RNN 言語モデルを用いた日本語形態素解析

森田 一（京大）

形態素解析

- 単語の境界と品詞を推定するタスク

単語	私	は	新しい	本	も	買った
品詞	名詞	助詞	形容詞	名詞	助詞	動詞

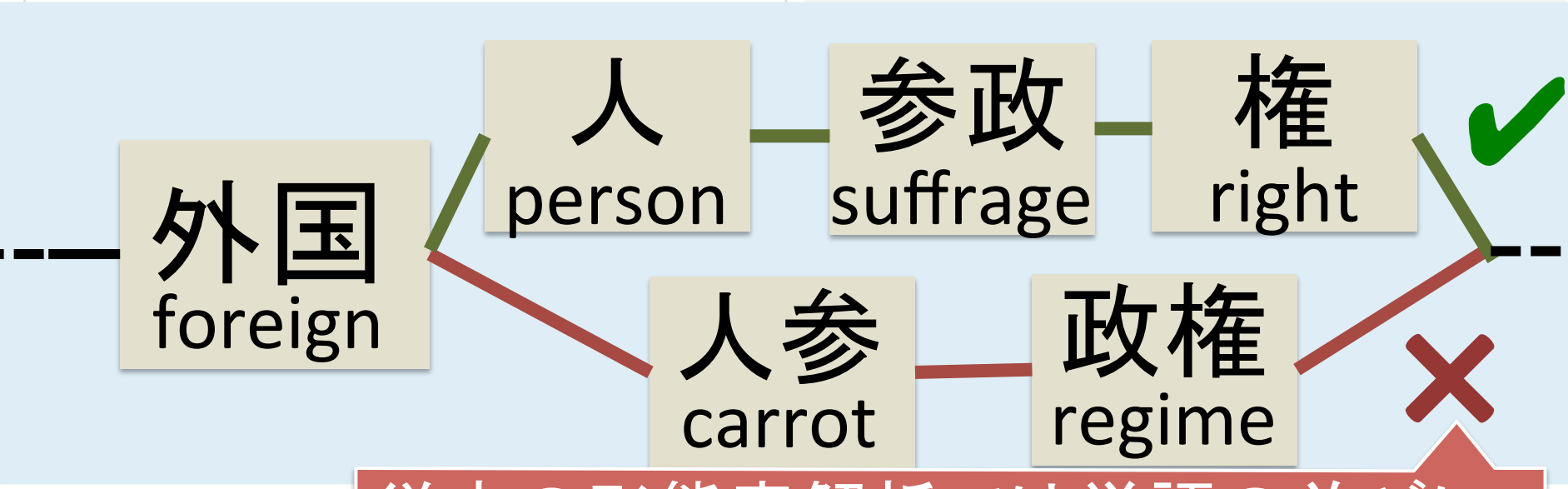
背景

English | Spanish | Japanese | Detect language

外国人参政権

foreigner suffrage right

Foreign carrot regime



従来の形態素解析では単語の並びに対する意味的自然さを考慮できない

概要

Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model [Morita+; 2015]

