EMNLP 参加報告

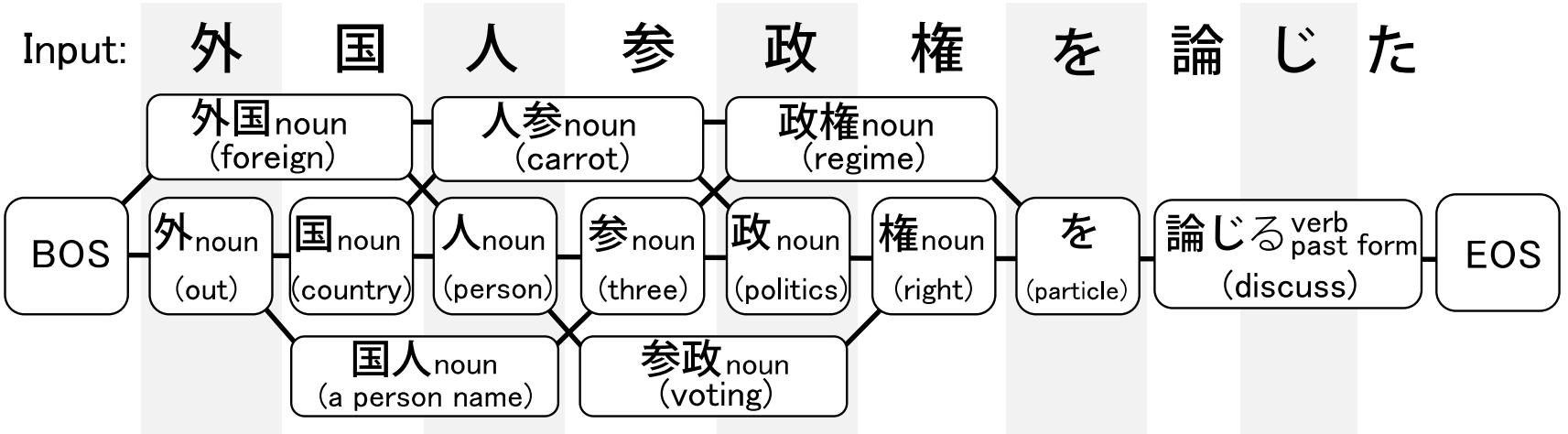
形態素解析における近年の傾向

Hajime Morita

Kyoto University

第224回自然言語処理研究会 2015/12/3

日本語の形態素解析



- 入力された文を，単語に分割し，品詞と活用を判定（原形を判定）する.
- 単語ベースのモデルが多い

中国語の場合

タグ	B	M	E	S	B	E
入力	田	雅	各	的	創	作
	Tian Yage			‘s	creations	

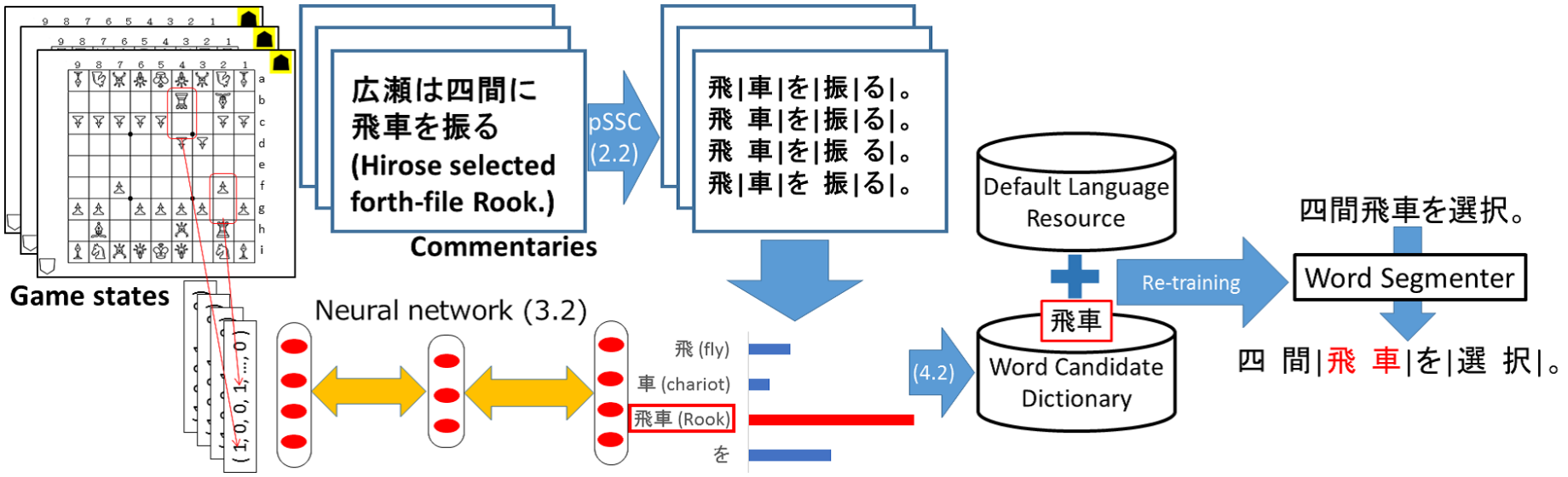
- 活用が無いため，単語分割と品詞の推定のみ.
- 文字ベースのモデルが多い
  - 各文字にタグを付加する
    単語の先頭:B 中間:M 末尾:E 一文字の語:S
  - 単語分割のみを扱うことも多い

