# Natural Language Processing (6)

## Sequence Labeling and Morphological Analysis

Daisuke Kawahara

Department of Communications and Computer Engineering,
Waseda University

# Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1)
8. Parsing (2)
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1)
13. Guest Talk (2)
14. Project: Survey or Programming
15. Project Presentation

# Sequence Labeling

- A process for assigning labels to data sequence
  - Part-of-speech tagging
  - Named entity recognition
  - Morphological analysis
  - ...

# Part-of-speech (POS) Tagging

- Estimate a grammatical category for each word
- Input: a sentence
- Output: a POS sequence
- Input length = output length

Time flies like an arrow.

noun verb pp det noun ⇒ 光陰矢のごとし

noun noun verb det noun ⇒ 時蠅は矢を好む

# Penn Treebank [Marcus+ 1993]

- Syntactically annotated corpus built by Pennsylvania Univ. in 1990
- Penn Treebank-3 (1999) is widely used now, which contains articles of Wall Street Journal in 1989 (1 million words) and documents of the Brown corpus, annotated with POS and syntactic structure

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

# Penn Treebank POS Tagset

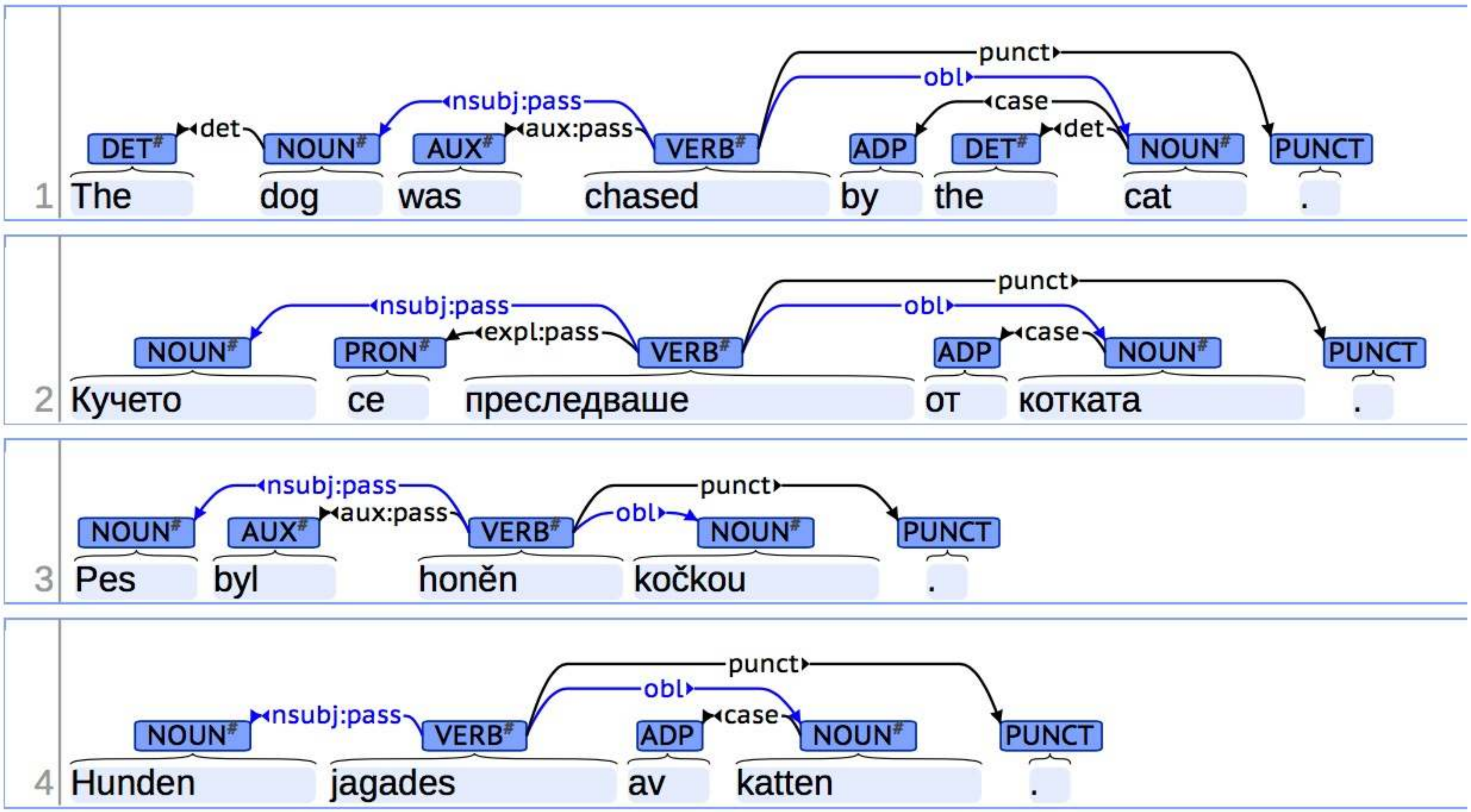
| | |
|---|---|
| CC | Coordinating conjunction |
| CD | Cardinal number |
| DT | Determiner |
| EX | Existential *there* |
| FW | Foreign word |
| IN | Preposition or subordinating conjunction |
| JJ | Adjective |
| JJR | Adjective, comparative |
| JJS | Adjective, superlative |
| LS | List item marker |
| MD | Modal |
| NN | Noun, singular or mass |
| NNS | Noun, plural |
| NNP | Proper noun, singular |
| NNPS | Proper noun, plural |
| PDT | Predeterminer |
| POS | Possessive ending |
| PRP | Personal pronoun |
| PRP$ | Possessive pronoun |

| | |
|---|---|
| RB | Adverb |
| RBR | Adverb, comparative |
| RBS | Adverb, superlative |
| RP | Particle |
| SYM | Symbol |
| TO | *to* |
| UH | Interjection |
| VB | Verb, base form |
| VBD | Verb, past tense |
| VBG | Verb, gerund or present participle |
| VBN | Verb, past participle |
| VBP | Verb, non-3rd person singular present |
| VBZ | Verb, 3rd person singular present |
| WDT | Wh-determiner |
| WP | Wh-pronoun |
| WP$ | Possessive wh-pronoun |
| WRB | Wh-adverb |
| , | Comma |
| . | Sentence-final punctuation |

# Exercise

- Find one POS tagging error in each of the following sentences that are tagged with the Penn Treebank POS tagset.
  1. I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
  2. Does/VBZ this/DT flight/NN serve/VB dinner/NNS
  3. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP

