ParaMetric: An Automatic Evaluation Metric for Paraphrasing

Chris Callison-Burch
Trevor Cohn    Mirella Lapata
Motivation for Paraphrasing

- Useful in NLP application such as:
  - **Generation** - creating more varied and fluent text
  - **Question answering** - identifying passages that match the expected answer type
  - **Multi-document summarization** - condensing repeated information

- Many recent data-driven approaches
  - Distributional similarity of dependencies (Lin and Pantel)
  - Aligning multiple translations of a foreign text (Barzilay and McKeown, Pang et al)
  - Statistical translation applied to comparable corpora (Quirk et al)
  - Pivoting through foreign phrases (Bannard and Callison-Burch)
Previous approaches to evaluation

- Subjective evaluation
  - Ask judges whether paraphrases
    - “are roughly interchangeable given the genre”
    - “are approximately conceptually equivalent”
    - “preserve meaning and remain grammatical”

- Task-based
  - Apply paraphrases to another task with its own evaluation metric
    - Machine translation
    - Question answering
    - Query expansion
**Goals of ParaMetric**

- Automatic evaluation of paraphrases
  - Objective
  - Repeatable
  - Low cost
  - Task-independent

- Only aiming to evaluate lexical and phrasal paraphrases, not sentential ones
Challenges for Automatic Evaluation

- Developing exhaustive list of paraphrases is difficult / impossible
- Calculating precision and recall is problematic if the list of valid paraphrases is incomplete

\[
Prec = \frac{|G \cap A|}{|A|} \quad \text{underestimate}
\]

\[
Recall = \frac{|G \cap A|}{|G|} \quad \text{relative to incomplete set}
\]
Challenges for Automatic Evaluation

- Context determines whether a particular paraphrase is valid

<table>
<thead>
<tr>
<th>Emma cried and he tried to comfort her.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emma cried and he tried to console her.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>George Bush said Democrats provide comfort to our enemies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Bush said Democrats provide console to our enemies.</td>
</tr>
</tbody>
</table>
Aligning Paraphrases

Develop gold standard paraphrases

- Manually align correspondences in equivalent sentences
- Drawn from multiple reference translations
- Extract paraphrases from manual alignments

<table>
<thead>
<tr>
<th>Some want to impeach him and others expect him to step down.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some people propose to impeach him, while others want him to resign.</td>
</tr>
<tr>
<td>There are those who propose impeaching him and those who want him to tender his resignation.</td>
</tr>
<tr>
<td>Some are proposing an indictment against him and some want him to leave office voluntarily.</td>
</tr>
</tbody>
</table>
Some people propose to impeach him, while others want him to resign.
there are those who propose impeaching him and those who want him to tender his resignation.

some want to impeach him and others expect him to step down.
Some want to impeach him and others expect him to step down. Some are proposing an indictment against him and some want him to leave office voluntarily.
<table>
<thead>
<tr>
<th>Basic paraphrases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>some</strong></td>
</tr>
<tr>
<td><strong>want</strong></td>
</tr>
<tr>
<td><strong>to impeach</strong></td>
</tr>
<tr>
<td><strong>and</strong></td>
</tr>
<tr>
<td><strong>others</strong></td>
</tr>
<tr>
<td><strong>expect</strong></td>
</tr>
<tr>
<td><strong>step down</strong></td>
</tr>
</tbody>
</table>
some people propose to impeach him, while others want him to resign.
some want to impeach him and others expect him to step down.

some people propose to impeach him, while others want him to resign.
<table>
<thead>
<tr>
<th>some</th>
<th>want to</th>
<th>impeach</th>
<th>him</th>
<th>and</th>
<th>others</th>
<th>expect</th>
<th>him</th>
<th>to</th>
<th>step</th>
<th>down</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>some</td>
<td>are proposing</td>
<td>an indictment against</td>
<td>him</td>
<td>and</td>
<td>some want</td>
<td>him</td>
<td>to</td>
<td>leave office</td>
<td>voluntarily</td>
<td>.</td>
</tr>
</tbody>
</table>
### Extended paraphrases

<table>
<thead>
<tr>
<th>some want</th>
<th>some people propose, there are those who propose, some are proposing</th>
</tr>
</thead>
<tbody>
<tr>
<td>some want to</td>
<td>some people propose to</td>
</tr>
<tr>
<td>some want to impeach</td>
<td>some people propose to impeach, there are those who propose impeaching, some are proposing an indictment against</td>
</tr>
<tr>
<td>want to</td>
<td>propose to</td>
</tr>
<tr>
<td>want to impeach</td>
<td>propose to impeach, propose impeaching, are proposing an indictment against</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- A total of 142 non-identical phrase pairs can be extracted from the 3 sentence pairs.
Manually aligned multiple reference translations from Chinese-English set