Out-of-vocabulary (OOV)

- 日々生まれる新しい語，特定の地域，組織，コミュニティのみで使われる語は解析を失敗しやすい

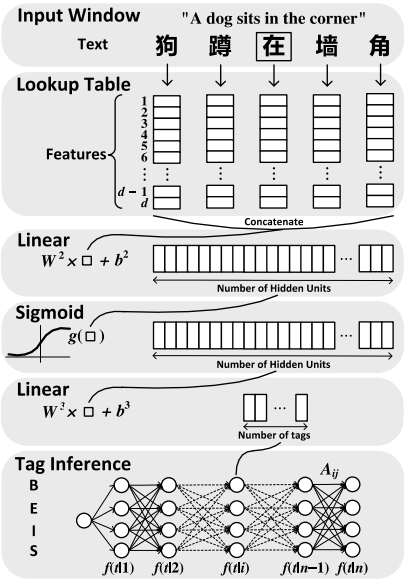
- 人手による辞書の拡張やコーパスの拡張によらない対策
  - Keyboard Logs as Natural Annotations for Word Segmentation
    - Fumihiko Takahashi, Shinsuke Mori
  - Can Symbol Grounding Improve Low-Level NLP? Word Segmentation as a Case Study
    - Hirotaka Kameko, Shinsuke Mori, and Yoshimasa Tsuruoka+

Can Symbol Grounding Improve Low-Level NLP? Word Segmentation as a Case Study

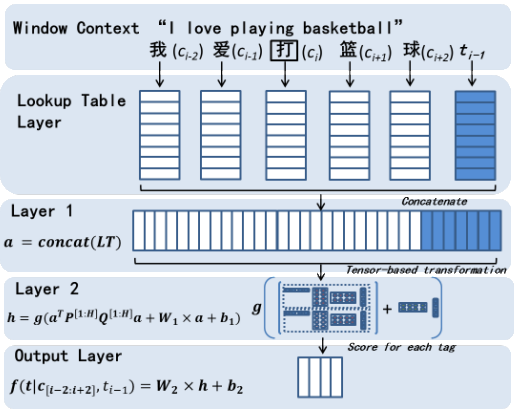


Neural Network (NN) を利用するモデルの発展

- NNを利用する手法が次々と発表されている

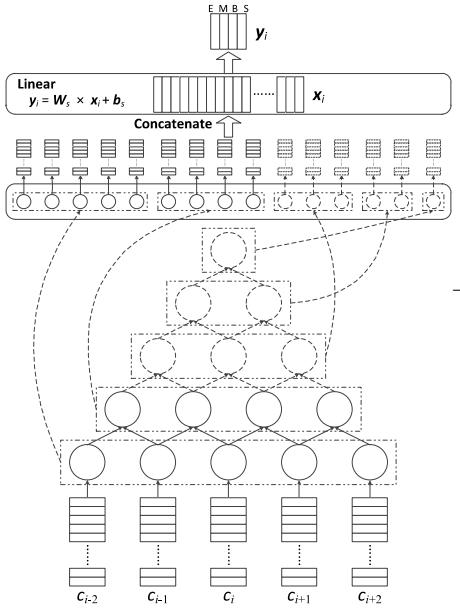


(Zheng et al., 2013)

