Automatically Acquired Lexical Knowledge Improves Japanese Joint Morphological and Dependency Analysis

Daisuke Kawahara Yuta Hayashibe[*1]
Hajime Morita[*2] Sadao Kurohashi[*]

Kyoto University
[*]JST CREST
1 Current affiliation is Fairy Devices Inc.
2 Current affiliation is Fujitsu Laboratories Ltd.

IWPT 2017 (September 21, 2017)

Knowledg
Knowle

Mary **ate** the salad

Mary ate **the salad**

Case frames

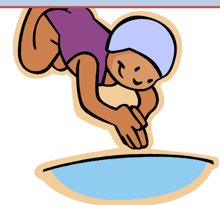
泳ぐ swim

{人 person, 子 child,...}が
{クロール crawl, 平泳ぎ,...}で
{海 sea, 大海,...}を

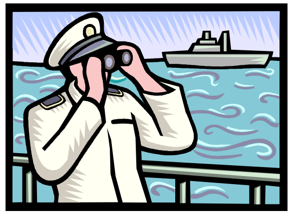
見る see

{人 person, 者,...}が
{望遠鏡 telescope, 双眼鏡 ,...}で
{姿 figure, 人 person,...}を

<u>クロール で</u> **泳いでいる** 女の子 を 見た
crawl swim girl saw



<u>望遠鏡 で</u> 泳いでいる 女の子 を **見た**
telescope swim girl saw



We Need to Segment a Sentence!

クロールで泳いでいる女の子を見た

crawlswimgirlsaw

クロール で 泳いで いる 女の子 を 見た

crawl swim girl saw

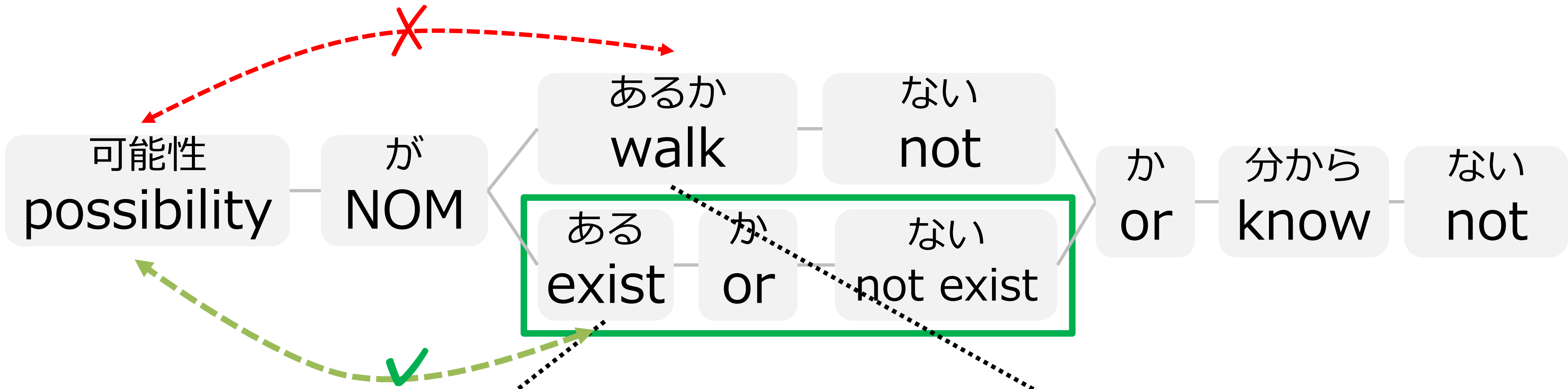
We Need to Segment a Sentence!

- Word segmentation is necessary before applying dependency parsing for unsegmented languages, such as Chinese and Japanese
- Such pipeline framework causes the problem of error propagation
- Several supervised joint models have achieved some success for Chinese but not for Japanese

Question: Can lexical knowledge improve Japanese joint morphological and dependency analysis?

可能性があるかないか分からない

I don’t know whether there is a possibility
or
I don’t know that a possibility doesn’t walk



[Case Frame] ある (exist):3	
が (NOM)	**possibility**:121867
に (DAT)	price:23, myself:20, you:18, ...
で (LOC)	step:4, influence:4, ...

[Case Frame] あるく (walk):1	
が (NOM)	person:57, I:13, ...
を (ACC)	road:24236, trail:4066, ...
から (ABL)	parking:175, station:88, ...

Related work (1/2)

- Joint transition-based parsing
  - POS tagging and parsing [Bohnet+, 2013] [Wang+, 2014]
  - Chinese word segmentation, POS tagging and dependency parsing [Hatori+, 2012] [Zhang+, 2014] [Kurita+, 2017]
- Lattice parsing [Goldberg+, 2009] [Green+, 2010] [Goldberg+, 2011]

Related work (2/2)

- Dependency parsing models using lexical knowledge [van Noord, 2007] [Koo+, 2008] [Chen+, 2009] [Bansal+, 2011]
- Japanese dependency parsing models
  - Transition-based (supervised) models [Kudo+, 2002] [Sassano, 2004] [Yoshinaga+, 2014]
  - Probabilistic model based on case frames [Kawahara+, 2006]