# Universal Dependencies [Nivre+ 2016]




http://universaldependencies.org/introduction.html

# Universal Dependencies POS Tagset

| | |
|---|---|
| ADJ | Adjective |
| ADV | Adverb |
| NOUN | Words for persons, places, things, etc. |
| VERB | Words for actions and processes |
| PROPN | Proper noun |
| INTJ | Interjection |
| ADP | Adposition (Preposition/Postposition) |
| AUX | Auxiliary |
| CCONJ | Coordinating conjunction |

| | |
|---|---|
| DET | Determiner |
| NUM | Numeral |
| PART | Particle |
| PRON | Pronoun |
| SCONJ | Subordinating conjunction |
| PUNCT | Punctuation |
| SYM | Symbol |
| X | Other |

# POS Tag Ambiguities

| | | WSJ | Brown |
|---|---|---|---|
| **Types** | Unambiguous (1 tag) | 86% | 85% |
| | Ambiguous (2+ tags) | 14% | 15% |
| **Tokens** | Unambiguous (1 tag) | 45% | 33% |
| | Ambiguous (2+ tags) | 55% | 67% |

[Jurafsky & Martin 2020]

- Particularly ambiguous common words:
  - *that, back, down, put, set*

# 6 Different POS for *back*

- earning growth took a **back/JJ** seat
- a small building in the **back/NN**
- a clear majority of senators **back/VBP** the bill
- Dave began to **back/VB** toward the door
- enable the country to buy **back/RP** debt
- I was twenty-one **back/RB** then

# Information Sources in POS Tagging

- Syntagmatic structural information
  - e.g., DT JJ NN > DT JJ VBP
    → 77% accuracy

- Lexical information
  - Most frequent tag for each word
  - e.g., *flour* can be used as a verb, but an occurrence of *flour* is much more likely to be a noun
    → 92% accuracy

# Markov Model Tagger

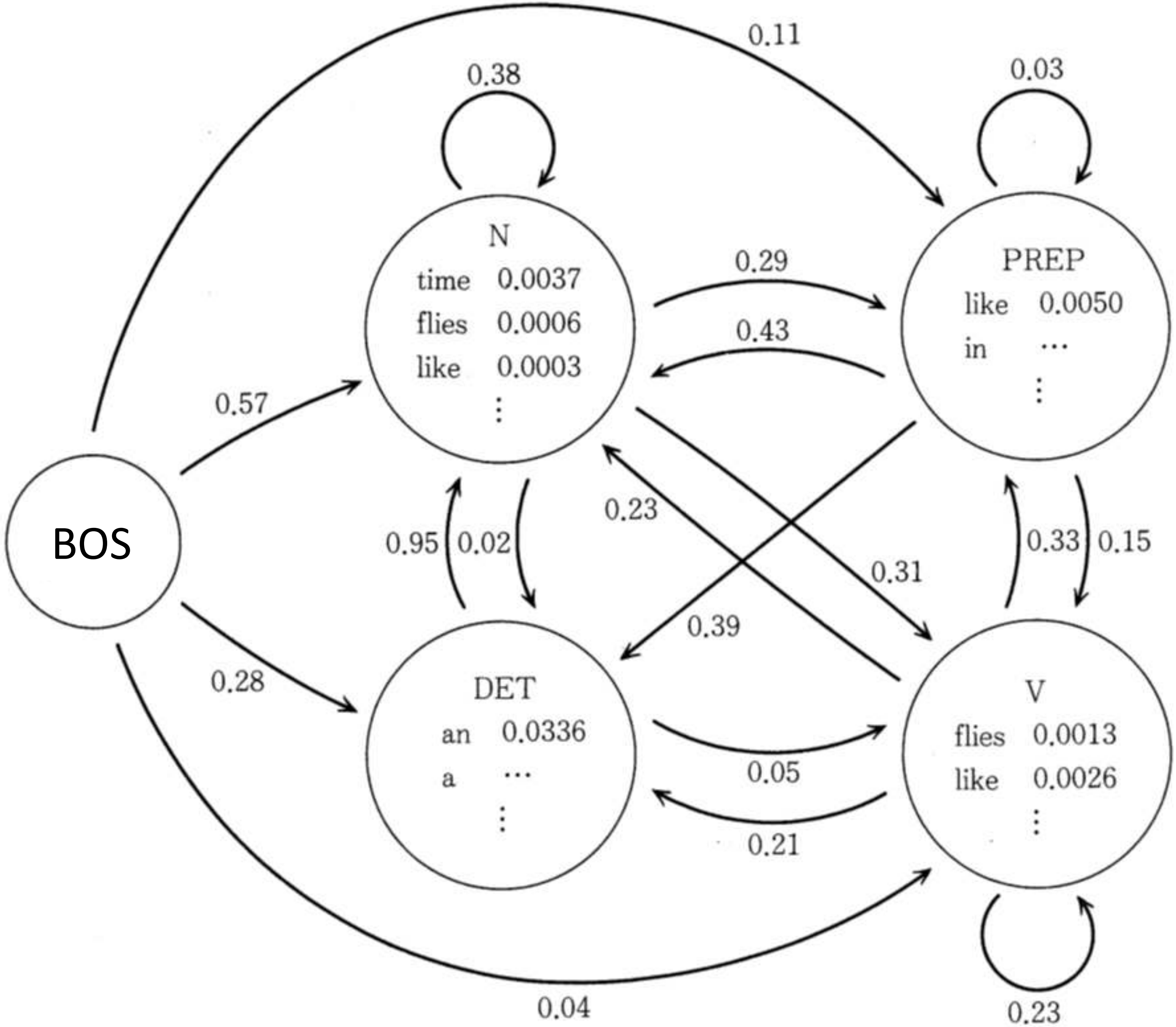
- Find the best tagging:

$$\arg\max_{t_{1,n}} P\left(t_{1,n} \mid w_{1,n}\right) = \arg\max_{t_{1,n}} \frac{P\left(w_{1,n} \mid t_{1,n}\right) P\left(t_{1,n}\right)}{P\left(w_{1,n}\right)}$$

$$= \arg\max_{t_{1,n}} P\left(w_{1,n} \mid t_{1,n}\right) P\left(t_{1,n}\right)$$

$$= \arg\max_{t_{1,n}} \prod_{i=1}^{n} P\left(w_i \mid t_i\right) P\left(t_i \mid t_{i-1}\right)$$

  – Limited Horizon
    $P\left(t_{i+1} \mid t_{1,i}\right) = P\left(t_{i+1} \mid t_i\right)$
  – Words are independent of each other
  – A word's identity only depends on its tag


0.11
0.38
0.03
N
time 0.0037
flies 0.0006
like 0.0003
⋮
0.29
0.43
PREP
like 0.0050
in ⋯
⋮
0.57
BOS
0.95
0.02
0.23
0.31
0.39
0.33
0.15
0.28
DET
an 0.0336
a ⋯
⋮
0.05
0.21
V
flies 0.0013
like 0.0026
⋮
0.04
0.23

