Toward Automatic Compilation of Phrasal Thesaurus
Atsushi Fujita and Satoshi Sato (Nagoya Univ., JAPAN)
{fujita,ssato}@nuee.nagoya-u.ac.jp, http://paraphrasing.org/

Summary
- Phrasal thesaurus: beyond the word-based semantic computing
  - Generating productive paraphrases:
    1. Generate candidate paraphrases
    2. Filter out incorrect instances with a statistical measurement
  - Collecting non-productive paraphrases:
    1. Determine the target vocabulary
    2. Collect their paraphrases (e.g. literal phrases for idioms)

Background & Goal
- Words are not necessarily the appropriate unit of meaning
- Phrasal thesaurus: beyond the word-based semantic computing
- Our approach: Phrasal thesaurus: beyond the word-based semantic computing
- Words are not necessarily the appropriate unit of meaning

Strategy
- Productive
  - Generating productive paraphrases:
    1. Generate candidate paraphrases
    2. Filter out incorrect instances with a statistical measurement
- Collect!!

Non-productive paraphrases
- The lack of discussion on the goal of building a static resource
- Essential from both viewpoints of engineering and lexicography
- Our approach: determine the target + collect their paraphrases
- Idiom/literal paraphrase dictionary
  1. Compile a list of Japanese basic idioms
  2. Collect the counterpart for each basic idiom (ongoing)
- From the gloss in those dictionaries
- From corpus based on DS etc.

Ongoing work
- Verb/VP paraphrase dictionary
- For Sino-Japanese deverbals nouns + * "saru (do)"
- Interaction between predicate phrases and functional expressions
  - With TUTSUJI: a dictionary of Japanese Functional Expressions

Non-productive
- Words are not necessarily the appropriate unit of meaning
- Phrasal thesaurus: beyond the word-based semantic computing
- a natural extension of conventional word-based thesaurus
- interacts with predicate phrases accompanying complements

Productive paraphrases
- Those traditionally represented with transformation patterns
  - Case/voice/verb alternation
  - Category-shifting (nominalization, light-verb construction)
  - Head-switching
- General patterns lead to plenty of incorrect instances

Over-generation step
- Generate candidate paraphrases based on 3 sorts of knowledge
- Trans. Pat. gives a skeleton of syntactic variant at the Synt level.
- Gen. Func. generates a set of the simplest phrases from 0-2 content words.
- Lex. Func. generates different lexical items in certain semantic relations.

Filtering/ranking step
- Measure the quality of phrasal pair (s and t) as paraphrases
  \[ P(t|s) = \frac{P(t)}{\sum_{f \in F} P(f|t)P(f|s)} \]
- Grammaticality factor: structured N-gram language models
- MDS
- CFDS
- Similarity factor: distributional similarity measures
- BOW and MOD features from Web snippets
- Given phrase

NSF Symposium @ New York, 2008.11.15