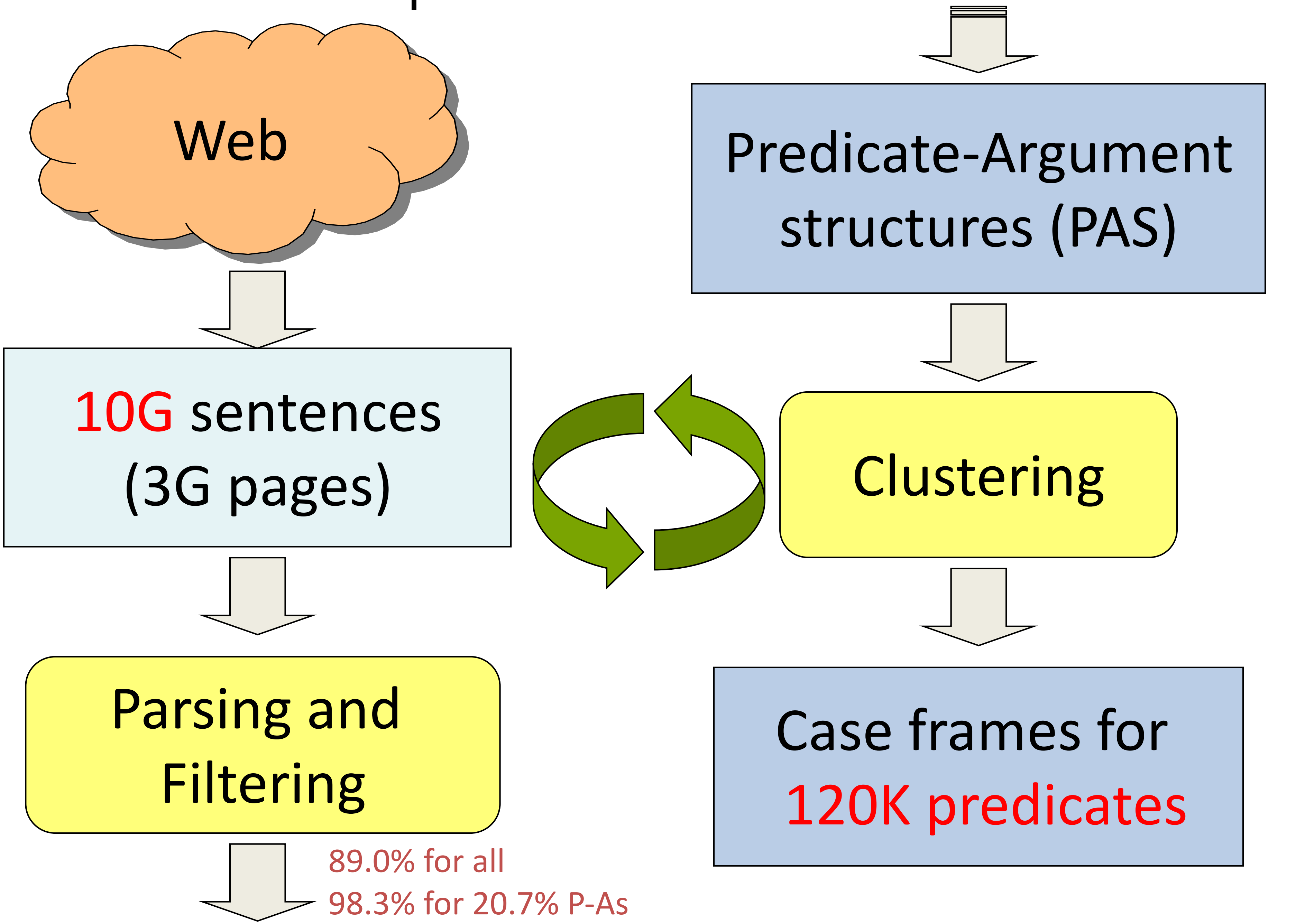
Lexical Knowledge

- Case frames
- Cooccurrence probabilities of noun-noun / predicate-predicate dependencies
- Word embeddings

[Case Frame] ある (exist):3	
が (NOM)	**possibility**:121867
に (DAT)	price:23, myself:20, you:18, ...
で (LOC)	step:4, influence:4, ...

[Case Frame] あるく (walk):1	
が (NOM)	person:57, I:13, ...
を (ACC)	road:24236, trail:4066, ...
から (ABL)	parking:175, station:88, ...

Case Frame Compilation



[Kawahara+, 2006] [Kawahara+, 2014]

