Online Acquisition of Japanese Unknown Morphemes using Morphological Constraints

Yugo Murawaki, Sadao Kurohashi (Kyoto University, Japan)

1. A new model of autonomous lexicon acquisition that is integrated into Japanese morphological analysis (no manual intervention is needed)
2. A method of online acquisition of unknown morphemes

Background

The Japanese language is
- agglutinative (like Finnish and Turkish)
- non-segmented (like Chinese and Thai)
- written with several different character types such as hiragana, katakana and kanji (isolated case)

Dictionary-based morphological analysis for Japanese
- High accuracy (~99% F-score)
- Character-type based heuristics to handle unknown morphemes
  - Simple and effective, but far from perfect

Lexicon Acquisition Task

Generate dictionary entries inductively from their examples in texts. We need to identify:
1. stem (front and rear boundaries in character sequences)
2. POS tag (noun, >10 verb types and 2 adjective types)

Experiments

1. Setting
  - Initial lexicon: the default dictionary of JUMAN
  - ~30K entries
  - Data: the first 1000 pages given by a search engine (for 5 queries)
    - 74,572—195,928 sentences

2. Results
  - Acquired morphemes: 107—913
  - Accuracy: 93.9—99.3%
  - Only 4—9 examples were required for acquisition (median)
  - Covered ~50% of detected examples in terms of frequency

Evaluation of “diff” blocks selected from 50 sentences

Future Work

Comprehensive unknown morpheme detection
- Some unknown morphemes are falsely segmented into registered morphemes