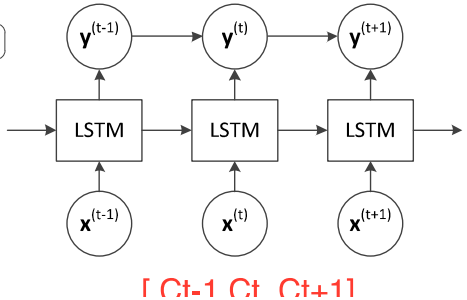


Tensor Neural Network

(Pei et al., 2014)



(Chen et al., 2015)



(Chen et al., 2015)

Long Short-Term Memory Neural Networks for Chinese Word Segmentation

- Long Short-Term Memory Neural Networks (LSTM) で単語分割
  - コンテクスト幅内におさまらない，長距離の関係を扱う
- 結果は，ほぼ全てのデータセットで State-of-the-art

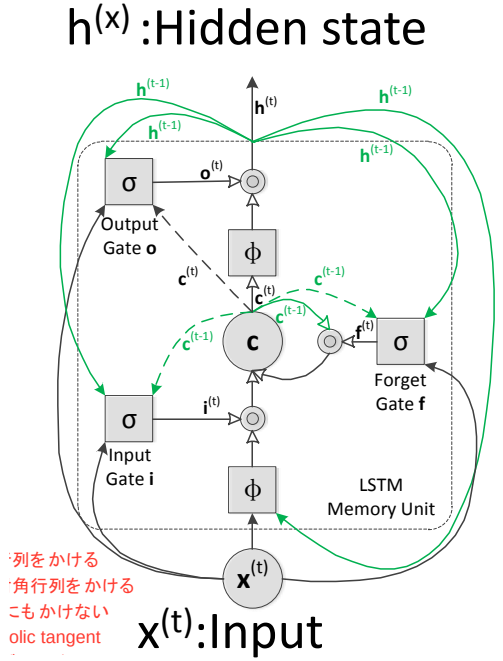
Introduction

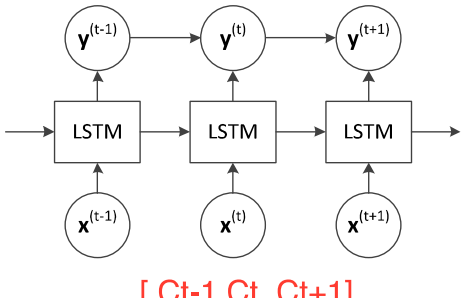
- 離れた単語が単語分割に影響する場合がまれにある
  - 中国語では…
    - /冬天/,/能/穿/多少/穿/多少/;
      /夏天/,/能/穿/多/少/穿/多/少/。
    - 冬はなるべくたくさん服を着て、
      夏は着る服が少なければ少ない方が良い。
  - 中国語に限った話ではない
    - 晩ごはん代が，タクシー乗って **/ある/か/ない/** かだ.
    - 駅まであるくか，タクシー乗って /あるか/ ない/ かだ.
  - ただし，実際に解決したとは言っているわけではない.
    - もともと，単語分割のレベルで解決するべき話ではない