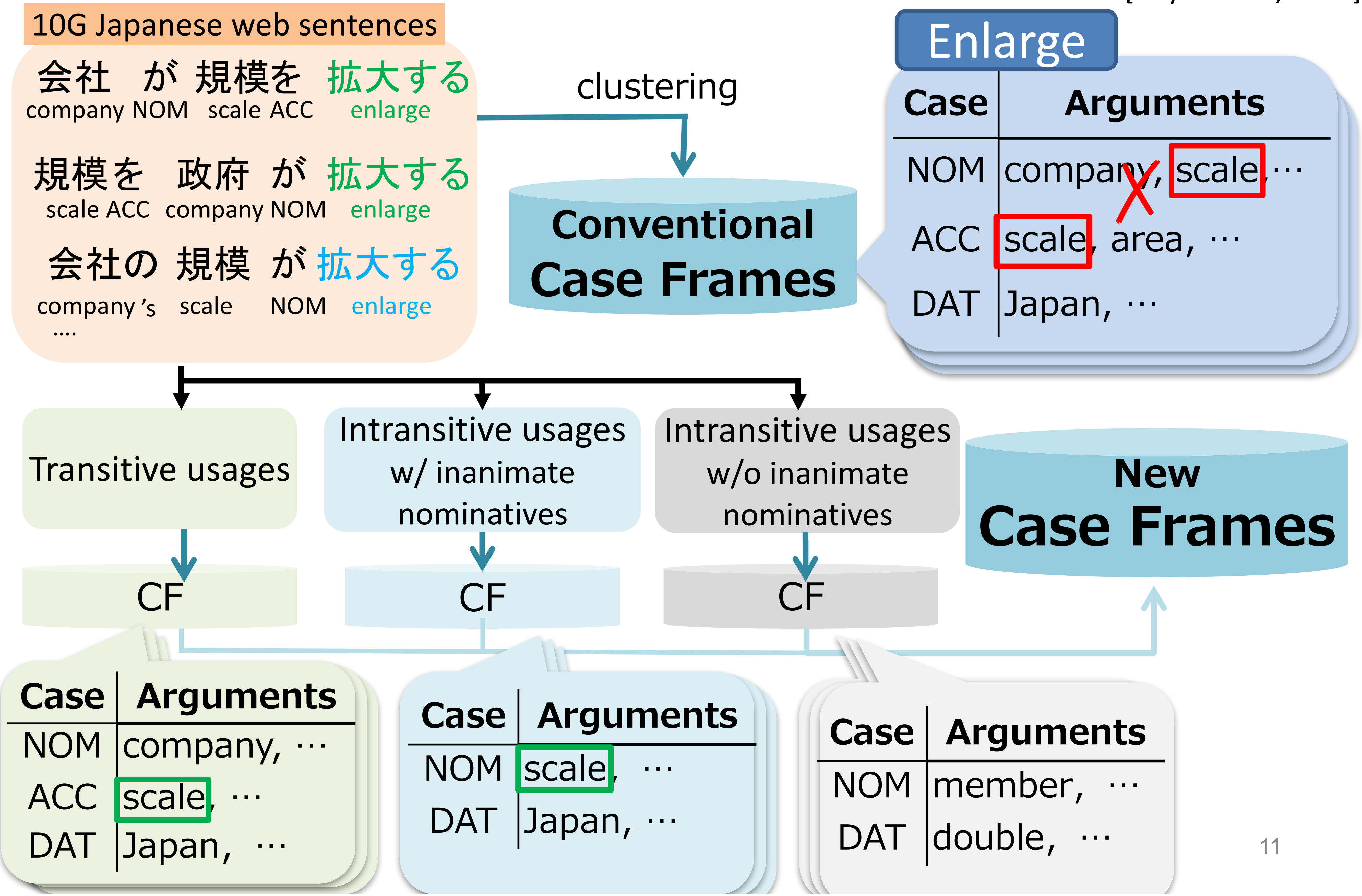
Case frame examples for *tsumu* (積む)

	CS	instances (translated into English)
tsumu (1) (accumulate experience)	*ga*	player:21, all:20, person:142, …
	wo	experience:100127, achievement:10350, …
	de	site:240, area:209, …
tsumu (2) (pursue/ devote)	*ga*	person:27, player:13, all:12, …
	wo	exercise:15579, study:13222, …
	de	basis:694, under:384, university:99, …
tsumu (3) (load)	*ga*	man:33, person:20, child:11, …
	wo	baggage:11294, luggage:2989, …
	ni	car:920, truck:160, bike:114, …
…		

ga: nominative, *wo*: accusative, *ni*: dative, *de*: instrument

Robust Case Frame Compilation

[Hayashibe+, 2015]