# Training

- Maximum Likelihood Estimate using training data

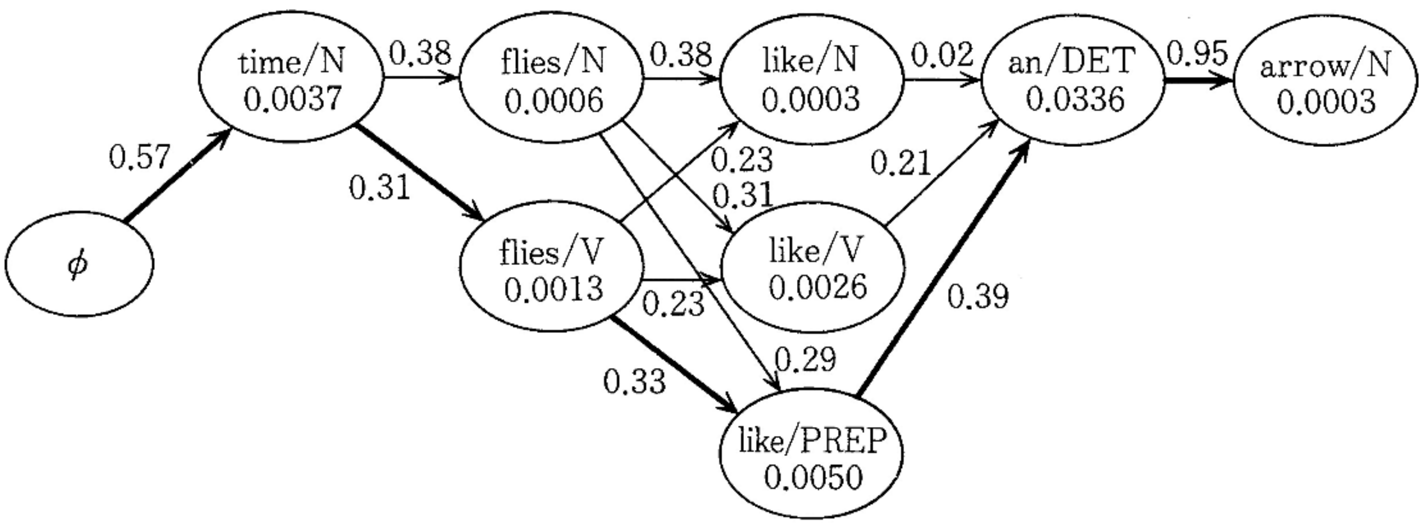
$$P\left(t^k \mid t^j\right) = \frac{C\left(t^j, t^k\right)}{C\left(t^j\right)}$$

$$P\left(w^l \mid t^j\right) = \frac{C\left(w^l, t^j\right)}{C\left(t^j\right)}$$

Time/N flies/V like/PREP an/DET arrow/N

# Tagging: Viterbi Algorithm

- We need to efficiently calculate $\arg\max_{t_{1,n}} P(t_{1,n}|w_{1,n})$
- We can use the Viterbi algorithm, which is based on dynamic programming



# Variations

- Models for unknown words
  - Unknown words can be any part of speech
    → Loss of lexical information
  - Use morphological and other cues

$$P\left(w^l \mid t^j\right) = \frac{1}{Z} P\left(\text{unknown word} \mid t^j\right) P\left(\text{capitalized} \mid t^j\right) P\left(\text{endings} \mid t^j\right)$$

| Feature | Value | NNP | NN | NNS | VBG | VBZ |
|---|---|---|---|---|---|---|
| unknown word | yes | 0.05 | 0.02 | 0.02 | 0.005 | 0.005 |
| | no | 0.95 | 0.98 | 0.98 | 0.995 | 0.995 |
| capitalized | yes | 0.95 | 0.10 | 0.10 | 0.005 | 0.005 |
| | no | 0.05 | 0.90 | 0.90 | 0.995 | 0.995 |
| ending | *-s* | 0.05 | 0.01 | 0.98 | 0.00 | 0.99 |
| | *-ing* | 0.01 | 0.01 | 0.00 | 1.00 | 0.00 |
| | *-tion* | 0.05 | 0.10 | 0.00 | 0.00 | 0.00 |
| | other | 0.89 | 0.88 | 0.02 | 0.00 | 0.01 |

# Variations

- Trigram taggers
  - RB (adverb) can precede both a verb in the past tense (VBD) and a past participle (VBN).
  - "*clearly <u>marked</u>*" is ambiguous in bigram taggers
  - Trigram taggers can disambiguate such cases
    - "*is clearly <u>marked</u>*": VBN
      - VBZ RB VBN > VBZ RB VBD
    - "*he clearly <u>marked</u>*": VBD
      - PRP RB VBD > PRP RB VBN

# Tagging Accuracy Depends on:

- The amount of training data available
  - Over 97% accuracy for Penn Treebank
- The tag set
- The difference between a training corpus and a dictionary on the one hand and the corpus of application on the other hand
- Unknown words

# Examples of Frequent Errors

| Correct tag | Tagging error | Example |
|---|---|---|
| noun singular | adjective | *an <u>executive</u> order* |
| preposition | particle | *He ran <u>up</u> a big …* |
| past tense | past participle | *load <u>needed</u> to meet* |
| past participle | past tense | *load <u>needed</u> to meet* |

Ambiguous sentences:

- <u>The load needed to meet</u> rising costs of health care.
- They cannot now handle <u>the load needed to meet</u> rising costs of health care.

# Confusion Matrix

| Correct Tags | Tags assigned by the tagger DT | IN | JJ | NN | RB | RP | VB | VBG |
|---|---|---|---|---|---|---|---|---|
| DT | 99.4 | .3 | | | .3 | | | |
| IN | .4 | 97.5 | | | 1.5 | .5 | | |
| JJ | | .1 | 93.9 | 1.8 | .9 | | .1 | .4 |
| NN | | | 2.2 | 95.5 | | | .2 | .4 |
| RB | .2 | 2.4 | 2.2 | .6 | 93.2 | 1.2 | | |
| RP | | 24.7 | | 1.1 | 12.6 | 61.5 | | |
| VB | | | .3 | 1.4 | | | 96.0 | |
| VBG | | | 2.5 | 4.4 | | | | 93.0 |

[Manning & Schütze 1999]

