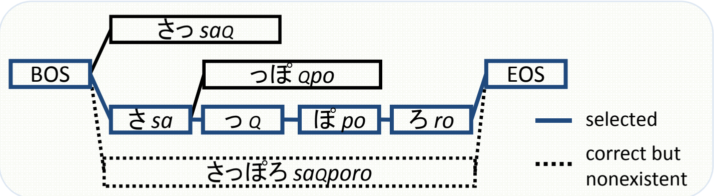
Semantic Classification of Automatically Acquired Japanese Nouns using Lexico-Syntactic Clues

Yugo Murawaki Sadao Kurohashi (Kyoto University, Japan)

1. Lexicon Acquisition from Text

Japanese morphological analysis

- Segmentation + POS tagging
  - Much more ambiguous than POS tagging
- Every morpheme needs to be in the dictionary in order to be selected as the output
- Unknown morpheme problem: Long tail of infrequently used morphemes cannot be maintained manually
- Solution: Automatically find unknown morphemes in raw text and add them to the dictionary



Morpheme Lattice: Morpheme “*saQporo*” need to be added to the dictionary

3. Solution: Two-stage Approach

First Stage: Boundary Identification + Morphological Classification

- Use morphological constraints on what can and cannot follow a morpheme

Constraint: Followed by a particle ⇒ noun

さっぽろを通り越す。 pass through **Sapporo**

Constraint: Followed by a verb ending + an aux. verb ⇒ verb

ぐぐってみた。 have tried to **google**

- Little information about semantic-level distinction
- Hard to use syntactic clues directly because even segmentation is unclear

Second Stage: Semantic Classification of Acquired Nouns

- With acquired nouns, apply morphological analysis + dependency parsing to text
- Use lexical preferences in syntactic relations
  - Reflects semantic-level distinction

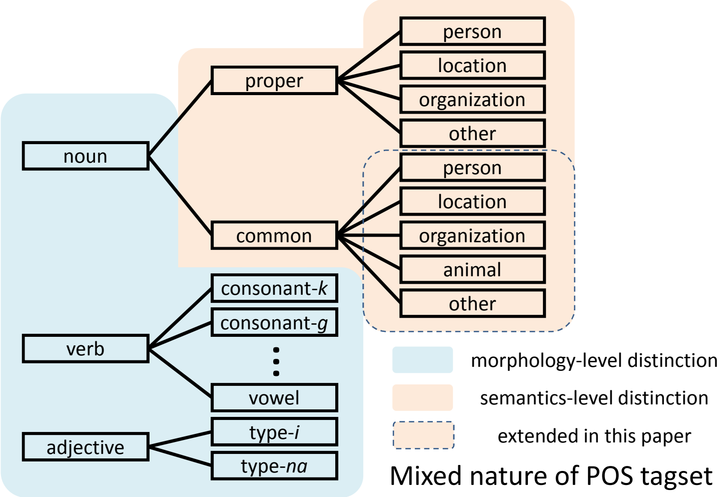
pass through ϕ ⇒ maybe a proper/common noun for place

さっぽろ を 通り越す 。 pass through **Sapporo**

all ϕ ⇒ probably a common noun

すべて の メル友 all **keypal**

2. Challenge in POS Classification



Mixed nature of POS tagset

- When originally designed, POS tags were expected to be assigned manually
- Mixed nature obstructs automation
- Partial POS classification as a compromise [Murawaki+ 2008]
- Demand for full POS classification in applications of morphological analysis
  - e.g. Named Entity Recognition requires semantic information of morphemes

4. Experiments and Discussion

- Multi-class linear classifier
  - Input: Lexico-syntactic clues extracted from multiple occurrences of a noun
  - Output: One of semantic classes
- Manually registered nouns for training
- Classify automatically acquired nouns

Confusion matrix of classification

		Actual								
		Proper nouns				Common nouns				
Predicted		PSN	LOC	ORG	OTH	PSN	LOC	ORG	ANI	OTH
Proper	person	**16**		1		4				1
	location									1
	organization			**4**						
	other									
Common	person	16				**39**			1	2
	location	2	2	1			**10**			4
	organization									2
	animal								**28**	
	other	3	1	1		1	13		9	**338**

- Overall accuracy was **87.0%** (435 / 500)
- Proper nouns often misidentified as common nouns
- Asymmetry between proper and common nouns
  - Common nouns have distinctive usages “ϕ **increase in number**”, “X **become** ϕ” “**many/much** ϕ”, and “**which** ϕ?”
  - Few distinctive usages for proper nouns