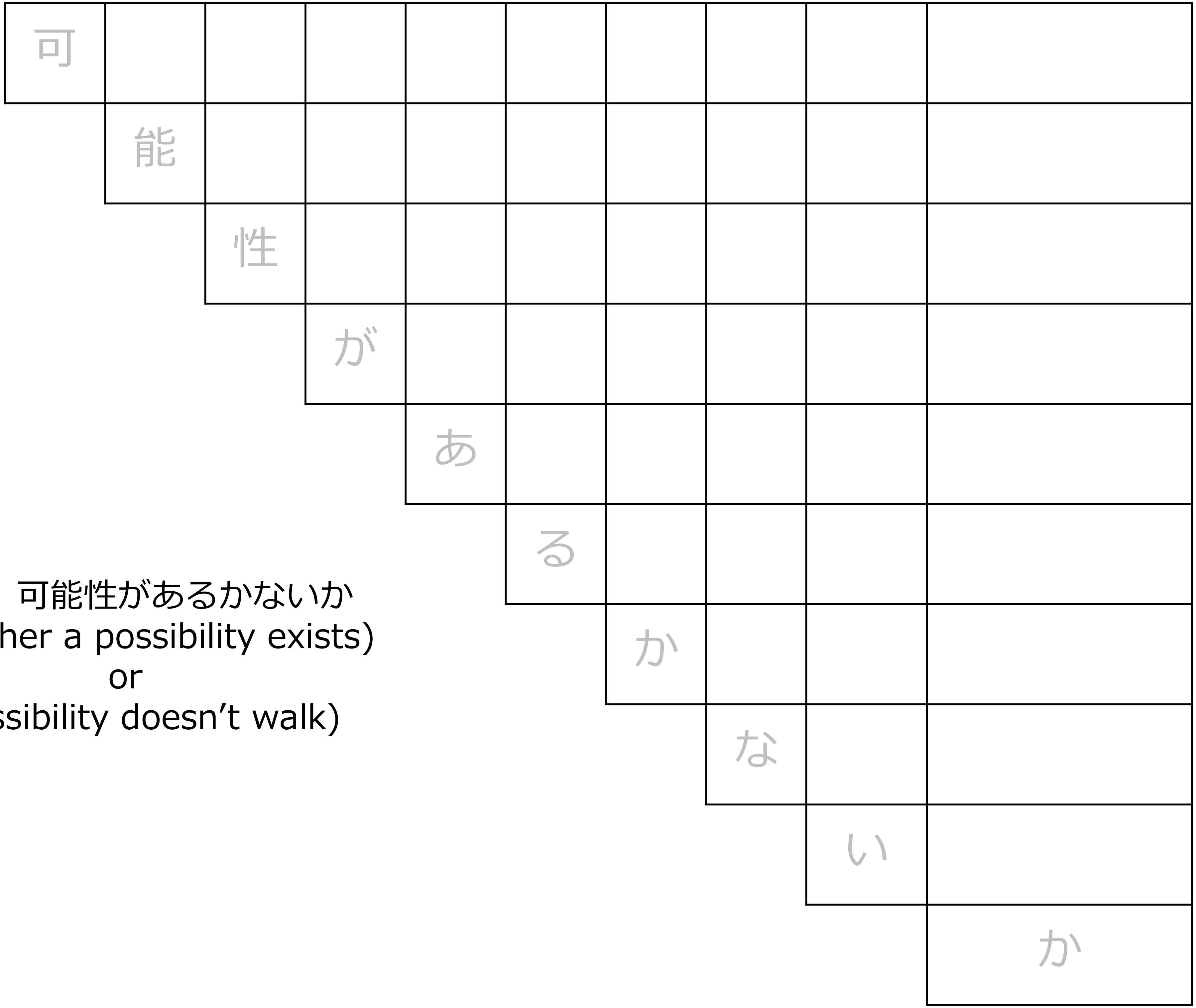


Lexical Knowledge

- Case frames
- Cooccurrence probabilities of noun-noun / predicate-predicate dependencies
  - Calculate $P(predicate_1|predicate_2)$ and $P(noun_1|noun_2)$ from automatic parses
- Word embeddings [Mikolov+, 2013]
  - Clues for coordinate structures

Parsing Model

- Using the well-known CKY algorithm
- Procedure
  1. Project candidate words onto the CKY table
  2. Generate base phrases
     - a base phrase = a content word + 0 or more function words
  3. Generate dependencies and calculate their scores based on lexical knowledge



Input: 可能性があるかないか
(whether a possibility exists)
or
(a possibility doesn’t walk)

可	可能								
	能								
		性							
			が		exist	walk			
				あ	有る	歩か			
					る				
						か			
							な	ない	
								い	
									か

