Toward Automatic Compilation of Phrasal Thesaurus

Atsushi Fujita and Satoshi Sato (Nagoya Univ., JAPAN)

{fujita,ssato}@nuee.nagoya-u.ac.jp, http://paraphrasing.org/

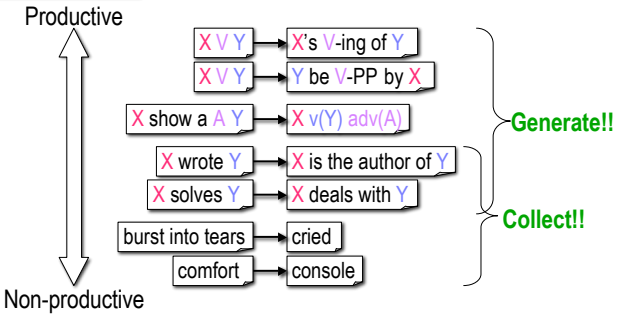
Summary

- 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 idiom)

Background & Goal

- 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
  - deals with predicate phrases accompanying complements

Strategy

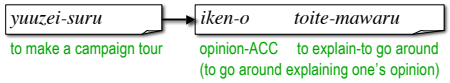


Non-productive paraphrases

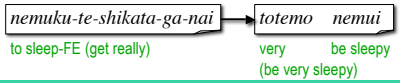
- The lack of discussion on the goal of building a static resource
  - Essential from both viewpoints of engineering and lexicography
- Our approach: **determine the target** + **collect their paraphrases**
  - Idiom/literal paraphrase dictionary
  1. Compile a list of Japanese basic idioms
     - 5 dictionaries for human → comparative list (3,629 idioms)
     - **http://kotoba.nuee.nagoya-u.ac.jp/**
  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 deverbal nouns + "*suru* (do)"

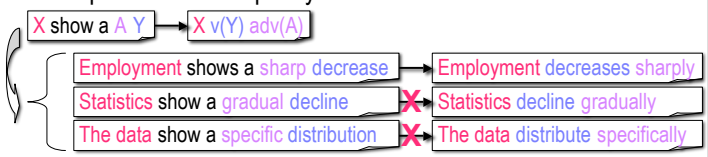


- Interaction between predicate phrases and functional expressions
  - w/ TSUTSUJI: a dictionary of Japanese Functional Expressions [Matsuyoshi & Sato, 08]



Productive paraphrases

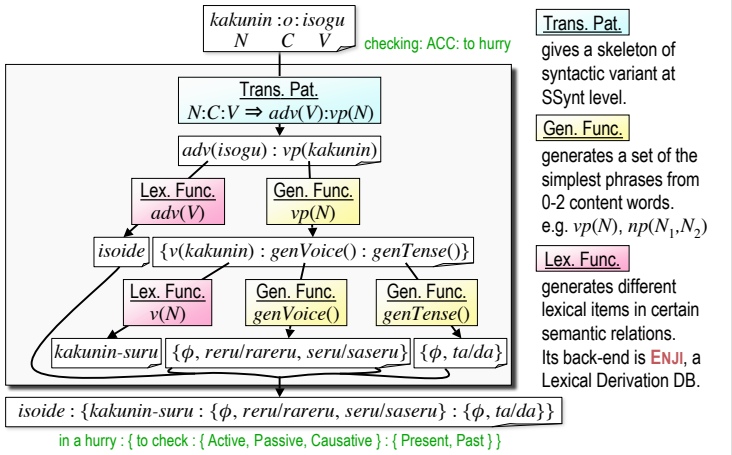
- Those traditionally represented with transformation patterns
  - Case/voice/verb alternation
  - Category-shifting (nominalization, light-verb construction)
  - Head-switching
- General patterns lead to plenty of incorrect instances



- Our approach: **over-generation** + **filtering/ranking**

Over-generation step

- Generate candidate paraphrases based on 3 sorts of knowledge

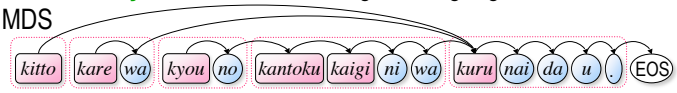


Filtering/ranking step

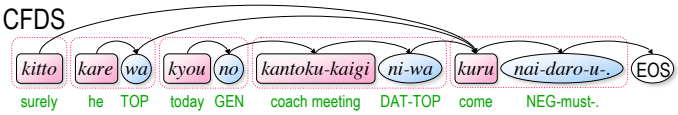
- Measure the quality of phrasal pair (s and t) as paraphrases

$$P(t|s) = P(t) \sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}$$

- **Grammaticality factor**: structured N-gram language models
  - MDS



  - CFDS



- **Similarity factor**: distributional similarity measures
  - BOW and MOD features from Web snippets