

『JUMAN++』の 大規模語彙獲得へ向けた 取り組み

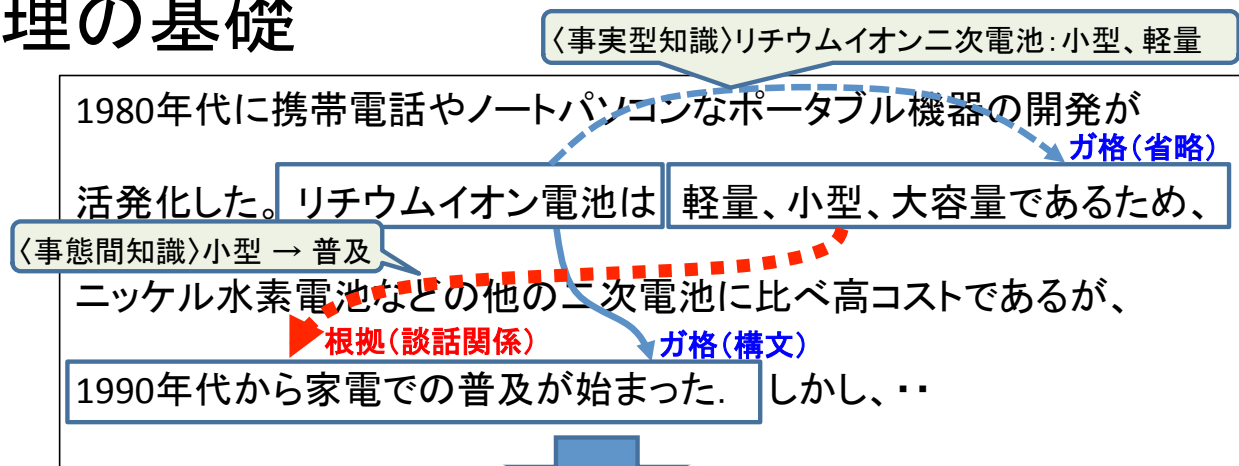
京都大学 森田 一

形態素解析とは

- 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

- 後段の言語処理の基礎



CREST
知識に基づく構造的言語処理
(黒橋, 京都大学グループ)

新たに獲得される知識

リチウムイオン電池は
軽量、小型、大容量

→

リチウムイオン電池は1990年代
から家電での普及が始まった

未知語

- 辞書に登録されていない語
 - 日本語の形態素解析は辞書に大きく依存しているため、未知語が存在する文では解析を誤りやすい。
- 形態素解析で残るエラーの40%が未知語によるもの(1-best)

例文:

「綺麗にへこんでるお腹になりたい」

きれいに	形容詞	"代表表記:綺麗だ/きれいだ"
へこ	未定義語	
ん	未定義語	
でる	動詞	"代表表記:出る/でる"
お腹	名詞	"代表表記:お腹/おなか"
に	助詞	
なり	動詞	"代表表記:成る/なる"
たい	接尾辞	"代表表記:たい/たい"

未知語の撲滅に向けて

- 大きければ大きいほどよいが、人手では語彙拡張に限界がある

自動的な語彙獲得

- しかし、単純に語彙を増やすと解析に副作用が
 - 例: 「ことは」(人名)
写真部だという|ことは|以前... (JUMANによる解析)

JUMAN++ による解析

目次

- JUMAN++
- 語彙獲得フロー
 - 語彙獲得の方針
 - Wikipedia からの語彙獲得
 - Wiktionary からの語彙獲得
 - Webテキストからの語彙獲得
- まとめ