Input: 可能性があるかないか

1. Project candidate words onto the CKY table

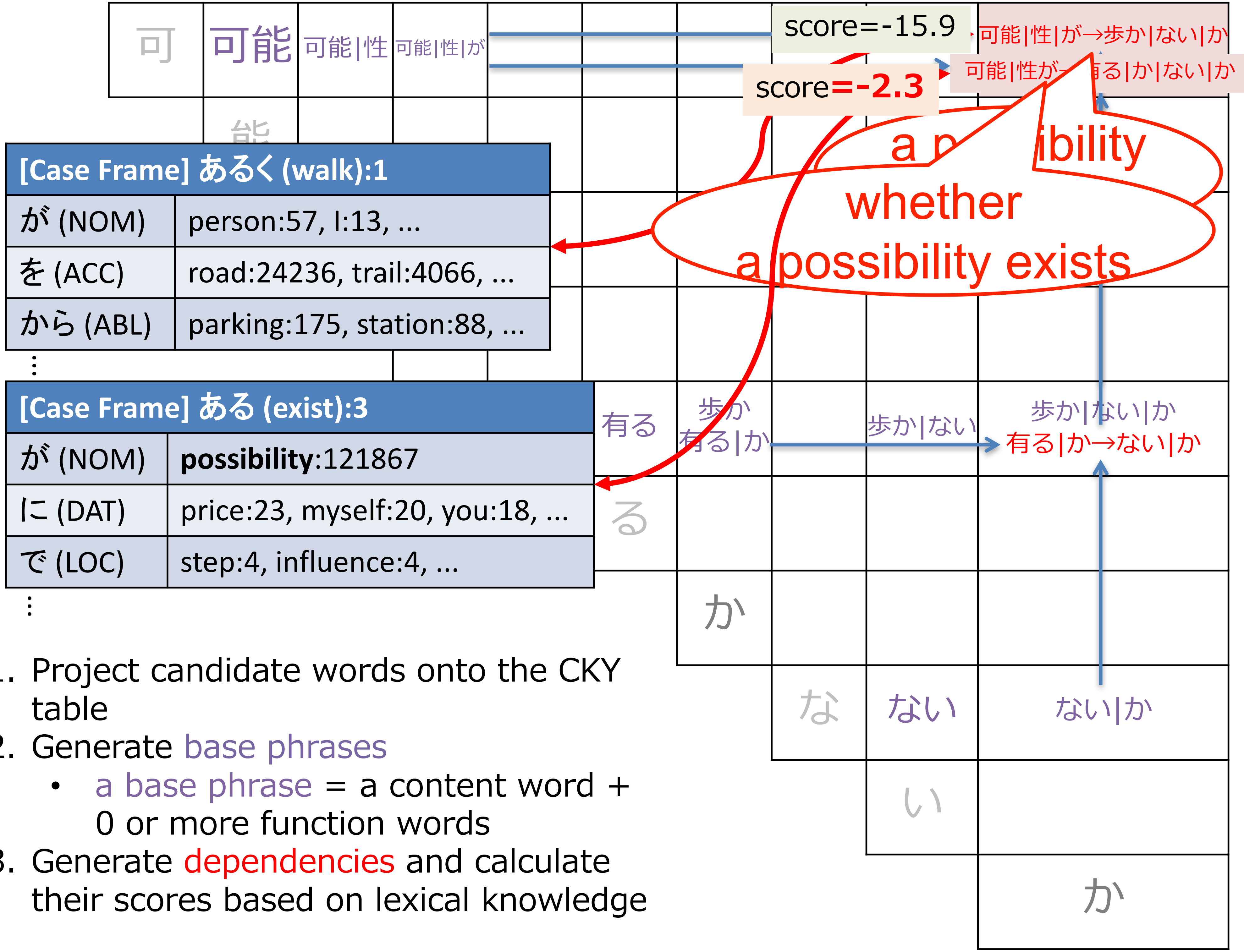
可	可能	可能|性	可能|性|が						
	能								
		性							
			が						
				あ	有る	歩か 有る|か		歩か|ない	歩か|ない|か
					る				
						か			
							な	ない	ない|か
								い	
									か

possibility NOM

doesn’t walk

Input: 可能性があるかないか

1. Project candidate words onto the CKY table
2. Generate base phrases
   - a base phrase = a content word + 0 or more function words



1. Project candidate words onto the CKY table
2. Generate base phrases
    - a base phrase = a content word + 0 or more function words
3. Generate dependencies and calculate their scores based on lexical knowledge

Features (1/2)

- Word feature
  - Marginal score of morphological analysis
- Base phrase features
  - Word 2,3-grams in a base phrase
  - # of base phrases in a sentence
  - Words at a base phrase boundary
  - # of predicates in a sentence
  - A predicate representation
- Dependency features
  - A dependency label
  - Content/function words and punctuations of a modifier
  - Content/function words and punctuations of a head
  - Distance between a modifier and its heads

Features (2/2)

- Features derived from lexical knowledge
  - # of predicates that do not have case frames
  - Probabilities calculated based on case frames
    - e.g., case frame/slot generating probability
  - A cooccurrence probability between nouns
  - A cooccurrence probability between predicates
  - Content word similarity between a modifier and its head
  - Similarity of word sequences for coordination

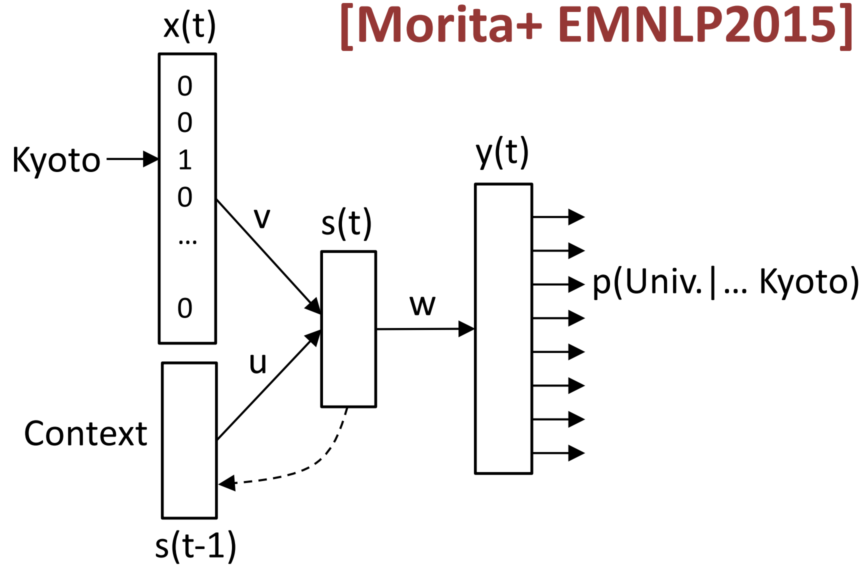
Experimental Settings (1/2)

