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Utterance 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Utterance 6</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>3</td>
</tr>
<tr>
<td>Utterance 7</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>2</td>
</tr>
<tr>
<td>Utterance 8</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Utterance 9</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Utterance 10</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>4</td>
</tr>
</tbody>
</table>

WD scores = 0.22 0.22 0.33 0.44 1.00
What do people think about language?

- Getting a lot of data from language users shows us heretofore unsuspected facts about segmentation
- ‘let’s see’
  - We had been grouping ‘let’s see’ utterances with the previous segment
  - Overwhelmingly, annotators group them with the next segment.
- As a group, annotators’ segments seem to fall along topical lines
  - rather than the back-and-forth ‘atoms’ of segments
- Responses become unpredictable when different speakers perform a series of similar speech acts one after the other
  - e.g. two or three ‘information’ utterances in a row
Conclusions

- The ‘satellite/nucleus’ description of dialogue segments is one that annotators can follow
  - Other variations did not lead to consistent results
- Annotation can be done reliably with large numbers of anonymous volunteers
  - People with language experience answer in very similar ways
  - Even the occasional malicious or clueless user does not detract from the annotation
  - Because this project uses a large number of volunteers, like many internet projects, it runs on the power of free time
Conclusions

- Using multiple annotators tells us more about what people think about language use
  - and can lead to insights that may be missed using fewer annotators
  - Getting input from 200 annotators rather than 3 suggests that these results are more representative of language users in general

- Annotator opinion agrees well with expert opinion
  - In some cases, the group performed as well as or better than the best annotator
  - Even though not all annotators match the expert, the combined responses of the group correlate extraordinarily well to our judgments
    - or should that be vice versa?
Future work

Several new areas of study present themselves:

- Examine dialogue structure within dialogue segments
  - What are these segments made up of?
- Turn this data into a machine learning task in which the computer learns to approximate human segmenting behaviour
- Use this information to improve performance on a dialogue act tagger
  - where the goal is to recognise intention in dialogue
- If other corpus linguists are interested, it may lead to the creation of a website where linguists can provide experiments for net users to try
  - May be more successful if presented in the form of language games, puzzles, competitions, and so on.
Thank you for your interest!

If you have any questions, please contact me at dmidgley@cyllene.uwa.edu.au