# Conditional Random Fields (CRFs)

- We want to use more flexible features than generative models, such as Markov models
- A linear chain CRF:
  - $P\left(y_{1,n} \middle| x_{1,n}\right) = \frac{1}{Z} \exp \sum_i \sum_j \left(\lambda_j f_j\left(x_{1,n}, y_{i-1}, y_i, i\right)\right)$
    - $Z = \sum_{y_{1,n}} P\left(y_{1,n} \middle| x_{1,n}\right)$
  - Feature function: $f_j\left(x_{1,n}, y_{i-1}, y_i, i\right)$
  - By limiting the scope to the current label $y_i$ and the previous label $y_{i-1}$, we can use the Viterbi algorithm

# Feature Functions of CRFs

- Feature templates
  - $\langle y_i, x_i \rangle, \langle y_i, y_{i-1} \rangle, \langle y_i, x_{i-1}, x_{i+1} \rangle$
- Feature functions
  - Input sentence:
    Janet/NNP will/MD <u>back/VB</u> the/DT bill/NN
  - Feature functions for this sentence:
    - $f_{3743}$: $y_i = \text{VB}$ and $x_i = \text{back}$
    - $f_{156}$: $y_i = \text{VB}$ and $y_{i-1} = \text{MD}$
    - $f_{99732}$: $y_i = \text{VB}$ and $x_{i-1} = \text{will}$ and $x_{i+1} = \text{the}$

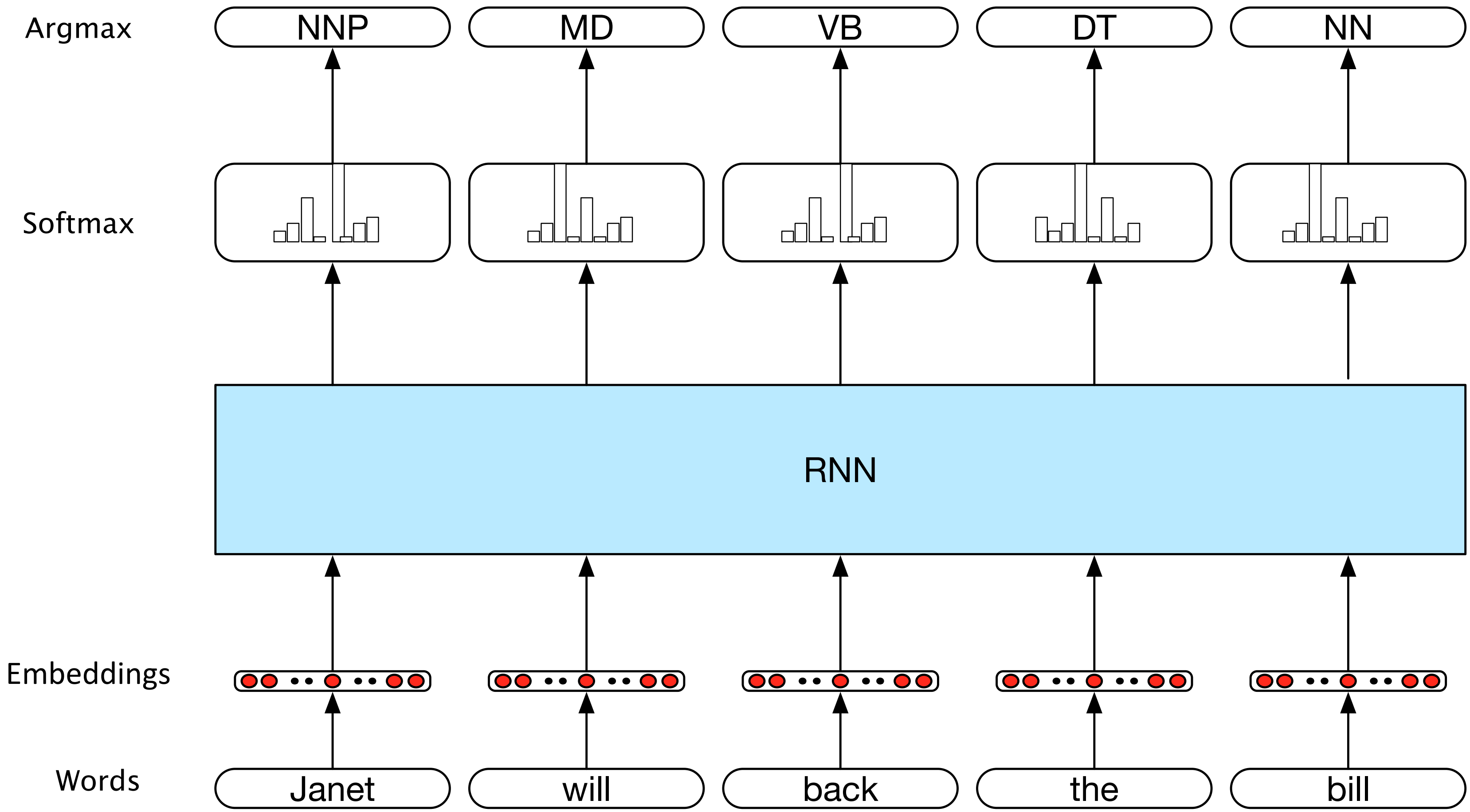
# Features for Unknown Words

- $x_i$ contains a particular prefix (perhaps from all prefixes of length $\leq 2$
- $x_i$ contains a particular suffix (perhaps from all suffixes of length $\leq 2$
- $x_i$'s word shape
- $x_i$'s short word shape

*well-dressed* →

- $\text{prefix}(x_i) = \text{w}$
- $\text{prefix}(x_i) = \text{we}$
- $\text{suffix}(x_i) = d$
- $\text{suffix}(x_i) = ed$
- $\text{word}-\text{shape}(x_i) = \text{xxxx}-\text{xxxxxxx}$
- $\text{short}-\text{word}-\text{shape}(x_i) = \text{x}-\text{x}$

# RNN-based Sequence Labeling




[Jurafsky & Martin 2020]

# Named Entities

- A named entity is anything that can be referred to with a proper name
- Generic named entity types

| Tag | Type | Sample Categories |
|---|---|---|
| PER | People | people, characters |
| ORG | Organization | companies, sports teams |
| LOC | Location | regions, mountains, seas |
| GPE | Geo-Political Entity | countries, states |

- TIME, MONEY, etc. are often added

# Example Document

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

# Named Entity Recognition (NER)

- The task of NER is to find and label **spans** of text
- Issues
  - Segmentation ambiguity
    - What’s an entity and what isn’t
    - Most words are not named entities
  - Type ambiguity
    - e.g., JFK can refer to a <u>person</u> and the <u>airport</u> in New York

# BIO (and BIOES) Tagging Scheme

| Words | BIO Label | BIOES Label |
|---|---|---|
| Jane | B-PER | B-PER |
| Villanueva | I-PER | E-PER |
| of | O | O |
| United | B-ORG | B-ORG |
| Airlines | I-ORG | I-ORG |
| Holding | I-ORG | E-ORG |
| discussed | O | O |
| the | O | O |
| Chicago | B-LOC | S-LOC |
| route | O | O |
| . | O | O |