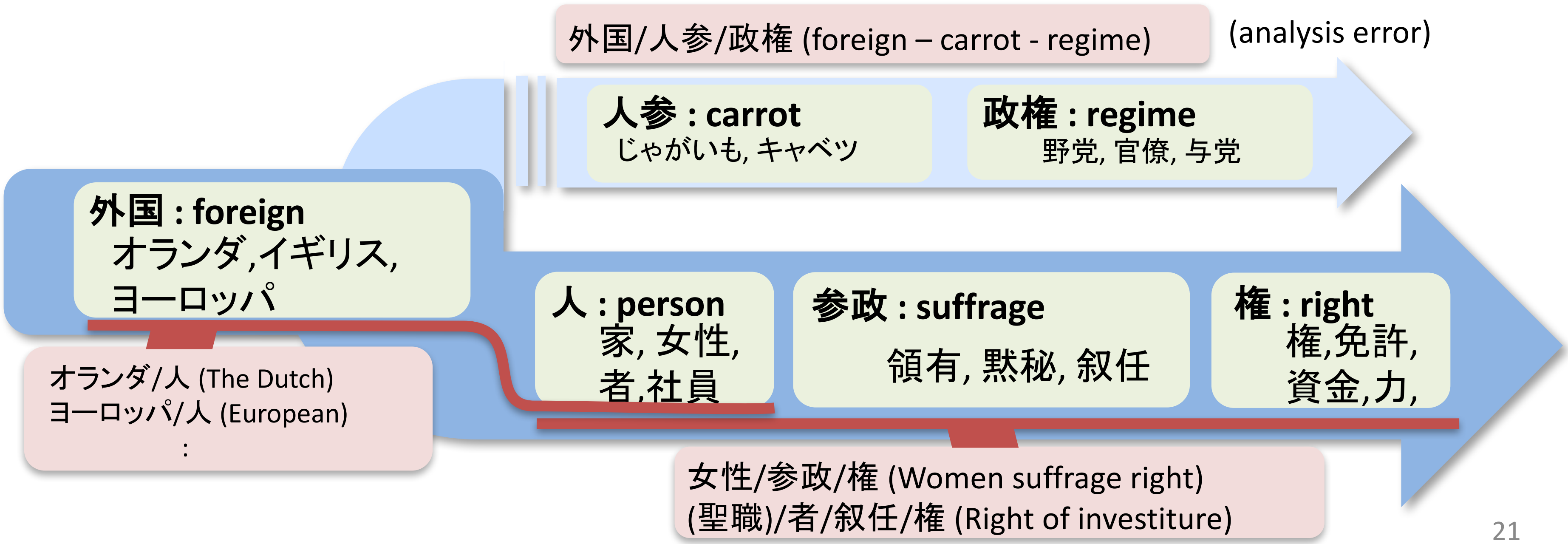
- Dependency treebank
  - Kyoto Univ. Text Corpus (NEWS)
  - Kyoto Univ. Web Document Leads Corpus (WEB)
- Dependency unit
  - Base phrase dependencies
- Input of the parser
  - N-best output of the Japanese morphological analyzer JUMAN++ [Morita+, 2015]

JUMAN++: RNN-based Japanese Morphological Analyzer

[Morita+ EMNLP2015]

- Recurrent Neural Network Language Model [Mikolov+, 2010]
  - A neural network based language model, with a hidden context layer
  - The model can calculate $p(w|context)$ based on **semantically generalized** vector representation





Experimental Settings (1/2)

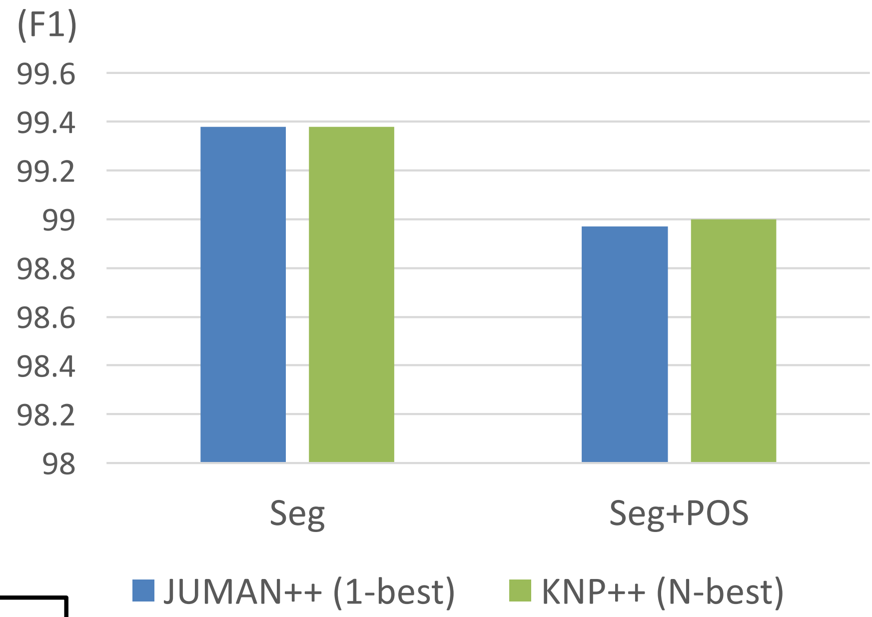
- Dependency treebank
  - Kyoto Univ. Text Corpus (NEWS)
  - Kyoto Univ. Web Document Leads Corpus (WEB)
- Dependency unit
  - Base phrase dependencies
- Input of the parser
  - N-best output of the Japanese morphological analyzer JUMAN++ [Morita+, 2015]
  - Apply 10-way jackkniffing to the training set
- Training of the parser
  - L-BFGS with L1 regularization
- Using beam search
  - Beam width = 10