- コンテキストサイズを単に広くすると，うまくいかない
- 人手で素性を作りこむのは負担が大きい

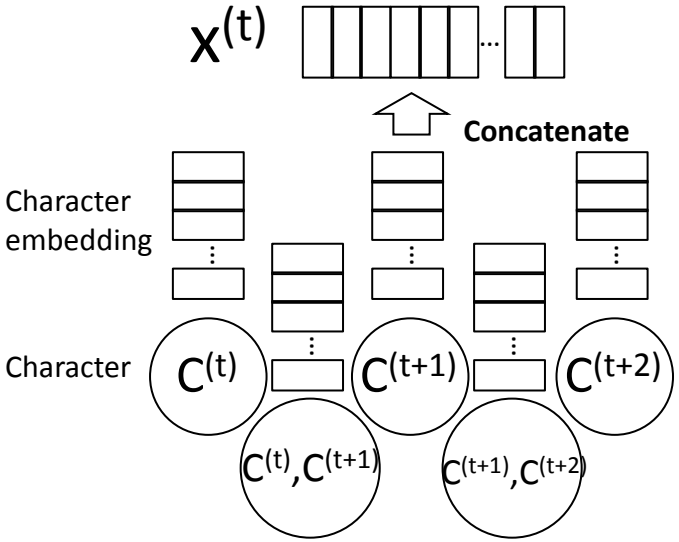
LSTM

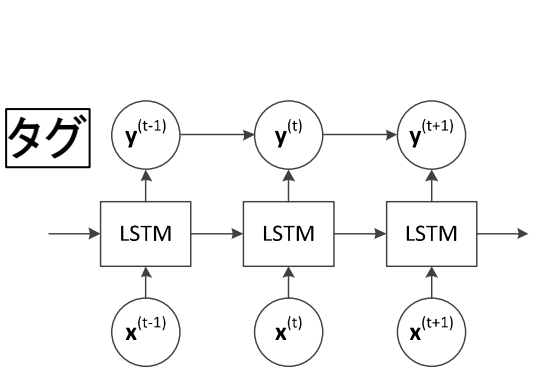
- Recurrent Neural Network と同様に，1つ前の状態を入力に受け取る
- 時間的に離れた依存関係を学習できることが強み





LSTM-1 LSTM-2

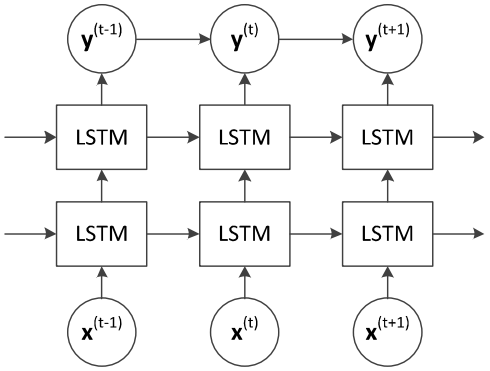




[Ct-1 Ct Ct+1]

(a) LSTM-1

PKU: 95.7 (F1)



(b) LSTM-2

PKU: 94.8 (F1)

- 時刻 t では 文字unigram と 文字bigram の embedding が入力
  - unigram embeddingの初期値は word2vec, bigram では2単語のembedding の平均
- $X^{(t)}$にそれ以前の文字の情報は含まれていない

Result

	Models	PKU	MSRA	CTB6
CRF ベース	(Tseng et al., 2005)	95.0	96.4	-
Averaged Perceptron	(Zhang and Clark, 2007)	95.1	97.2	-
CRF ベース	(Sun and Xu, 2011)	-	-	95.7
CRF ベース	(Zhang et al., 2013)	96.1	97.4	-
Tensor Neural Network	(Pei et al., 2014)*	95.2	97.2	-
Gated Recursive Neural Network	(Chen et al., 2015)*	96.4	**97.6**	95.8
	提案手法	**96.5**	97.4	**96.0**

(F1)

* は単語ベクトルの初期値を提案手法と同じ方法で初期化した場合の結果

- コンテクスト幅内の組合せ明示的に考える Pei+, Chen+ の手法と比較しても，PKU, CTB6 では少しずつ上回っている

Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model

Semantic knowledge of language

Problem

In unsegmented languages, language models are learned from automatically segmented texts and **inevitably contain errors**.

