Paraphrase Fragment Extraction from Monolingual Comparable Corpora

Rui Wang
German Research Center on Artificial Intelligence (DFKI)

Chris Callison-Burch
Johns Hopkins University
Paraphrase Fragments (outputs)

+ Applications
  + Machine translation (Marton et al., 2009)
  + Document summarization (Barzilay et al., 1999)
  + Recognizing textual entailment (Dagan et al., 2005)
  + etc.

+ Sentence-level: e.g., Barzilay and McKeown (2001), Quirk et al. (2004)

+ Sub-sentence-level: e.g., Shinyama et al. (2002), Bannard and Callison-Burch (2005)
Monolingual Comparable Corpora (inputs)

- Monolingual: e.g., Shinyama et al. (2002)
- Comparable: Shinyama et al. (2002)
- Parallel: N/A

- Bilingual: e.g., Bannard and Callison-Burch (2005)

<table>
<thead>
<tr>
<th>MT</th>
<th>Sentence</th>
<th>Sub-sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>N/A</td>
<td>Phrase tables</td>
</tr>
<tr>
<td>Comparable</td>
<td>e.g., Fung and Lo (1998)</td>
<td>e.g., Munteanu and Marcu (2006)</td>
</tr>
</tbody>
</table>
In a Nutshell

- Adapt Munteanu and Marcu (2006)’s approach to monolingual corpora

- Construct a collection of paraphrase fragment pairs with acceptable quality

- Evaluate both the intermediate and final results using the crowd-sourcing technique
Architecture
Data

+ **Source**
  + Gigaword (1995), 6 news agencies

+ **Document pair extraction**
  + MoreLikeThis $>0.9$: 3896 document pairs

+ **Sentence pair extraction**
  + Modified BLEU $0.2-0.8$: 276,120 sentence pairs

+ **Fragment pair extraction**
  + Ca. 90,000 fragment pairs
MTurk Task Description

- Human Intelligence Tasks (HITs)
  - Positive and negative controls

- Document pair extraction
  - Q: Are these two documents about the same topic?
  - A: Yes, No, and Not sure

- Sentence pair extraction
  - Q: Are these two sentences about the same thing?
  - A: Paraphrase, Related, and Non-related
Fragment Pair Extraction

1. Word alignment
   + Identical words

2. Positive and negative associations (Munteanu and Marcu, 2006)

3. Averaging/smoothing with neighboring words

4. Finalization
   + N-gram fragments
   + Linguistic phrases/chunks
If the Bosnian Serbs withdraw their heavy weapons from Sarajevo's outskirts, NATO would be obliged to end its bombardment of Bosnian Serb military installations.

However, Burns said Milosevic unveiled a detailed peace plan calling for the Bosnian Serbs to pull their heavy weapons back from Sarajevo.
MTurk Task Description

+ How are the two highlighted phrases related?
  + These phrases refer to the same thing as each other (PARAPHRASE)
  + These phrases are overlap but contain different information (RELATED)
  + The phrases are unrelated or invalid (INVALID)

+ Controls
  + Positive and negative
## Results

<table>
<thead>
<tr>
<th>Configurations</th>
<th>Aligner+ Phrase Extraction</th>
<th>LCS+ Chunk</th>
<th>Word+ N-Gram</th>
<th>LCS+Word+ Chunk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our Corpus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARAPHRASE</td>
<td>15%</td>
<td>36%</td>
<td>32%</td>
<td></td>
</tr>
<tr>
<td>RELATED</td>
<td>21%</td>
<td>26%</td>
<td>21%</td>
<td></td>
</tr>
<tr>
<td>SUM</td>
<td>36%</td>
<td><strong>62%</strong></td>
<td>53%</td>
<td></td>
</tr>
<tr>
<td><strong>The MSR Corpus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARAPHRASE</td>
<td>38%</td>
<td>44%</td>
<td>49%</td>
<td></td>
</tr>
<tr>
<td>RELATED</td>
<td>20%</td>
<td>19%</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>SUM</td>
<td><strong>58%</strong></td>
<td>63%</td>
<td>67%</td>
<td></td>
</tr>
</tbody>
</table>

Bannard and Callison-Burch (2005): 60%
Zhao et al. (2008): 67%
Examples

+ Our corpus
  + In San Juan, Puerto Rico, Governor Pedro Rosello said the storm could hit the US territory by Friday, ...
  + In Puerto Rico, Gov. Pedro Rossello announced that banks will be open only until 11 a.m. Friday and ...

  + Kunstler rose to fame as the lead attorney for the “Chicago Seven,” ...
  + The highlight of his career came when he defended the Chicago Seven ...
Examples (cont.)

- MSR corpus
  - Iraq’s nuclear program had been dismantled, and there “was no convincing evidence of its reconstitution.”
  - Iraq’s nuclear program had been dismantled and there was no convincing evidence it was being revived, ...

- ... Super Wireless Media Router, which will be available in the first quarter of 2004, at under $200.
- The router will be available in the first quarter of 2004 and will cost around $200, the company said.
  (Related)
## Statistics

<table>
<thead>
<tr>
<th>N-Gram</th>
<th>Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Our</td>
<td>N/A</td>
</tr>
<tr>
<td>MSR</td>
<td>N/A</td>
</tr>
<tr>
<td>CCB</td>
<td>10.7%</td>
</tr>
<tr>
<td>MSR’</td>
<td>8.1%</td>
</tr>
</tbody>
</table>

21% overlapping with Callison-Burch (2008) on MSR

Bannard and Callison-Burch, 2005 (bilingual)
## Statistics (cont.)

<table>
<thead>
<tr>
<th>N-Gram</th>
<th>Para-Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1&amp;1</td>
</tr>
<tr>
<td>Our</td>
<td>N/A</td>
</tr>
<tr>
<td>MSR</td>
<td>N/A</td>
</tr>
<tr>
<td>CCB</td>
<td><strong>34.7%</strong></td>
</tr>
<tr>
<td>MSR'</td>
<td><strong>29.0%</strong></td>
</tr>
</tbody>
</table>
Summary and Future Work

+ Paraphrase fragment pair extraction
  + Acceptable quality
  + Able to scale up

+ MTurk evaluation
  + Quality control
  + Evaluation

+ Future work
  + More data
  + More Syntactic constraints (e.g. Callison-Burch, 2008)
  + Applications
Thank You!

Questions?