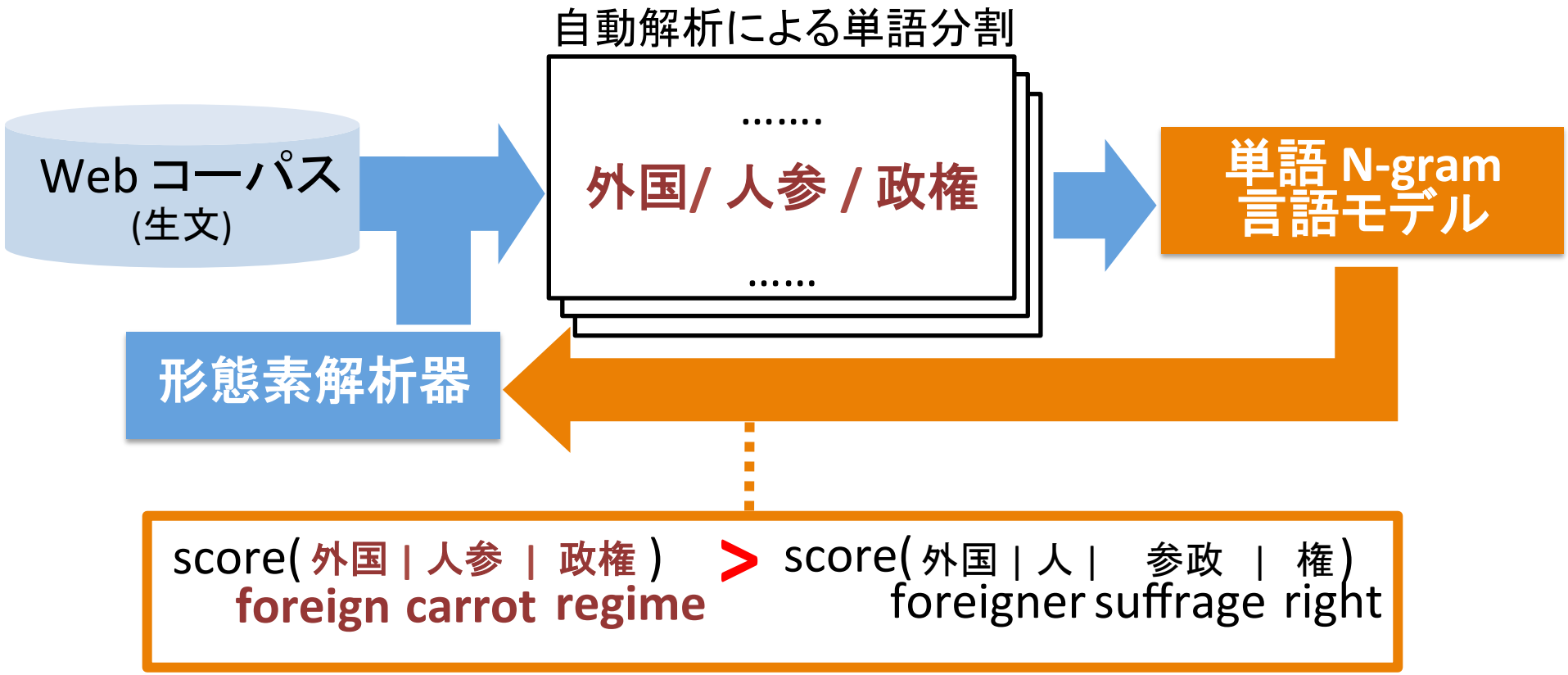
1. 意味的に汎化された言語モデル(RNNLM)の利用
2. Wikipedia, Wiktionary, 大規模Webコーパス等からの大規模語彙獲得