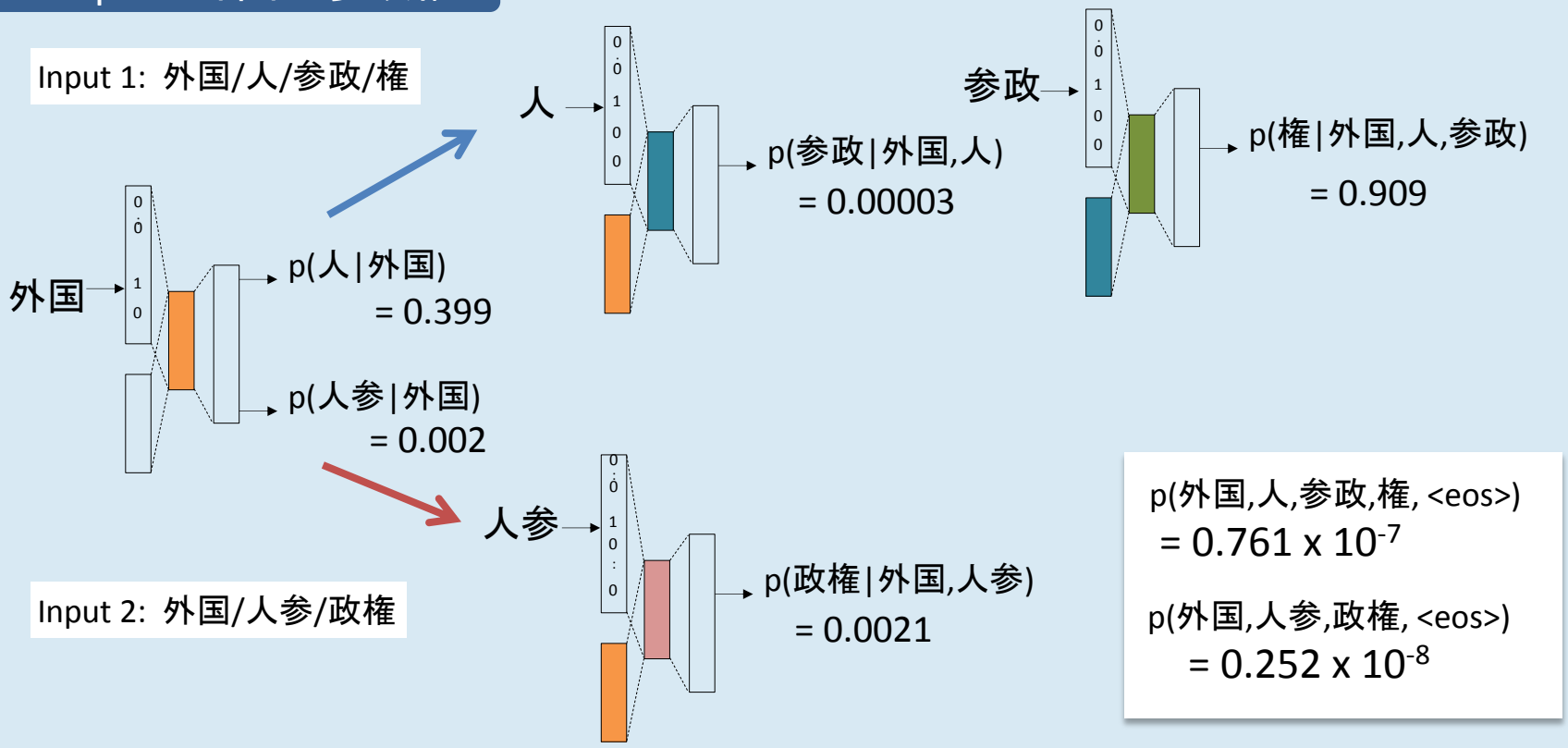
- A language model trained by incorrectly segmented "外国(foreign)/人参(carrot)/政権(regime)" just supports that incorrect and semantically unnatural word sequence.

Solution

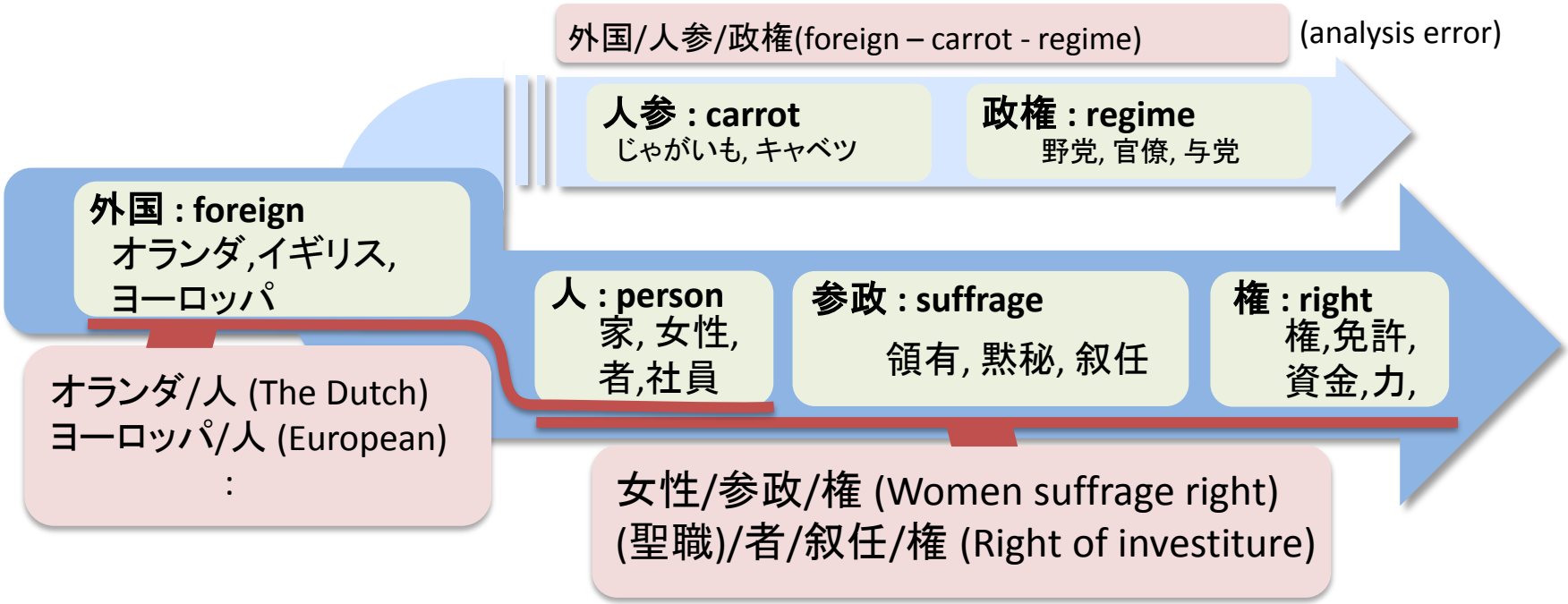
- **Semantically generalized language model**
  - The generalized model reduces influence of errors on automatically segmented corpus
- **Retraining using manually labeled corpus**
  - The retraining aims to cope with errors related to function word sequences

