Expanding Paraphrase Lexicons by Exploiting Lexical Variants

Atsushi Fujita (NICT, Japan; fujita@paraphrasing.org) and Pierre Isabelle (NRC, Canada)

- Given a paraphrase lexicon → Obtain a large number of new paraphrase pairs
  - Generalization: Obtain paraphrase patterns by generalizing corresponding words
  - Instantiation: Collect phrases that match the paraphrase patterns from monolingual corpus
- Promising results
  - Large coverage: 100x expansion is possible, although the rate depends on the size of corpora
  - High accuracy: comparable to state-of-the-art, without harnessing rich resources

**Motivation**

- How to create a paraphrase lexicon with both high coverage and high accuracy?

**Exploiting Lexical Correspondences within Paraphrases**

- **S\textsubscript{Seed}:** seed paraphrase pairs
  - airports in Europe ⇒ European airports
  - amendment of regulation ⇒ amending regulation
  - should be noted that ⇒ is worth noting that

  **Step 1. Learning Paraphrase Patterns**
  - Paraphrase patterns
    
    | Word\textsubscript{1} | Word\textsubscript{2} | Affix\textsubscript{1} | Affix\textsubscript{2} | Stem |
    |----------------------|----------------------|----------------------|----------------------|------|
    | aimed aims          | X:ed X:ts          | 0                   | 0                   | No   |
    | aimed achieve       | X:ed X:ts          | 0                   | 0                   | No   |
    | achieving aims      | X:ing X:ms          | 0                   | 0                   | No   |
    | achieving achieve   | X:ing X:e          | 0                   | 0                   | Yes  |

- **Step 2. Harvesting New Paraphrase Pairs**
  - S\textsubscript{Dev}: new paraphrase pairs
    - cohesion in Europe ⇒ European cohesion
      - European democracy in Europe ⇒ European democracy increase in Haiti ⇒ Haitian increase
      - transportation in suburb ⇒ suburban transportation economy in Uruguay ⇒ Uruguayan economy
      - amendment of documents ⇒ amending documents
      - amendment of protocol ⇒ amending protocol
      - investment of resources ⇒ investing resources
      - recruitment of engineers ⇒ recruiting engineers
      - should be highlighted that ⇒ is worth highlighting that should be reiterated that ⇒ is worth reiterating that should be stated that ⇒ is worth stating that

**Quantity: Large Multiple of Seed Paraphrases**

- English experiments using the approach (B) for obtaining S\textsubscript{Seed}
  - (a) Europarl (En-Fr, 55.7 MW) + News Crawl 2013 (1.20 BW)
  - (b) NTCIR Patent: Parallel (En–ja, 107 MW) + NTCIR En Monolingual 2006-2007 (1.36 BW)
  - Moses, MeCab, SYMGIZA++, UniNe stoplists

**Quality: Reasonably High**

- Human evaluation of paraphrase substitutions
  - Apply paraphrases in (a) to WMT "newstest" data
  - 200 phrases * 3-best paraphrases chosen by 5g LM
  - Unit-wise, 2-phased, classification-based protocol

**Future Work**

- Effective combination of heterogeneous paraphrase lexicons
- Application to further large S\textsubscript{Seed} and monolingual corpus
- Application to various languages
- Integration into NLP applications, e.g., MT