Paraphrasing and Translation
Chris Callison-Burch
16 March 2006

Talk Overview

- Paraphrases
  - What they’re useful for
  - How other people generate them
  - How we do it
- Applying Paraphrases to Translation
  - Problem of unseen words in SMT
  - Using paraphrases to alleviate this
  - Evaluation

Usefulness of paraphrases

- Paraphrases are alternative ways of conveying the same information
- Useful in NLP application such as:
  - Generation - producing paraphrases allows for the creation of more varied and fluent text
  - Multidocument summarization - identifying paraphrases allows information repeated across documents to be condensed
  - Question answering - paraphrasing is important when going beyond simple keyword matching to find answers
  - Machine translation - as we will see later

Paraphrasing with monolingual parallel data

- Previous work by Regina Barzilay and others has focused on monolingual parallel corpora
- Monolingual parallel data comes from multiple translations of the same thing:
  - Multiple translations of classic French novels into English
  - Evaluation data for Bleu method of scoring MT systems
- People have also used comparable corpora (encyclopedia articles on the same topic)

Potential problems with this method

- Parallel monolingual texts are relatively uncommon
- Limits what paraphrases we can generated
  - Limited number of paraphrases
  - Constrained to a few genres

Paraphrasing with bilingual parallel corpora

- Our Methodology:
  - Use statistical MT techniques to align a bilingual parallel corpus
  - Get foreign phrases aligned to the English phrase we want to paraphrase
  - Find other English phrases that foreign phrases align with
  - Treat those English phrases as potential paraphrases, and rank them
- Example:
  - Emma burst into tears and he tried to comfort her, saying things to make her smile.
  - Emma cried, and he tried to console her, adorning his words with puns.
  - Extract burst into tears = cried and comfort = console

More examples

- military force → armed forces, defence, force, forces, peace-keeping personnel, military forces
- sooner or later → at some point, eventually
- great care → a careful approach, greater emphasis, particular attention, specific attention, special attention, very careful
- at work → at the workplace, employment, held, holding, in the work sphere, organised, operate, taken place, took place, working
Paraphrase Probability

- Since we have multiple paraphrases, we rank them with a paraphrase probability

\[
e_2 = \arg \max_{e_2 \in \mathbb{P}} p(e_2 | e_1)
\]

(1)

\[
e_2 = \arg \max_{e_2 \in \mathbb{P}} \sum_{f} p(f | e_2) p(e_2 | f)
\]

(2)

\[
e_2 = \arg \max_{e_2 \in \mathbb{P}} \sum_{f} \frac{\text{count}(f, e_1)}{\sum_{e} \text{count}(e, f)} \frac{\text{count}(e_2, f)}{\text{count}(e_2, f)}
\]

(3)

- Can also rank paraphrases in context by weighting paraphrase probability by language model score

Judging paraphrase quality

- Substituted each paraphrase into 2 - 10 sentences which contained original phrase

<table>
<thead>
<tr>
<th>Under control</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is more, the relevant cost dynamic is completely in check.</td>
</tr>
<tr>
<td>What is more, the relevant cost dynamic is completely checked.</td>
</tr>
<tr>
<td>What is more, the relevant cost dynamic is completely slow down.</td>
</tr>
<tr>
<td>What is more, the relevant cost dynamic is completely curbed.</td>
</tr>
<tr>
<td>What is more, the relevant cost dynamic is completely limit.</td>
</tr>
</tbody>
</table>

- Judged whether new sentences preserved meaning and grammaticality

Results

Judging paraphrase quality

<table>
<thead>
<tr>
<th>Condition</th>
<th>Meaning and Grammaticality</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>automatic alignments</td>
<td>49%</td>
<td>55%</td>
</tr>
<tr>
<td>+ language model</td>
<td>55%</td>
<td>65%</td>
</tr>
<tr>
<td>+ multiple corpora</td>
<td>57%</td>
<td>65%</td>
</tr>
<tr>
<td>+ word sense disambiguation</td>
<td>62%</td>
<td>70%</td>
</tr>
<tr>
<td>manual alignments</td>
<td>75%</td>
<td>85%</td>
</tr>
</tbody>
</table>

Using paraphrases to improve SMT

- Statistical machine translation learns the translations of words and phrases from examples

- Currently if a word is unseen then SMT will be unable to translate it

- If a phrase is unseen, but its individual words are, then SMT won’t be as likely to produce a correct translation for it

We will try to use paraphrases to alleviate this problem

Behavior on unseen words

- A system trained on 10,000 sentences (≈200,000 words) may translate

*Es positivo llegar a un acuerdo sobre los procedimientos, pero debemos encargarnos de que este sistema no sea susceptible de ser usado como arma política.*

- It is good reach an agreement on procedures, but we must **encargarnos** that this system is not susceptible to be **usado** as political weapon.

- Since the translations of **encargarnos** and **usado** were not learned, they are either reproduced in the translation, or omitted entirely.

Substituting paraphrases then translating

- It is good reach an agreement on procedures, but we must **encargarnos** that this system is not susceptible to be **usado** as political weapon.
Improvements in coverage

<table>
<thead>
<tr>
<th>Coverage of</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique 1-grams</td>
<td>48%</td>
<td>92%</td>
</tr>
<tr>
<td>Unique 2-grams</td>
<td>25%</td>
<td>73%</td>
</tr>
<tr>
<td>Unique 3-grams</td>
<td>10%</td>
<td>41%</td>
</tr>
<tr>
<td>Unique 4-grams</td>
<td>3%</td>
<td>20%</td>
</tr>
</tbody>
</table>

For a Spanish-English SMT system trained in 10,000 sentence pairs (approx. 210,000 words in each language), with paraphrases generated from parallel corpora between Spanish and Danish, Dutch, Italian, French, Finnish, German, Greek, Portuguese, and Swedish,

Average quality of translated paraphrase

<table>
<thead>
<tr>
<th>Corpus size (sentences)</th>
<th>Single word Paraphrases</th>
<th>Multi-word Paraphrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>47%</td>
<td>48%</td>
</tr>
<tr>
<td>20,000</td>
<td>61%</td>
<td>52%</td>
</tr>
<tr>
<td>40,000</td>
<td>58%</td>
<td>55%</td>
</tr>
</tbody>
</table>

Prior to paraphrasing none of the unseen words were translating correctly.

Final thoughts

- The data for statistical MT can be used for other tasks, such as paraphrasing
- Paraphrases can be applied to many natural language processing tasks
- Paraphrases can help to overcome the lack of generalization in SMT