Recurrent Neural Network Language Model (RNNLM)

Example: “外国人参政権”

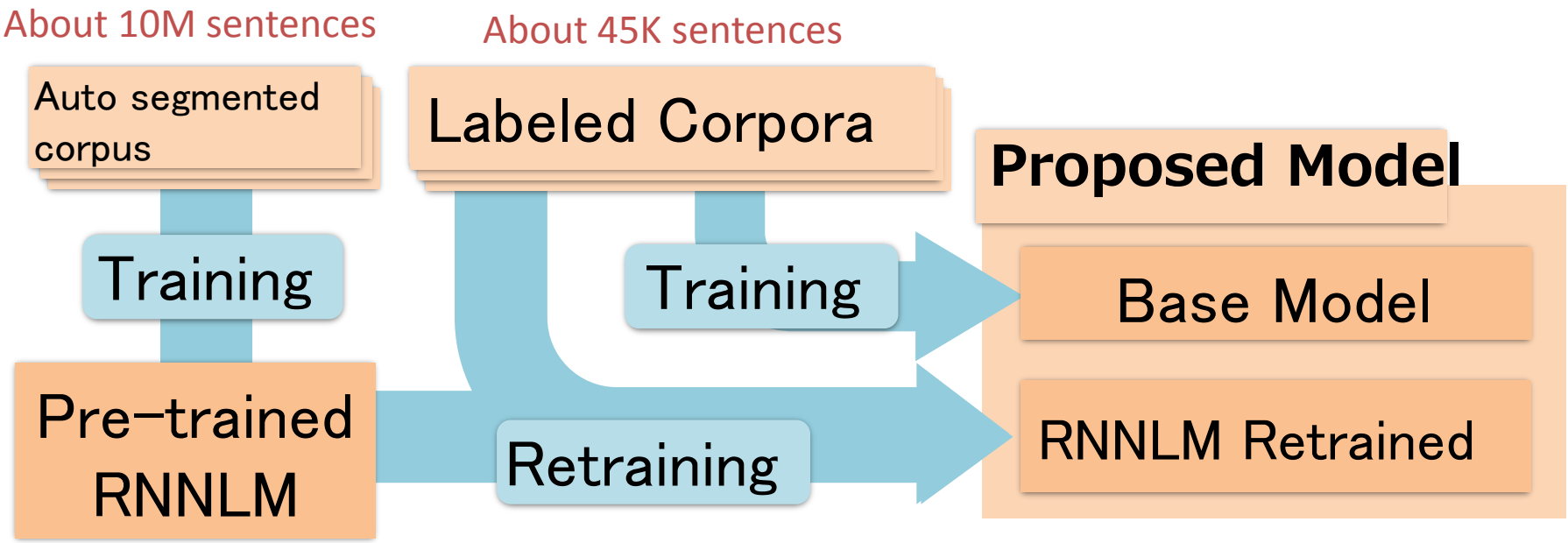


Semantically Generalized Language Model



On semantically generalized level, natural and correct semantic sequences are supported by many other similar word sequences.

