Enlarging Paraphrase Collections through Generalization and Instantiation

Atsushi Fujita (Future University Hakodate) fujita@fun.ac.jp
Pierre Isabelle, Roland Kuhn (National Research Council Canada)

Summary

- Paraphrase acquisition
  - Through generalization and instantiation
  - Using both bilingual and monolingual data
- Resources
  - Corpora (bilingual parallel and monolingual)
  - Tokenizer
  - SMT system
  - Lists of stop words
  - (optional) Morphological resources

Translation table

<table>
<thead>
<tr>
<th>Bilingual Parallel Corpus</th>
<th>Seed paraphrases ( P_{\text{Seed}} )</th>
<th>Monolingual Non-parallel Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>health issue ( \rightarrow ) ( \text{problème de santé} )</td>
<td>health problem ( \rightarrow ) ( \text{health issue} )</td>
<td>look like ( \rightarrow ) ( \text{look like} )</td>
</tr>
<tr>
<td>regional issue ( \rightarrow ) ( \text{problème régional} )</td>
<td>regional problem ( \rightarrow ) ( \text{regional problem} )</td>
<td>regional issue ( \rightarrow ) ( \text{regional issue} )</td>
</tr>
<tr>
<td>( \text{problème régional} ) ( \rightarrow ) ( \text{regional problem} )</td>
<td>( \text{look like} ) ( \rightarrow ) ( \text{look like} )</td>
<td>Morphological resources</td>
</tr>
</tbody>
</table>

Step 1. Seed Paraphrase Acquisition

- Pivot-based PA using generic SMT systems
  - e.g., Phrase-based SMT system [Koehn, 03]
  1. Clean up phrase table: sig. pruning [Johnson+, 07]
  2. Pair phrases that get translated to the same phrases [Bannard and Callison-Burch, 05]
  3. Filter paraphrase candidate pairs
    - 3a. stop word differences, word super-sequences
    - 3b. conditional probability and contextual similarity

Step 2. Paraphrase Pattern Induction

- Identical words of LHS and RHS \( \rightarrow \) Variable slots
  - Ignore morphological variation
    - e.g., number (sg./pl.), gender, case, person, tense
  - Related work
    - Develop patterns manually [Jacquemin, 99][Fujita+, 07]
    - Add contextual constraints [Callison-Burch, 08][Zhao+, 09]

Step 3. Paraphrase Instance Acquisition

- Harvest novel instances of the patterns
  1. Collect expressions that match both sides of the pattern
  2. Score each instance by contextual similarity
  - Related work
    - Learn class-dependent patterns [De Saeger+, 09;11]
    - (Pattern-dependent) set expansion

Large multiple of \# of seeds

- (A) Europarl + GigaFrEn, (B) NTCIR Patent data
- One-variable patterns and single words
- High leverage rate (\# pair): \( \geq 1580\% \), \( \geq 2140\% \)
- Paraphrases for many novel phrases

Good quality

- Human evaluation of phrase substitutions
  - Europarl paraphrases on WMT “newstest” data
  - Comparable to the state-of-the-art

<table>
<thead>
<tr>
<th>( n )</th>
<th>5-pt</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{\text{Seed}} )</td>
<td>55</td>
<td>4.60</td>
</tr>
<tr>
<td>( P_{\text{Hast}} )</td>
<td>295</td>
<td>4.22</td>
</tr>
<tr>
<td>Total</td>
<td>350</td>
<td>4.28</td>
</tr>
</tbody>
</table>
### Recipe for contextual similarity

#### Ingredients
- Extract contextual features: adjacent n-grams
  - cf. Bag-of-words (cheap but noisy)
  - cf. Dependency trees (accurate but expensive)
- Weight and filter features: nothing
- Aggregate into a single value: cosine of vectors

#### Examples
- multi-lateral ⇒ multilateral
- i would like to start by congratulating ⇒ let me first of all congratulate
- transitional (task, strategy, phase, costs, …) ⇒ (task, strategy, phase, costs, …) of transition
- in the course of the last few (months) ⇒ during recent (months)
- overall structure ⇒ entire configuration
  - in accordance with the structure mentioned above ⇒ due to such a constitution
  - (bypass, chip) condensers ⇒ (bypass, chip) capacitors
  - will be described with reference to (drawings) ⇒ is explained based on the (drawings)
- [layer, ceramic, ferroelectric, solid, …] condensers ⇒ [layer, ceramic, ferroelectric, solid, …] capacitors
  - will be described with reference to (embodiments) ⇒ is explained based on the (embodiments)

### Additional statistics

- **Paraphrase patterns**
  - Coverage depends on corpus/domain
  - Mostly 1-var patterns
- **Leverage rate**
  - Small bilingual data → High leverage
- **Phrases tend to be short**
  - Our filters tend to discard long phrases
  - Setting: 1-var patterns & single-word fillers

### Human evaluation: details

- Show 5 alternatives at the same time
  - To make results more consistent
  - To reduce the human labor
- “Grammaticality” and “Meaning equiv.”
- 5-pt scales and binary prec. [Callison-Burch, 08]
- [Callison-Burch, 08]
  - Europarl (10 langs-En) + CCG + LM
  - WMT 2007 “newtest” data
  - Binary prec.: .68 for G, .62 for M, .55 for both
  - Ours (total): .76 for G, .71 for M, .58 for both
- [Chan+, 11]
  - Europarl (10 langs-En) + CCG + Google N-gram
  - Europarl (i.e., closed)
  - Score for 1-best: 4.2 pts for G and 3.7 pts for M
  - Ours (1-best): 4.57 pts for G and 3.96 pts for M

### Limitations

- **Our method (current version)**
  - Does not cover totally different expressions
- **Type-based approaches**
  - Do not properly deal with polysemy
  - Tend to miss rare expressions
- **Corpus-based approaches**
  - Do not acquire expressions that do not appear

### Future work

- In-depth analyses of the proposed method
  - Similarity metrics
  - Paraphrase patterns with more than one variable
  - Size & type of monolingual corpora
- Sophisticated paraphrase patterns
  - Hierarchical pattern induction
  - Deeper level of lexical correspondences
- Use for NLP applications