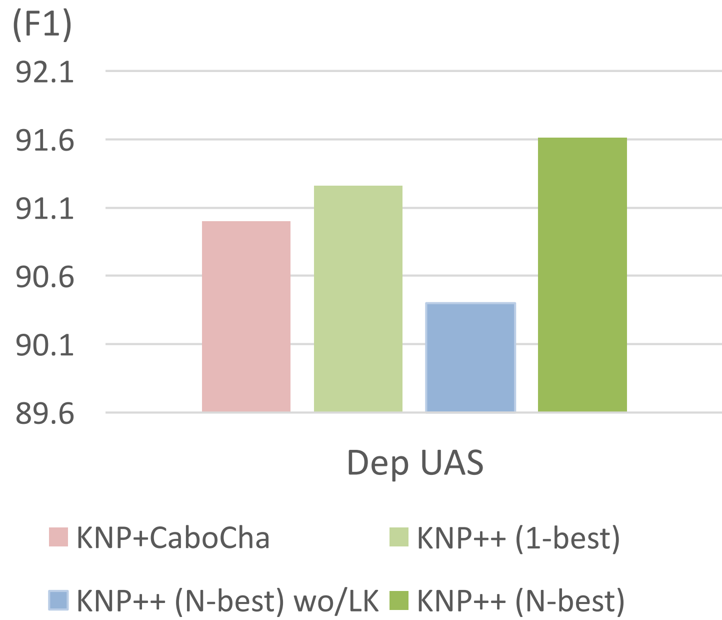
Experimental Settings (2/2)

- Baseline for word segmentation and POS tagging
  - JUMAN++ [Morita+, 2015] (1-best)
- Baselines for dependency parsing
  - KNP [Kawahara+, 2006]
  - CaboCha (using the transition-based algorithm of [Sassano, 2004])
  - KNP+CaboCha
    - Base phrase chunking by KNP and dependency parsing by CaboCha
  - Our model without lexical knowledge (LK)