Proposed Method



- We trained Base model for morphological analysis using manually labeled corpus.
- We train RNNLM using auto segmented web texts, and retrain the model using manually labeled corpus
  - The retraining aims to cope with errors related to function word sequences

Experiment and Results

Experiment

Baseline:

JUMAN, MeCab : Japanese popular morphological analyzer
a model using conventional language model (SRILM)

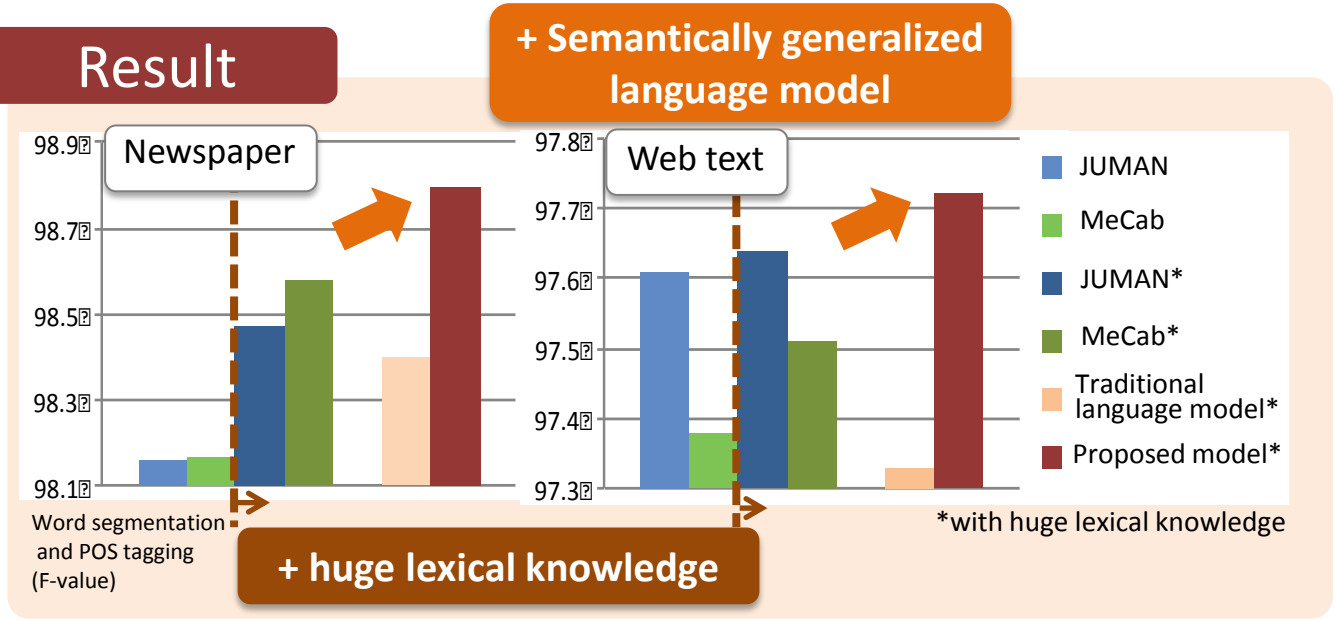
Dataset for training RNNLM:

10 million raw sentences
crawled from the web

Dataset for training base model and evaluation:

Kyoto University Text Corpus (Newspaper),
Kyoto University Web Document Leads Corpus (Web text)

Result



Example

解析結果の比較

JUMAN, Base Model

× 羽田/皓市/長/と/両/市/の/交流

Proposed model

○ 羽田/皓/市長/と/両/市/の/交流

JUMAN

× 自民党/の/森/喜/朗/幹事/長

× 東京/千/駄/ケ/谷/の/党/本部

× 歴史/の/流れ/となり/、/未来/の …

× 感想/やご/要望 …

Proposed model

○ 自民党/の/森/喜朗/幹事/長

○ 東京/千駄ケ谷/の/党/本部

○ 歴史/の/流れ/と/なり/、/未来/の …

○ 感想/や/ご/要望 …

Score of SRILM	**- 2.87**	>>	- 3.69
Score of RNNLM	- 5.61	<	**- 5.34**

まとめ

- OOV（未知語）の対策に，新しいリソースを利用する手法が現れた.

- （特に中国語の）形態素解析でニューラルネットベースの手法が従来のCRF ベースの手法を上回ってきている.