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

Purposes of Paraphrase Corpus

- Beneficial to activate the research field
- Deep analysis of phenomena
- Knowledge discovery
- Example-based paraphrasing
- Design better evaluation methods
- Paraphrase rule induction
- Gold-standard for evaluation
- Our aim

Outline

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

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
  - Shinzawa et al., 2002
  - Pang et al., 2003

Reliability: Cost-efficiency
  - Coverage is not ensured
  - No focus on sorts/variety of paraphrases
A class-oriented approach
- Separately collect examples for each class
- Semi-automatic example collection
  - Automatic generation + human judgment
  - **Step 1:** Define a paraphrase class based on morpho-syntactic transformation patterns
  - **Step 2:** Collect all candidates using a paraphrase engine
  - **Step 3:** Judge candidate paraphrases in hand

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

Step 1: Pattern description
- Morpho-syntactic paraphrasing patterns
  - Pairs of dependency trees
  - Implemented on a paraphrase generation system
[Takehshita et al., 2011]
- X: variable for any word
- N: variable for a noun
- V: variable for a verb
- V(N): verbalized form of N

Step 3: Manual judgment (mutual judgment)
- “Correct”-preferred judgment process
  - To reduce labor cost

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

Outline
- Judgment guidelines
- Manual description
- Lexical resources
- Paraphrasing rules
- Text collection
- Paraphrase generation system
- Manual judgment (mutual judgment)
  - 1st annotator
  - 2nd annotator
  - Discussion
  - Candidate paraphrase
  - Correct
  - Incorrect
  - Deferred
  - Correct
  - Incorrect
  - Unchecked
  - Deferred
  - Incorrect
  - Deferred
  - Incorrect
Target classes

Paraphrases of light-verb constructions (LVC)
- The film made an impression on him.
- The film impressed him.

Transitivity alternation (TransAlt)
- The breeze caused the trees to sway.
- The trees were caused to sway by the breeze.

Resources

LVC
- 4 paraphrasing patterns (e.g. (7) in paper)
- 20,155 pairs of \(<\text{deverbal noun, verb}>\)
  - \langle\text{impression}, \text{impress}\rangle
  - \langle\text{invitation}, \text{invite}\rangle

TransAlt
- 8 paraphrasing patterns (e.g. (10) in paper)
- 212 pairs of \(<\text{intransitive verb, transitive verb}>\)
  - \langle\text{sway}, \text{break}\rangle
  - \langle\text{break}, \text{sway}\rangle

Outline

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

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

Results of trials

| Step | Paraphrase class | LVC | TransAlt |
|------|------------------|-----|----------|
| 1    | \# of paraphrasing patterns | 4   | 8        |
|      | Size of dictionary | 20,155 | 212  |
| 2    | \# of source sentences | 10,000 | 25,000 |
|      | \# of generated candidates | 2,566 | 985    |
|      | \# of judged candidates | 1,067 | 964    |
|      | \# of incorrect candidates | 520   | 503    |
|      | \# of correct candidates | 547   | 461    |
| 3    | \# of paraphrase examples | 591   | 484    |
|      | Working hours | 118 | 169.5   |

Working hours: 2 annotators’ working time for (1) Judgment, (2) Discussion, and (3) Re-judgment after refining guidelines
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% (158/210)
  - 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