# Features for (CRF-based) NER

- Please think about features for NER
  - Identities of $w_i$ and neighboring words
  - Embeddings of $w_i$ and neighboring words
  - POS of $w_i$ and neighboring words
  - Presence of $w_i$ in a gazetteer
  - $w_i$ contains a particular prefix
  - $w_i$ contains a particular suffix
  - Word shape of $w_i$ and neighboring words
  - Short word shape of $w_i$ and neighboring words

# Some NER features for a sample sentence

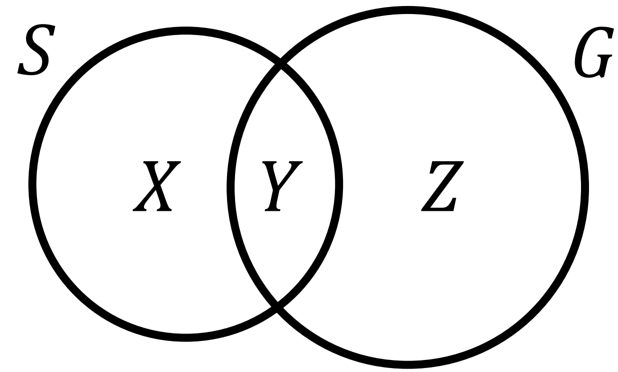
| Words | POS | Short shape | Gazetteer | BIO Label |
|---|---|---|---|---|
| Jane | NNP | Xx | 0 | B-PER |
| Villanueva | NNP | Xx | 1 | I-PER |
| of | IN | x | 0 | O |
| United | NNP | Xx | 0 | B-ORG |
| Airlines | NNP | Xx | 0 | I-ORG |
| Holding | NNP | Xx | 0 | I-ORG |
| discussed | VBD | x | 0 | O |
| the | DT | x | 0 | O |
| Chicago | NNP | Xx | 1 | B-LOC |
| route | NN | x | 0 | O |
| . | . | . | 0 | O |

# Evaluation of NER

- Precision
  - Ratio of the number of correctly labeled responses to the total labeled (output)
- Recall
  - Ratio of the number of correctly labeled responses to the total that should be labeled
- F1
  - Harmonic mean of precision and recall

# Precision, Recall, and F1

- $\text{Precision} = \frac{Y}{X+Y}$
- $\text{Recall} = \frac{Y}{Y+Z}$
- $F1 = \frac{1}{\frac{\frac{1}{P}+\frac{1}{R}}{2}} = \frac{2PR}{P+R}$



# Japanese Morphological Analysis

外国人参政権


※ Currently, this is translated correctly.

# Japanese Morphological Analysis

- Word boundary is not obvious!

  e.g., 外国人参政権

  くるまで待つ

- Joint process of word segmentation and POS tagging = Morphological Analysis
- Three components:
  - Dictionary
  - Connection matrix
  - Decoding algorithm

# Basic Word Dictionary (JUMAN)

…
(名詞 (普通名詞 ((読み からくさ)(見出し語 唐草 (から草 1.6) (からくさ 1.6))(意味情報 “代表記:唐草/からくさ"))))
(名詞 (普通名詞 ((読み からくち)(見出し語 辛口 (から口 1.6) (からくち 1.6))(意味情報 “代表表記:辛口/からくち"))))
(副詞 ((読み からくも)(見出し語 辛くも からくも)(意味情報 “代表表記:辛くも/からくも")))
(名詞 (普通名詞 ((読み からくり)(見出し語 からくり)(意味情報 “代表表記:からくり/からくり"))))
(動詞 ((読み からす)(見出し語 枯らす からす)(活用型 子音動詞サ行)(意味情報 “代表表記:枯らす/からす")))
(名詞 (普通名詞 ((読み からす)(見出し語 烏 カラス (からす 1.6))(意味情報 “代表表記:烏/からす"))))
(名詞 (普通名詞 ((読み からだ)(見出し語 身体 体 (からだ 1.6))(意味情報 “代表表記:身体/からだ"))))
(名詞 (普通名詞 ((読み からだつき)(見出し語 体付き 体付 体つき (からだつき 1.6))(意味情報 “代表表記:体付き/からだつき"))))
(名詞 (普通名詞 ((読み からっかぜ)(見出し語 空っ風 (からっかぜ 1.6))(意味情報 “代表表記:空っ風/からっかぜ"))))
(副詞 ((読み からっきし)(見出し語 からっきし)(意味情報 “代表表記:からっきし/からっきし")))
…

# Dictionary (JUMAN)

| | Vocab Size | Word Examples |
|---|---|---|
| **Basic Word** | 30K | 走る, 行く, 明日 |
| **Wikipedia** | 850K | アベノミクス, Dentsu, 山極, 豊洲 |
| **Wiktionary** | 8K | インセンティヴ, 糾す |
| **Web** | 10K | ググる, ねんどろいど |
| **Total** | 900K | |

# Connection Matrix (JUMAN)

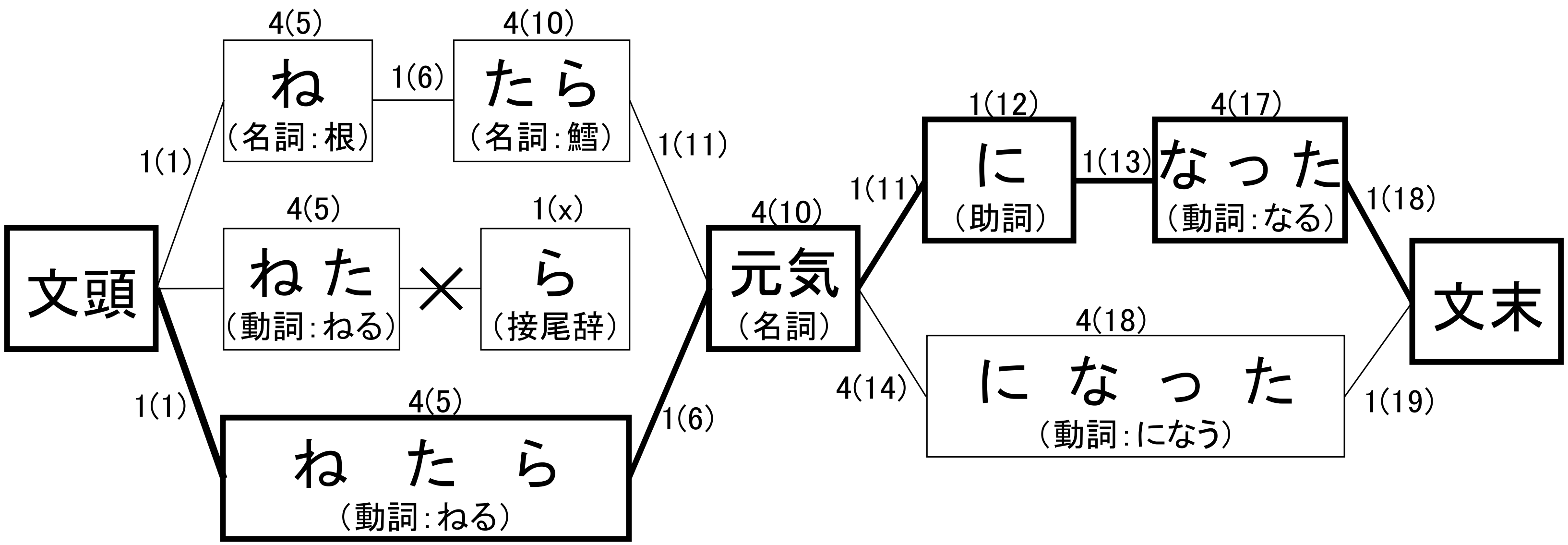
```
・・・

((BunsetsuEndSentenceEnd
  BunsetsuEnd
  (助詞 接続助詞 * * の))
      ((名詞))
4

((VerbBasicForm
  IAdjBasicForm
  NaAdjAllBasicForm
  AuxBasicForm
  NaAdjGuessForm
  (* * * タ系推量形)
  (動詞 * * タ系連用テ形)
  (接尾辞 動詞性接尾辞 * タ系連用テ形))
      ((助詞 接続助詞 * * から)))
・・・
```

