Online Acquisition of Japanese Unknown Morphemes using Morphological Constraints

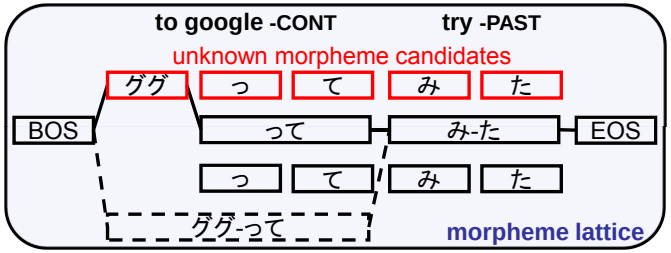
Yugo Murawaki, Sadao Kurohashi (Kyoto University, Japan)

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

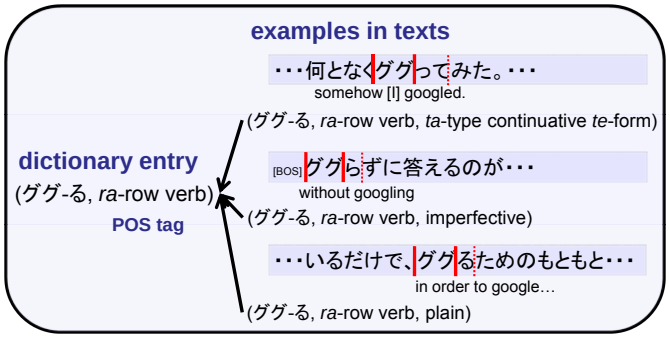
The Japanese language is
- agglutinative (like Finnish and Turkish)
- non-segmented (like Chinese and Thai)
- written with several different character types such as hiragana, katakana and kanji (isolated case)

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



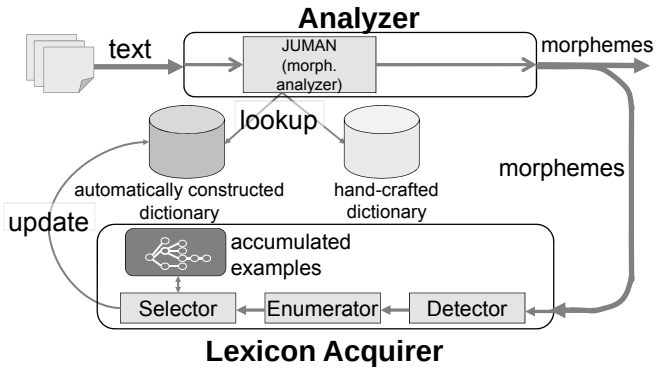
Lexicon Acquisition Task

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)



In previous methods, an extraction module runs independently of the analyzer and in off-line (batch) mode.

Integrated Lexicon Acquisition



Lexicon Acquisition Algorithm

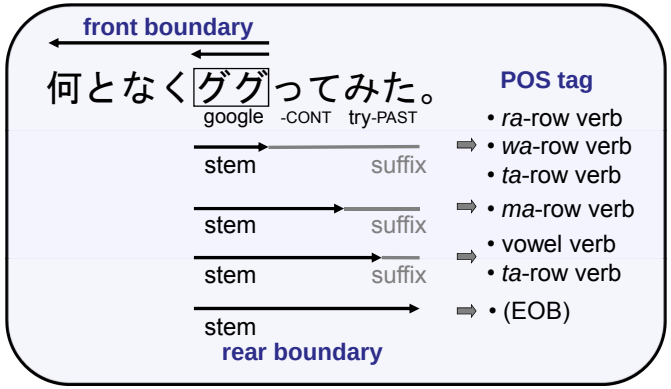
1. Detection

Detect examples of unknown morphemes
- Use special POS tag "Undefined" given by the analyzer

2. Enumeration

Enumerate one or more candidates of the dictionary entry, or the combination of
- front boundary
- rear boundary
- POS tag

Use "**suffixes**" to enumerate rear boundary-POS tag pairs
- Suffixes represent morphological constraints because a stem can be followed only by one of the suffixes specified by its POS tag



3. Selection

1. Accumulate examples that may represent the same entry
2. Select the best candidate that can interpret multiple examples
   - Identify the front boundary, rear boundary and POS tag in this order
3. Update the dictionary if the best candidate satisfies the termination conditions

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

Query	Examples of acquired morphemes
whaling issue	モラトリアム (moratorium), ツチクジラ (giant beaked whale), 混獲 (bycatch)
baby hatch	ダンナ (husband), 助産師 (midwife), 棄て-る (to abandon), 訊-く (to inquire)
JASRAC	ソフ倫 (an organization), シャ乱Q (a pop-rock band), ヲタ (geek)
tsundere	アキバ (abbr. of Akihabara), 腐女子 (*fujoshi*, a slang word), モテる (to be popular)
agaricus	サプリ (abbr. of supplement), アロマ (aroma), 食効 (enhanced nutritional function)

- Effects of acquisition
  - Check the differences of the analyses with the initial lexicon and the augmented lexicon
  - **1.04—15.4%** of sentences were affected by the acquisition

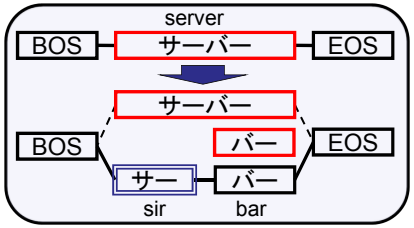
Evaluation of "diff" blocks selected from 50 sentences

Query	segmentation				POS tagging				total
	E→C	C→C	E→E	C→E	E→C	C→C	E→E	C→E	
whaling issue	11	**45**	0	2	11	**45**	0	2	58
baby hatch	**37**	12	9	3	**37**	12	0	3	52
JASRAC	16	**23**	1	12	16	**23**	1	12	52

C: correct; E: error

3. Error analysis

- oversegmentation of katakana morphemes (loan words) by acquired shorter morphemes



Future Work

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