A Class-oriented Approach to Building a Paraphrase Corpus

Atsushi FUJITA(1), Kentaro INUI(2)
(1) Kyoto University
(2) Nara Institute of Science and Technology

Outline
1. Background
2. Issues and our class-oriented approach
3. Semi-automatic example collection
4. Preliminary trials
   1. Specification
   2. Discussion
5. Conclusion

Requirements for handling paraphrases
- Transformation rules / patterns
  - Handcrafting: $X_{\text{hand}} \leftrightarrow S_{\text{X}}(X) \cdot \text{Oper}(S_{\text{X}}(X))$
  - Automatic acquisition: [Iordanajka et al., 1991] [Dra\v{s}, 1999] [Sato et al., 1999] [Kondo et al., 1999] [Kondo et al., 2001] [Iida et al., 2001] etc.
- Coverage is not ensured
- No focus on sorts/variety of paraphrases
- Reliability
- Cost-efficiency

Variety of paraphrases
- Lexical paraphrase ⇒ automatic acquisition
- Syntactic paraphrase ⇒ describable
- Lexically compositional paraphrase ⇒ syntactically regular
- Semantically compositional paraphrase ⇒ descriptively

Building paraphrase corpus
- Issues
  - to consider: variety, source, organization
  - to maximize: coverage, reliability, cost-efficiency
- Previous work
  Manual production
  - [Shirai et al., 2001]
  - [Kinjo et al., 2003]
  - [Shimohata et al., 2004]
  Automatic acquisition
  - [Barzilay et al., 2003]
  - [Shinyama et al., 2002]
  - [Pang et al., 2003]

- Reliability:
  - Cost-efficiency
- Coverage is not ensured
- No focus on sorts/variety of paraphrases
Step 1: Pattern description

- Morpho-syntactic paraphrasing patterns
  - Pairs of dependency trees
  - Implemented on a paraphrase generation system

Step 2: Collect all candidates using a paraphrase engine
Step 3: Judge candidate paraphrases in hand

A class-oriented approach

- Separately collect examples for each class
- Sub-corpus for class A, sub-corpus for class B, sub-corpus for class C...

Outline

1. Overview
2. Issues and our class-oriented approach
   - Semi-automatic example collection
3. Semi-automatic example collection
4. Preliminary trials
   1. Specification
   2. Discussion
5. Conclusion

Aim of this study

- Confirm the feasibility of the method through practice
  - Given
    - A paraphrase class
    - A text collection
  - Collect paraphrase examples belonging to the class
    - At a minimal human labor cost
    - As exhaustively as possible from the text collection
    - As reliable as humanly possible

Semi-automatic example collection

1. Manual description
2. Automatic generation
3. Evaluation and correction

Paraphrase generation system

Judgment guidelines

Paraphrase candidates

Text collection

Lexical resources

Manual description

Automatic generation

Correct

Incorrect

Deferred

Correct

Incorrect

Unchecked

Correct

Deferred

Incorrect

Unchecked

Incorrect
Step 3: Manual judgment (I/F)

(a) source sentence
(b) automatically generated paraphrase
(c) annotator’s judge (correct / incorrect)
(d) error tags
(e) second opinion (correct / incorrect)
(f) free comments.

Target classes
- Paraphrases of light-verb constructions (LVC)
- Transitivity alternation (TransAlt)

Outline
1. Overview
2. Issues and our class-oriented approach
3. Semi-automatic example collection
4. Preliminary trials
   1. Specification
   2. Discussion
5. Conclusion

Results of trials

<table>
<thead>
<tr>
<th>Step</th>
<th>Paraphrase class</th>
<th>LVC</th>
<th>TransAlt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of paraphrasing patterns</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Size of dictionary</td>
<td>20,155</td>
<td>212</td>
</tr>
<tr>
<td>Step 2</td>
<td># of source sentences</td>
<td>10,000</td>
<td>25,000</td>
</tr>
<tr>
<td></td>
<td># of generated candidates</td>
<td>2,566</td>
<td>985</td>
</tr>
<tr>
<td></td>
<td># of judged candidates</td>
<td>1,067</td>
<td>964</td>
</tr>
<tr>
<td></td>
<td># of incorrect candidates</td>
<td>520</td>
<td>503</td>
</tr>
<tr>
<td></td>
<td># of correct candidates</td>
<td>547</td>
<td>461</td>
</tr>
<tr>
<td></td>
<td># of paraphrase examples</td>
<td>591</td>
<td>484</td>
</tr>
<tr>
<td></td>
<td>Working hours</td>
<td>118</td>
<td>169.5</td>
</tr>
</tbody>
</table>

Aim of this study (reminder)
- Confirm the feasibility of the method through practice
- Given
  - A paraphrase class
  - A text collection
- Collect paraphrase examples belonging to the class
  - At a minimal human labor cost
  - As exhaustively as possible from the text collection
  - As reliable as humanly possible

Resources
- LVC
  - 4 paraphrasing patterns (e.g. (7) in paper)
  - 20,155 pairs of \(<\text{deverbal noun, verb}>\)
  - \(<\text{impression}, \text{to be impressed}>\)
  - \(<\text{invitation}, \text{to invite}>\)
- TransAlt
  - 8 paraphrasing patterns (e.g. (10) in paper)
  - 212 pairs of \(<\text{intransitive verb, transitive verb}>\)
  - \(<\text{to sway-intransitive, to sway-transitive}>\)
  - \(<\text{to break-intransitive, to break-transitive}>\)
Cost-efficiency

- Not so wasteful human labor cost
  - 7.1 candidates / man-hour
  - 3.7 paraphrase examples / man-hour
- TransAlt is 1.75 times more difficult than LVC due to test

Exhaustiveness

- The initial resource is not necessarily optimal
  - Paraphrasing patterns
  - Derivation pairs
- How are they optimal?
  - Estimated coverage: 77% (158/(158+48))
    - 158 paraphrases for 750 excerpted sentences
    - Manual examination obtained another 48 paraphrases
    - 47 misses can be salvaged by resource enhancement
    - Errors of shallow parsers hurt only once
  - Use of patterns is realistic approach
  - Manual examination ensures coverage

Reliability

- Strategy
  - Classification bases on guideline & linguistic intuition
  - Inter-annotator discussion refined judgment guidelines
- Agreement ratio increased (in case of LVC)
  - 74% (3rd day) → 77% (6th day)
  → 88% (9th day) → 93% (11th day)
  - It’s still not easy to explain *why this is correct / incorrect*
- Future plan
  - Involve an expert to make sure of judgment guidelines
  - Involve the 3rd annotator for judgment

Conclusion

- Feasibility of a semi-automatic example collection
  - Class-oriented example collection
  - Employing a paraphrase generation system
- Promising results
  - Reasonable human labor cost, but need reduction
  - Moderately exhaustive at initial stage
  - Typically reliable, but some marginal cases
- Paraphrase sub-corpora consist of
  - LVC: 1067 candidates / 591 examples
  - TransAlt: 964 candidates / 484 examples

Future work

- Discussion on required expertise
  - It is not easy to explain *why this is correct / incorrect*
  - Involve an expert to make sure of judgment guidelines
- Build sub-corpora for other paraphrase classes
- Extrinsic evaluation through case studies
  - Resultant provides both correct and incorrect examples
  - Immediately available for analysis and system evaluation
- Publicly open the resource
  - Paraphrase corpus, Lexical resources, Judgment guidelines