# Connection Matrix: Cost Setting

- By human
  - JUMAN
- Variable Memory Markov Model (VMMM)
  - ChaSen
- Conditional Random Fields (CRFs)
  - MeCab
- Support Vector Machines (SVMs)
  - KyTea
- Online learning (confidence weighted) + RNN language model
  - Juman++

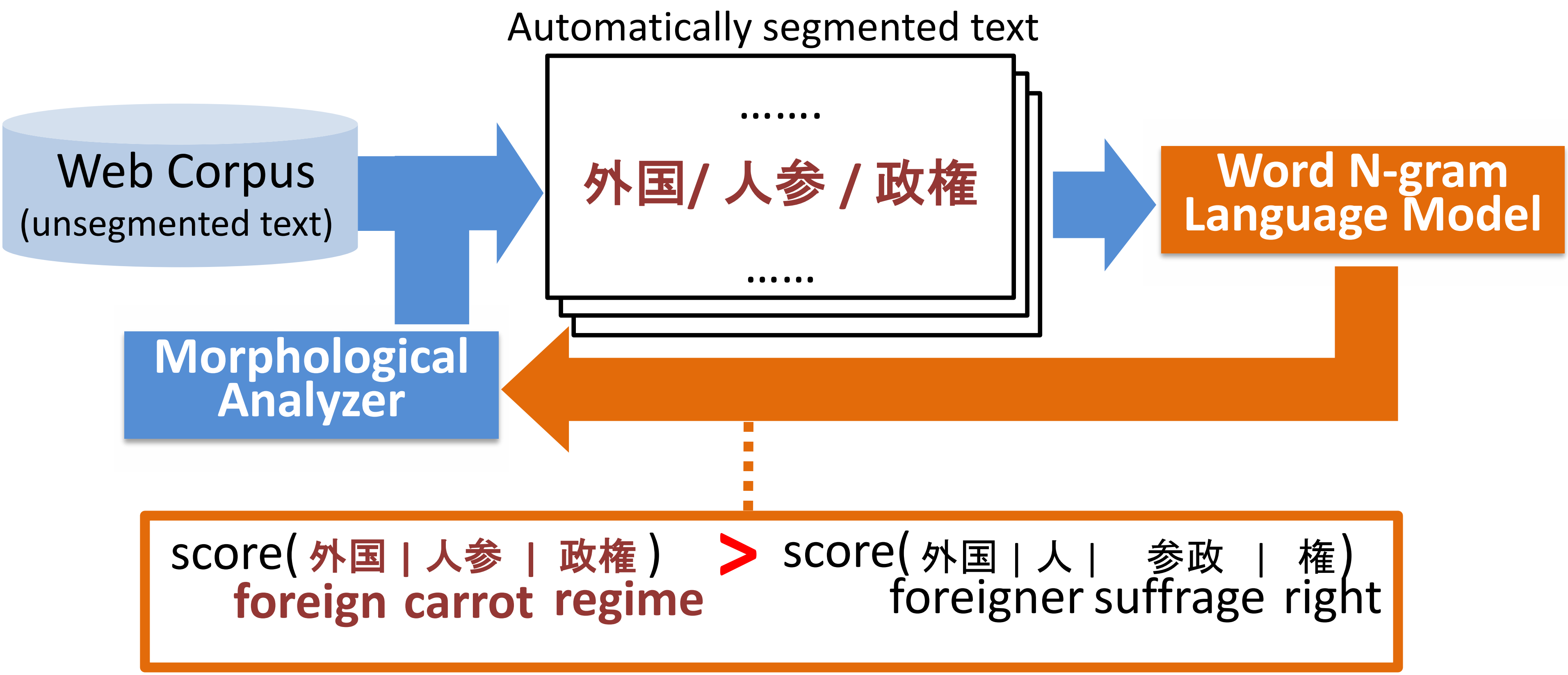
# Decoding

- Using the Viterbi Algorithm

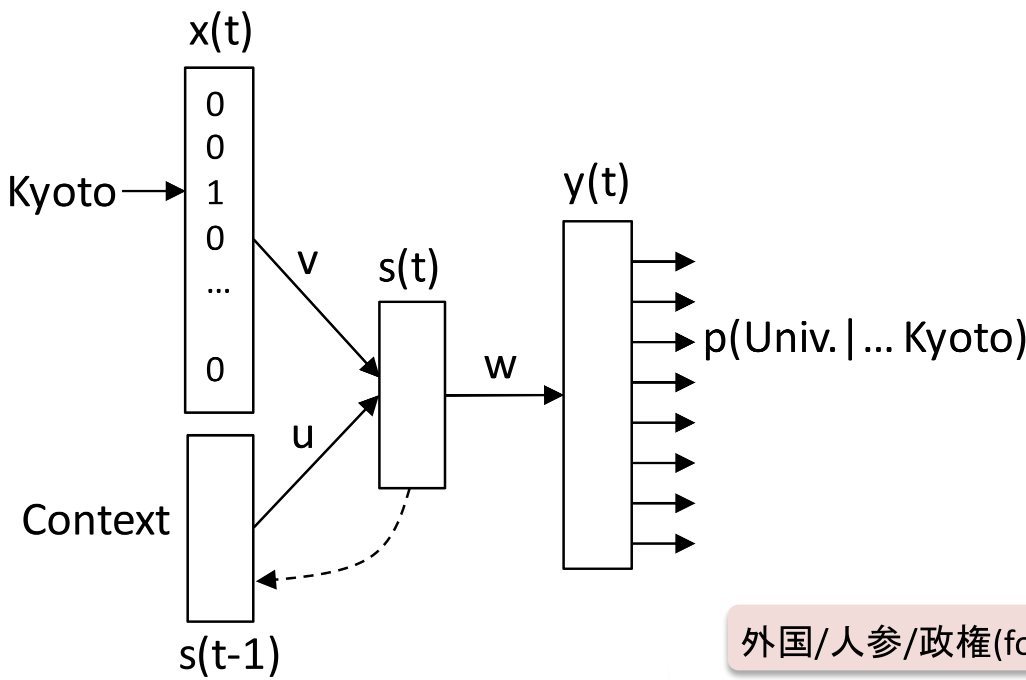