実用化に向けた分析

- 形態素解析の誤りの徹底分析
- 部分アノテーションによる改善枠組みの実現

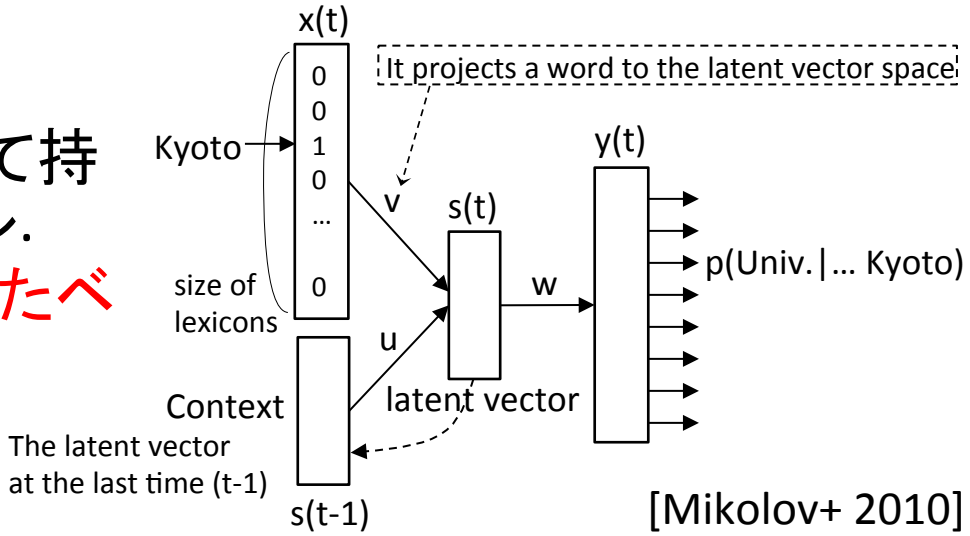
Chicken and Egg Problem

言語モデルは形態素解析に利用できない？



Recurrent Neural Network Language Model (RNNLM)

- コンテクストを隠れ層として持つNNベースの言語モデル.
- 単語を意味的に汎化されたベクトルとして扱う



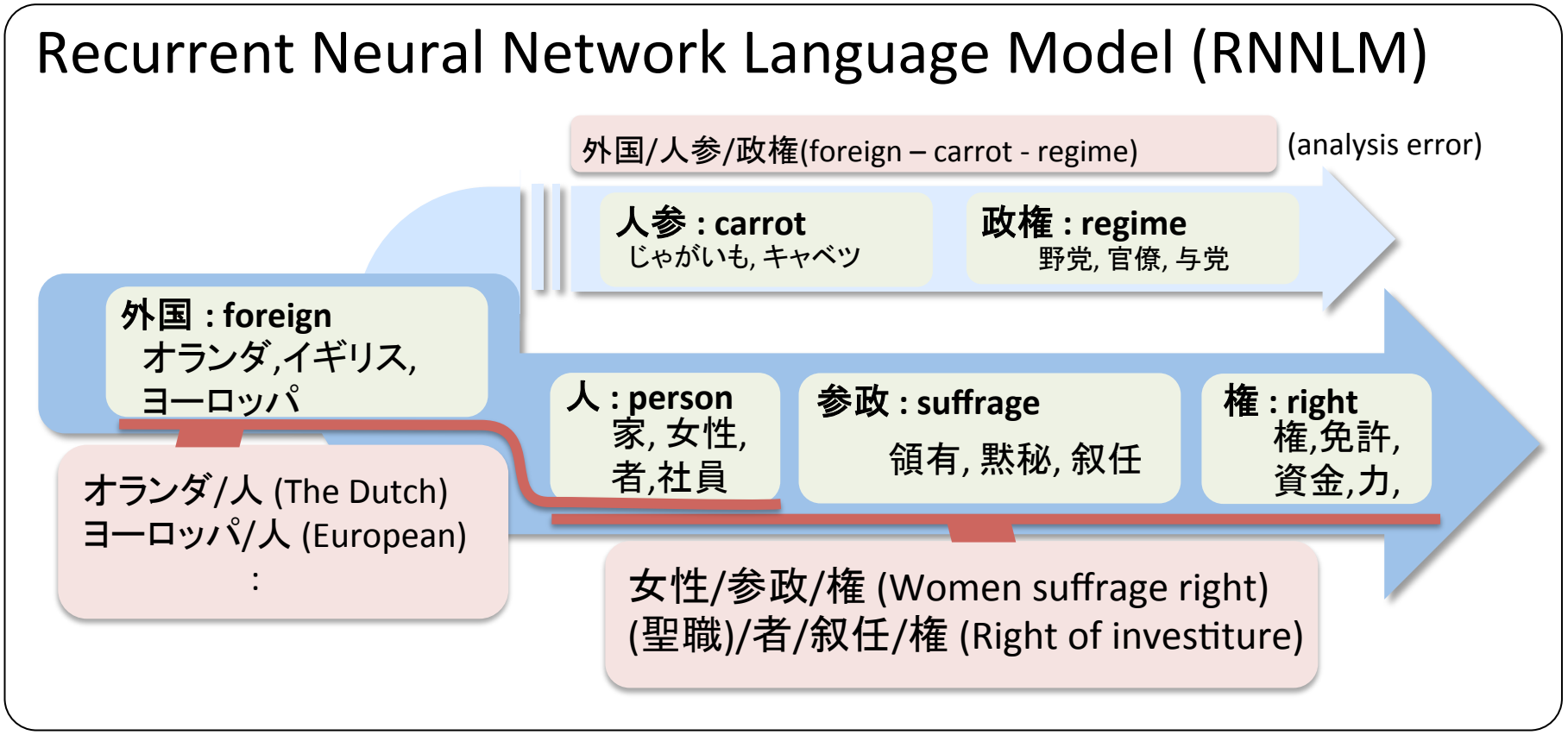
[Mikolov+ 2010]