- Chinese translated by 11 different human translators
- The first reference was paired with each of the other 10 (i.e. 1-2, 1-3, 1-4, ..., 1-11)
- A total of 500 pairs were annotated
- Started with automatic alignments and corrected them
- Avg annotation time was 77 seconds for a total of 11 hours worth of effort

14,000+ unique paraphrases containing 5 words or less
Evaluation Metrics

- Two types of metrics
Evaluation Metrics

- Two types of metrics
- First type:
  - *Alignment precision* and *alignment recall*
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
Evaluation Metrics

- Two types of metrics
- First type:
  - *Alignment precision* and *alignment recall*
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
  - Restricted to methods that can align sentences
Evaluation Metrics

- Two types of metrics

- First type:
  - Alignment precision and alignment recall
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
  - Restricted to methods that can align sentences

- Second type:
  - Enumerate paraphrase lists from manual alignments
  - Calculate precision and recall against these lists
  - Can be applied to any paraphrasing method
Evaluation Metrics

- Two types of metrics

  First type:
  - *Alignment precision* and *alignment recall*
  - Paraphrasing method creates alignments between sentence pairs similar to the manual alignments
  - Precision and recall can be accurately calculated
  - Restricted to methods that can align sentences

- Second type:
  - Enumerate paraphrase lists from manual alignments
  - Calculate precision and recall against these lists
  - Can be applied to any paraphrasing method
  - Provides *lower bound on precision* and a *recall estimate relative to test set*
Applied to three different paraphrasing methods

- Monolingual SMT (Quirk et al 2004)
- Pivoting through foreign phrases in a bilingual corpus (Bannard and Callison-Burch 2005)
- Syntactic alignment (Pang et al 2003)
Dzeirkhanov said 36 people were **injured** and that four people, including **a child**, **had been** hospitalized.

Of the 36 **wounded**, four people including **one child**, **were** hospitalized, Dzheirkhanov said.

- Trained on 100k sentence pairs / 3 million words worth of paired reference translations for AlignPrec and AlignRecall
- Used MSR Paraphrase Phrase Table for LB-Prec and Rel-Recall
I do not believe in mutilating dead bodies.

no soy partidaria de mutilar cadáveres
de inmigrantes ilegales ahogados a la playa...

El mar arroja tantos cadáveres de inmigrantes ilegales ahogados a la playa...

So many of drowned illegals get washed up on beaches...

- Trained on 40 million word Chinese-English parallel corpus for AlignPrec and AlignRecall
- Trained on 10 Europarl parallel corpora for LB-Prec and Rel-Recall
Syntactic Alignment

- Trained on full set of multiple translations for AlignPrec & AlignRecall
- Test set excluded for LB-Prec and Rel-Recall
Results

Paraphrase length = 1

<table>
<thead>
<tr>
<th></th>
<th>Monolingual SMT</th>
<th>Parallel Corpora</th>
<th>Syntactic Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlignPrec</td>
<td>64</td>
<td>54</td>
<td>48</td>
</tr>
<tr>
<td>AlignRecall</td>
<td>56</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>LB-Precision</td>
<td>59</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>Rel-Recall</td>
<td>2</td>
<td>20</td>
<td>16</td>
</tr>
</tbody>
</table>
Results

Paraphrase length <= 5

<table>
<thead>
<tr>
<th>Measure</th>
<th>Monolingual SMT</th>
<th>Parallel Corpora</th>
<th>Syntactic Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlignPrec</td>
<td>73</td>
<td>62</td>
<td>65</td>
</tr>
<tr>
<td>AlignRecall</td>
<td>46</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>LB-Precision</td>
<td>68</td>
<td>14</td>
<td>33</td>
</tr>
<tr>
<td>Rel-Recall</td>
<td>1</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>
Conclusions

- Presented an automatic evaluation metric for paraphrasing
- Overcomes problem of incomplete list of paraphrases by redefining the problem as one of alignment
- Overcomes problem of context by using extracting paraphrases from equivalent sentences
- Allows repeatable, task-independent, objective measure of paraphrase quality
Shameless Plug

- We’ve got an article appearing in Computational Linguistics which details the creation of the data and its other uses.
- Coming soon:

  Constructing Corpora for the Development and Evaluation of Paraphrase Systems
Thanks!