# Juman++: Chicken and Egg Problem

Can we use a language model as semantic knowledge?



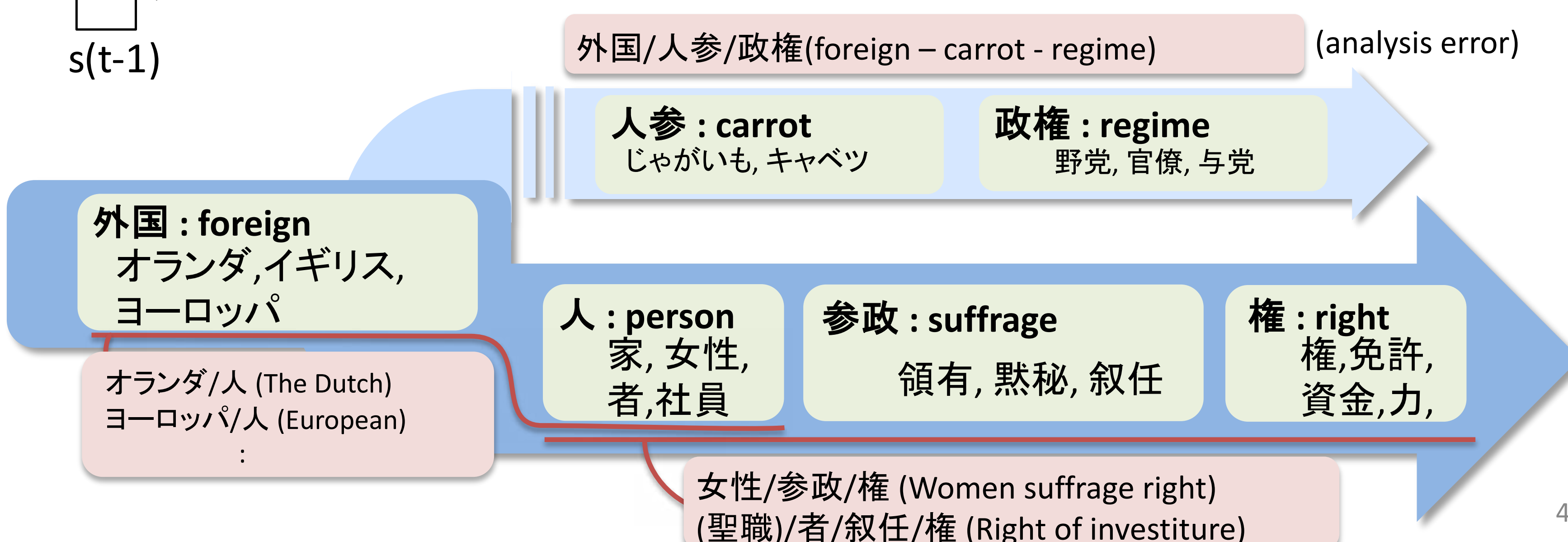
# Juman++: Use of RNN Language Model (RNNLM)




The model can calculate p(w|context) based on **semantically generalized** vector representation

Let’s use RNNLM for morphological analysis of unsegmented text!

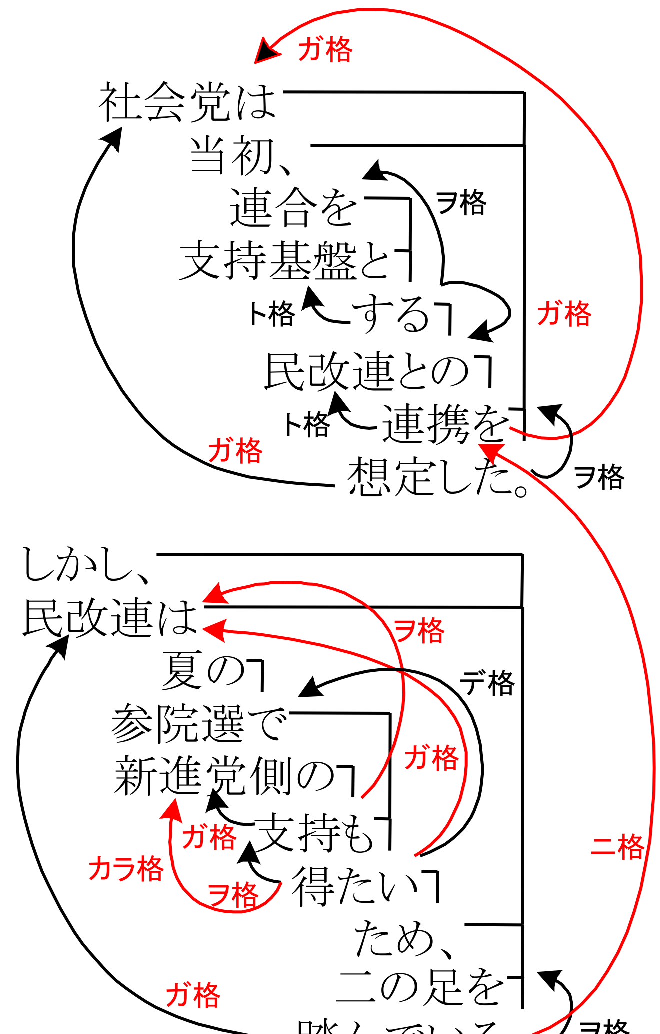
[Morita+ EMNLP2015]



# Kyoto University Text Corpus

[Kurohashi&Nagao 1998]