$$s_j(t) = f\left(\sum_i x_i(t)v_{ji} + \sum_l s_l(t-1)u_{jl}\right), \qquad y_k(t) = g\left(\sum_j s_j(t)w_{kj}\right)$$

$$f(z) = \frac{1}{1+e^{-z}}, g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}, \qquad v_{ji}, u_{jl}, w_{kj} \text{ are elements of V, U, W respectively.}$$

[Morita+,2015]

1.意味的に汎化された言語モデルの利用



- RNNLMでは単語を意味的に汎化したベクトルとして扱う
- 自然な意味・単語の系列には多くの似た単語列が存在
  - 品詞等の素性によるスコアと合わせて解析に利用

[Morita+,2015]

2.大規模語彙知識

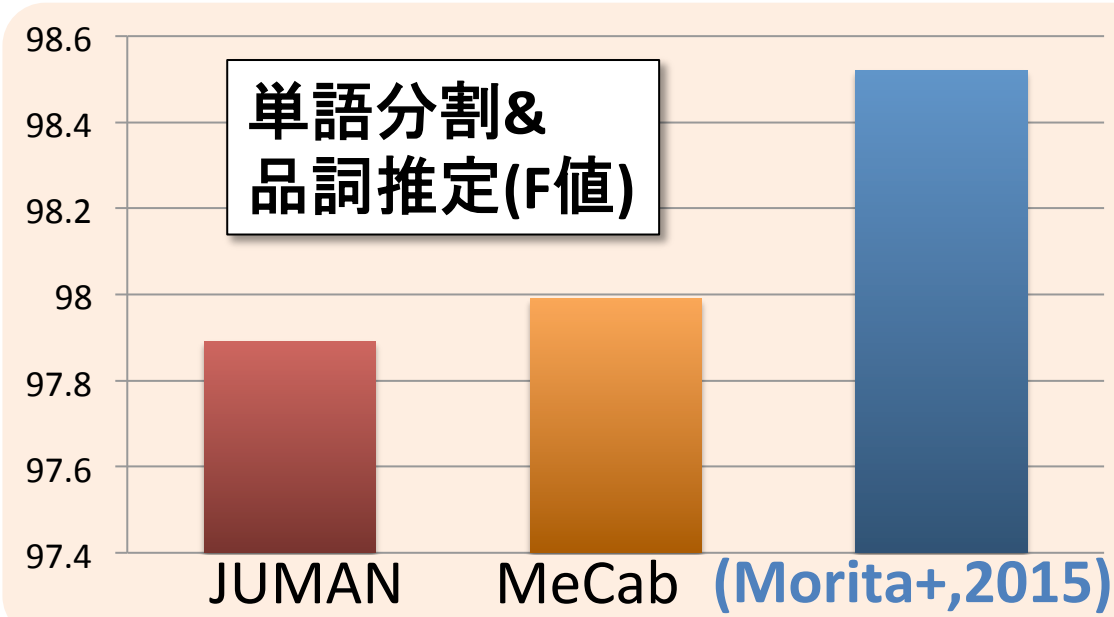
- Wikipedia, Wiktionary, Web テキストコーパスから，語彙知識を獲得.