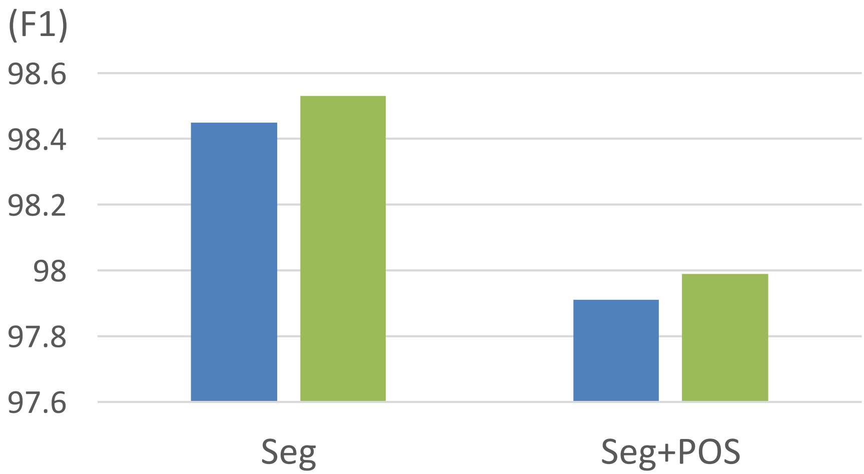
Results

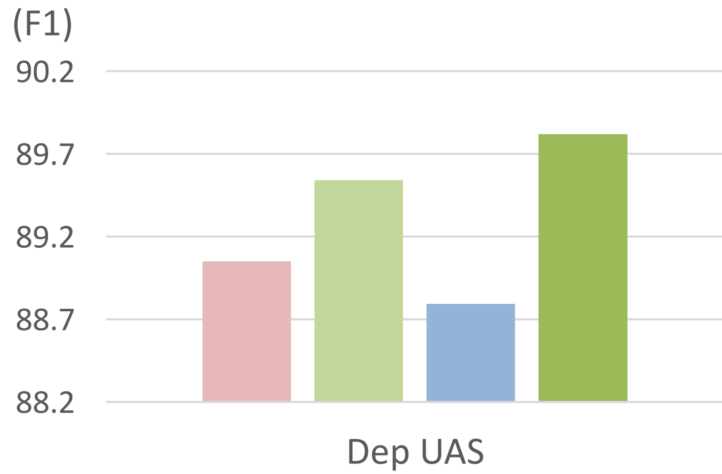
NEWS



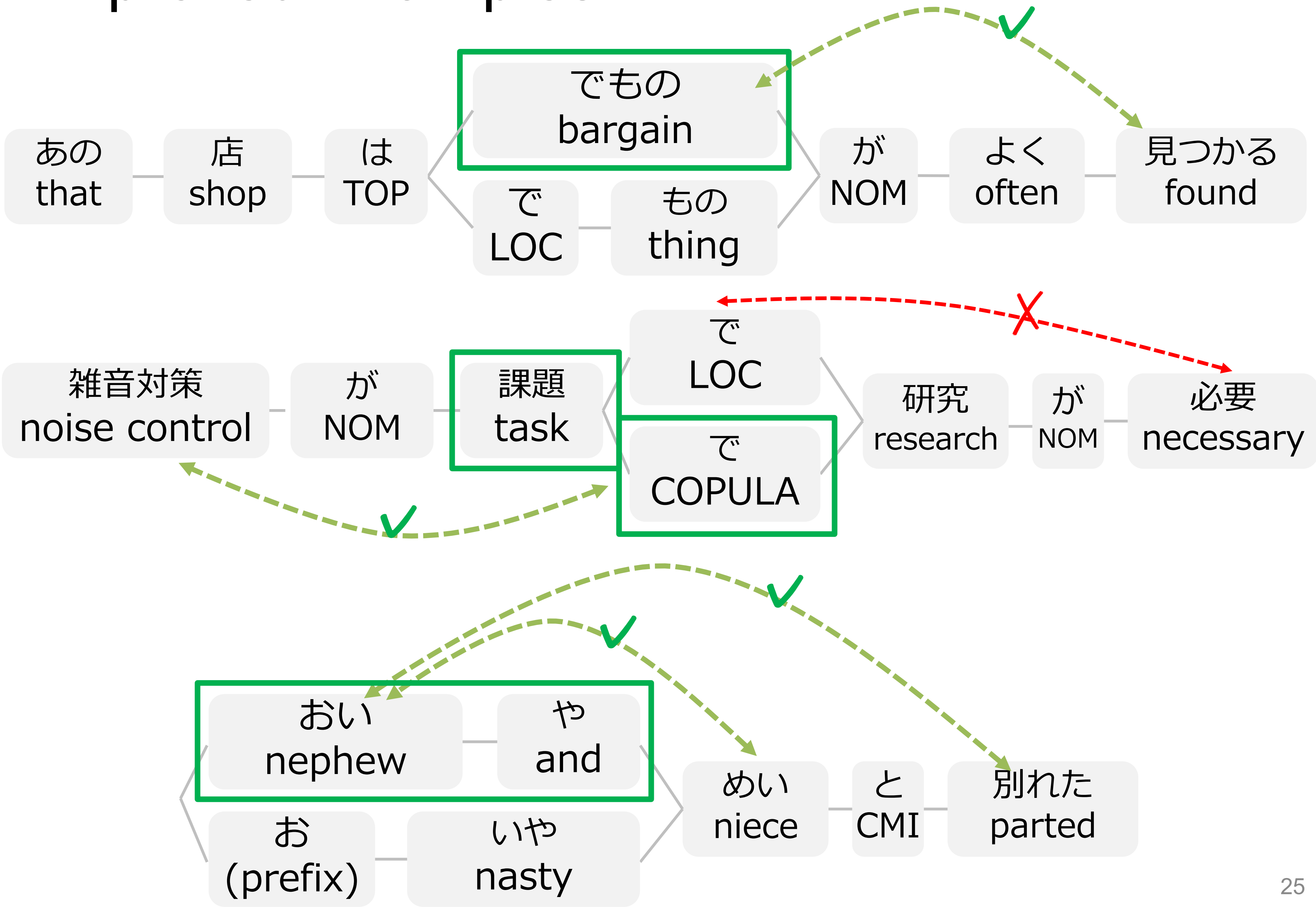


WEB



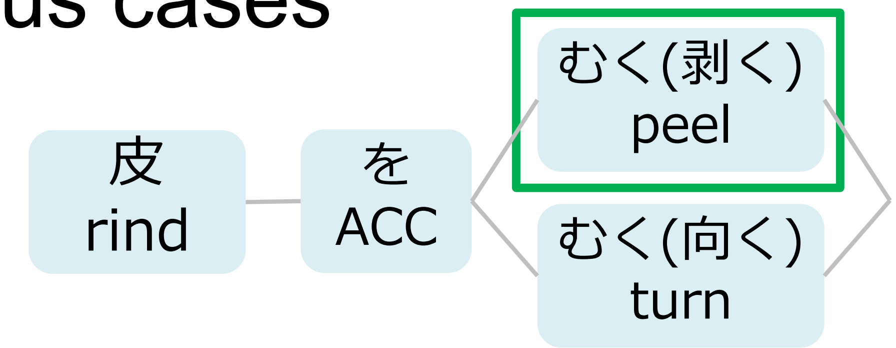


Improved Examples



Discussion

- The 1-best accuracy of segmentation and POS tagging is already very high, especially for NEWS
  - However, we can improve it by reranking N-best outputs based on lexical knowledge, especially for WEB
- The gold does not distinguish some ambiguous cases



Summary

- Automatically acquired lexical knowledge actually improved Japanese joint morphological and dependency analysis!
- We will release lexical resources and analyzers
  - RNN-based Japanese morphological analyzer (JUMAN++)
  - Case frames compiled from 10G Japanese sentences
  - Joint Japanese morphological and syntactic analyzer based on lexical knowledge (KNP++)

Future Work

- Neuralize it!
- Integrate PAS analysis (including zero anaphora resolution) into our joint morphological and syntactic analysis