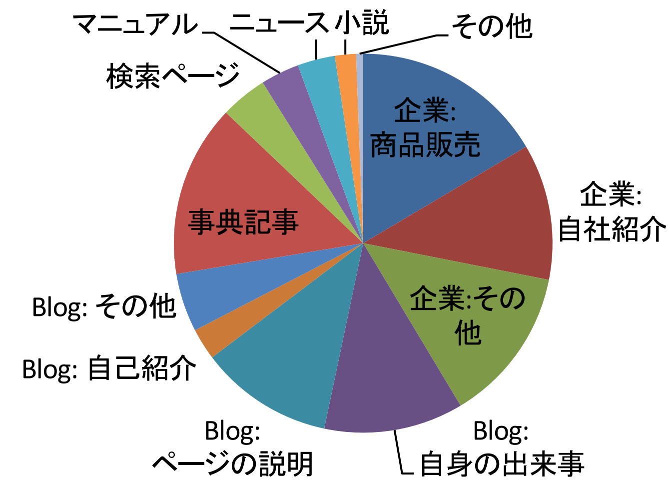
- 40K Mainichi newspaper articles annotated with syntactic information
  - Word segmentation
  - POS
  - Dependency
- 10K articles annotated with relation information
  - Predicate-argument structures
  - Relations between nouns
  - Anaphora and coreference



# KU Web Document Leads Corpus

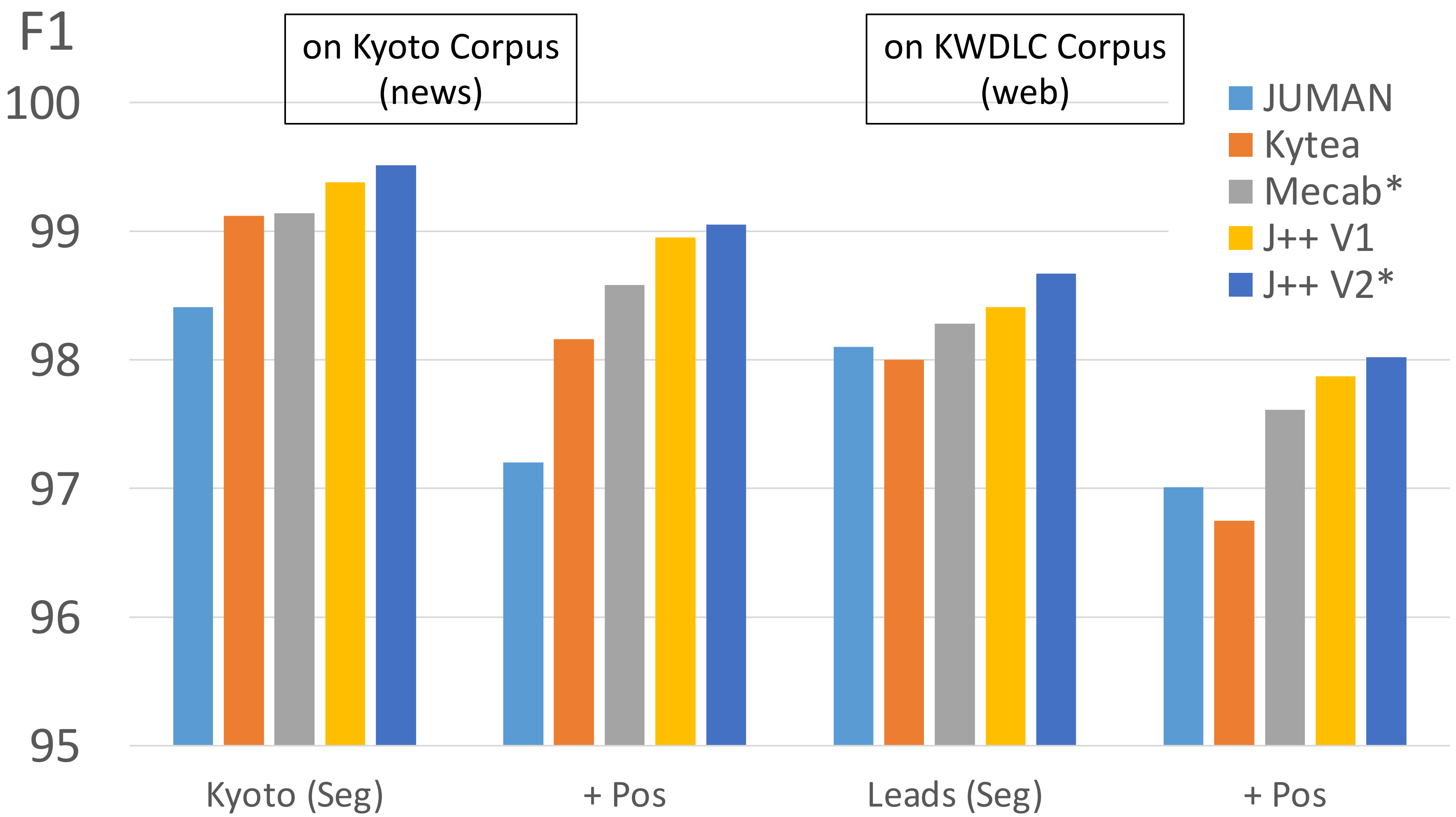
[Hangyo+ 2014]

- Lead 3 sentences of 5K web documents annotated with various linguistic information
  - Annotated by linguists
    - Word segmentation
    - POS
    - Dependency
    - Predicate-argument structures
    - Anaphora and coreference
  - Annotated by crowdworkers
    - Discourse relations




今回は様々な保険について([著者]ガ)([読者]ニ)説明しています。丁寧に([著者]ガ)([読者]ニ)(保険ヲ)解説したつもりですが、逆接([読者]ガ)分からない部分もあるかもしれません。原因・理由 疑問点はどんどん([読者]ガ) ([著者]ニ)コメントしてください。

# Accuracy




* = Optimized hyper-parameters on 10-fold cross-validation
Using the same Jumandic + concatenation of Kyoto/KWDLC corpora for training

# Remaining Problems

- Reading
  - 金(かね or きん)メダル
- Unknown words
  - 素晴しい ようつべ 待受
- 2-1 vs. 1-2 for 3-letter strings
  - 水/分子 ↔ 水分/子
- Ambiguities of postposition + verb
  - 部屋/に/はいる ↔ 部屋/に/は/いる
- Ambiguities of adverbs and verbs
  - あまり 極めて 改めて

# Assignment

Try to use one of POS taggers or morphological analyzers for any languages. Report analysis errors and think of an idea to solve them.

Deadline: May 26 (Thu) 23:59

※ You can write it in English or Japanese.

# Summary

- Sequence labeling tasks
  - POS tagging
  - Named entity recognition
  - Morphological analysis
- Methods
  - Viterbi decoding
  - BIO tagging scheme
  - CRF-based sequence labeling