	語彙数	Examples
基本辞書	3万語	走る，行く，明日
Wikipedia	85万語	橋下，お茶の水， 東プレ，アベノミクス
Wiktionary	2千語	インセンティヴ，糾す
Web text	1万語	逆進税，政独委
合計	90万語	

[Morita+,2015]

JUMAN,MeCabとの比較

- 利用データ
  - 京都大学テキストコーパス(news)，京都大学ウェブ文書リート゛コーパス(web) を合わせて訓練・評価用データとして利用
    - 訓練データ 49,774 文
    - エラー分析用データ 995 文
    - 精度評価用データ 2,983 文



✕ JUMAN

感想|やご|要望

✔ Proposed

感想 |や|ご|要望

形態素解析に残る問題点の分析

- 許容できる誤り:　2種類
- 許容できない誤り: 3種類

許容できる誤り

- **基準の違い**
  - コーパス・アノテーションと複合語の分割や品詞が違うが,解釈の誤りとはいえないもの.
    - | x北極/点 ← o北極点 |
- **意味的曖昧性に起因する誤り**
  - 文法的に問題のない単語列に分割されていて,形態素解析では区別しづらい意味的な曖昧性がある場合.
    - 単身赴任のようと | よく (x形容詞 ← o副詞) | 言われる
    - さかのぼって | みる (x接尾辞 ← o動詞) |

（凡例：x解析結果 ←oアノテーション）

許容できない誤り

- **未知語による誤り**
  - 薄日が | x射/して ← o射して |
- **複合語の分割誤り**
  - | x新名/人 ← o新/名人 |
- **その他の誤り**
  - | xおす/す/めな ← oお/すすめ/な | (他の解析器の例)
  - 増加の | x一途で(形容詞) ← o一途(名詞)/で (助詞) |

(凡例: x解析結果 ← oアノテーション)

誤りの分類内訳

- 連続した形態素の解析誤りを 1 箇所の誤りとしてカウント

	分析データ (995 文)	1-best		5-best
		JUMAN	[Morita+, 2015]	[Morita+, 2015]
許容できる誤り	基準の違い	203	139	-
	意味的曖昧性	42	24	7
許容できない誤り	未知語による誤り	12	8	8
	複合語の分割誤り	27	3	2
	その他の誤り	28	9	2

誤りの大部分は，許容できる誤り

残る許容できない誤りは見つけ次第修正したい

部分アノテーションを用いた学習

- 与えた単語境界に従って解析を行う機能を実装
- エラー分析用データ内の誤りについて単語境界を部分アノテーション
  - 1-best での誤り6ヶ所 (複合語分割:3, その他:3)
  - アノテートを利用した解析の結果を学習データに追加し，再度学習

解析誤り　この 木 の　実は　、発芽 し ない
(副詞)

部分アノテーション　この木の ¥t 実 ¥t は ¥t 、発芽しない

部分アノテーションの効果

	エラー分析用データ (6文を部分アノテーション)	精度評価用データ
	995 文	2,983文
JUMAN	97.89	97.91
MeCab	97.99	98.00
[Morita+,2015]	98.52	98.42
再学習(+部分アノテーション)	**98.53**	**98.41**

部分アノテーションを行っていないデータに対しても，大きな副作用はなかった

誤りの分類内訳
（部分アノテーション追加後）

	分析用データ (995 文)	1-best JUMAN	1-best [Morita+, 2015]	1-best +部分アノテーション	5-best [Morita+,2015]	5-best +部分アノテーション
許容できる誤り	基準の違い	203	139	157	-	-
	意味的曖昧性	42	24	19	7	6
許容できない誤り	未知語による誤り	12	8	**8**	8	**8**
	複合語の分割誤り	27	**3**	**0**	2	**0**
	その他の誤り	28	**9**	**6**	2	**2**

部分アノテーションした箇所は全て正しく解析された

残る問題はほぼ未知語の問題

まとめ

- 部分アノテーションを用いた学習
- 形態素解析に残る問題点を分析
  - Web の多様なテキストが含まれるコーパスで,
    5-best 解で許容できない誤りは 1,000文中
    10 箇所程度 (誤り率: 0.08 %)
  - 残るほとんどの解析誤りは未知語に起因