JUMAN++

背景

English Spanish Japanese Detect language ▼

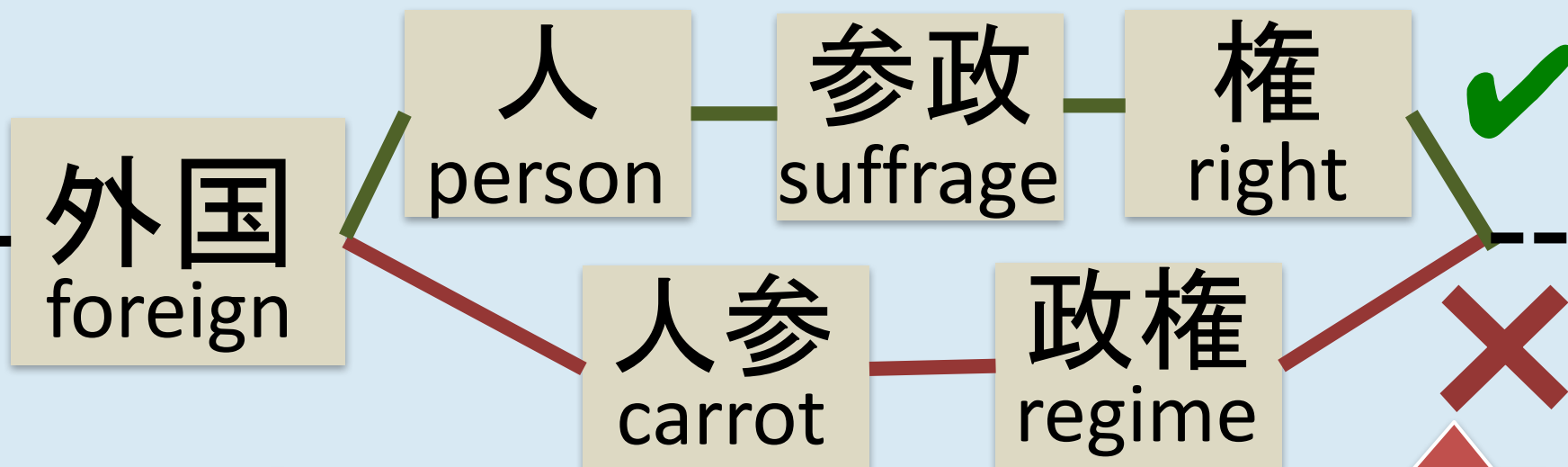
外国人参政権



foreigner suffrage right



Foreign carrot regime

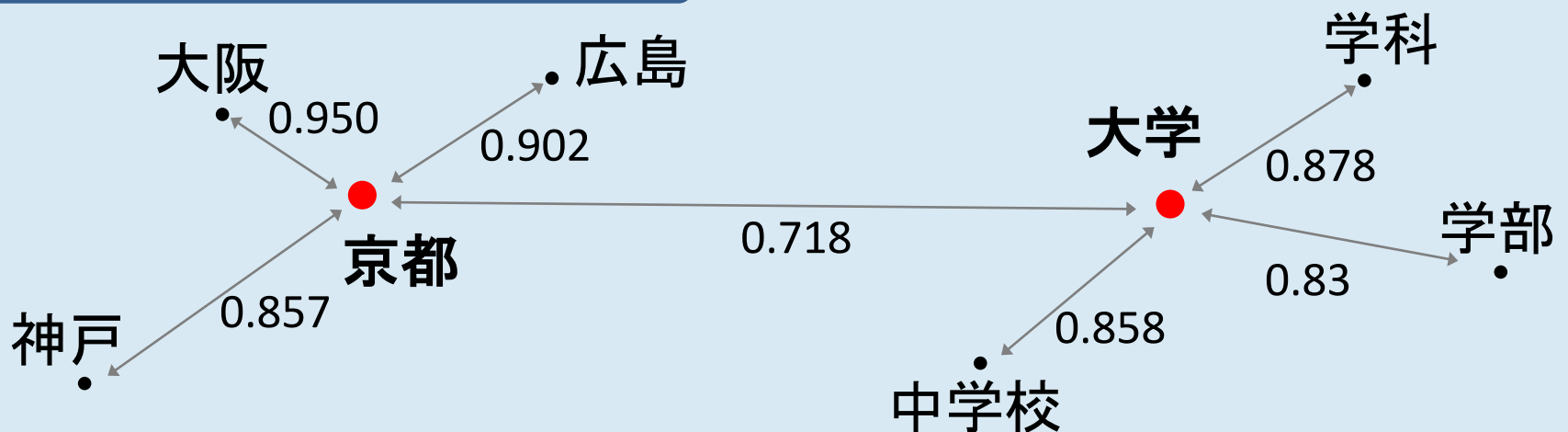


従来の形態素解析では単語の並びに
対する意味的自然さを考慮できない

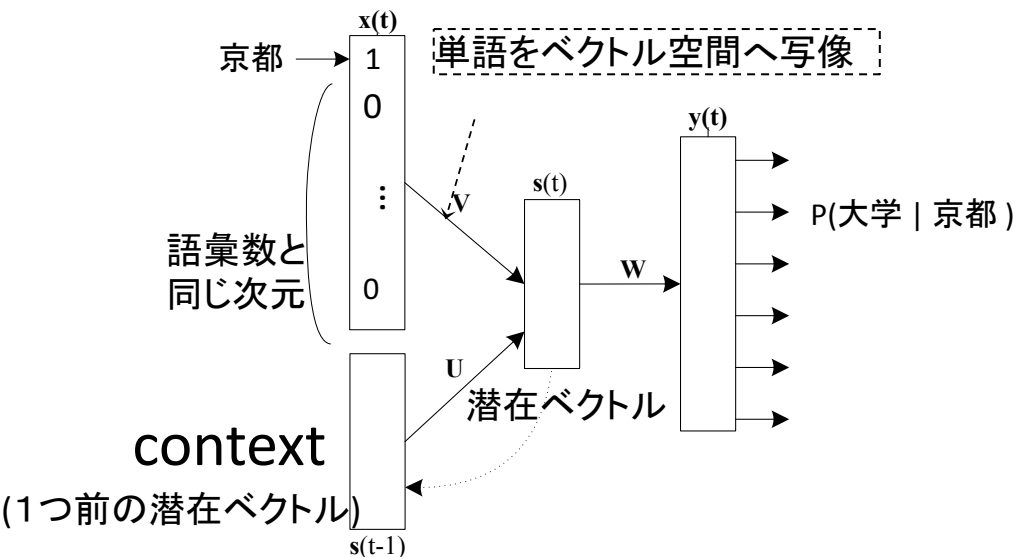
単語ベクトル

- 近年の言語処理では、ベクトルで単語を表現することがよく行われる
- 似ている単語, 似ていない単語をベクトルの類似度で表現することができる

単語ベクトル間の類似度

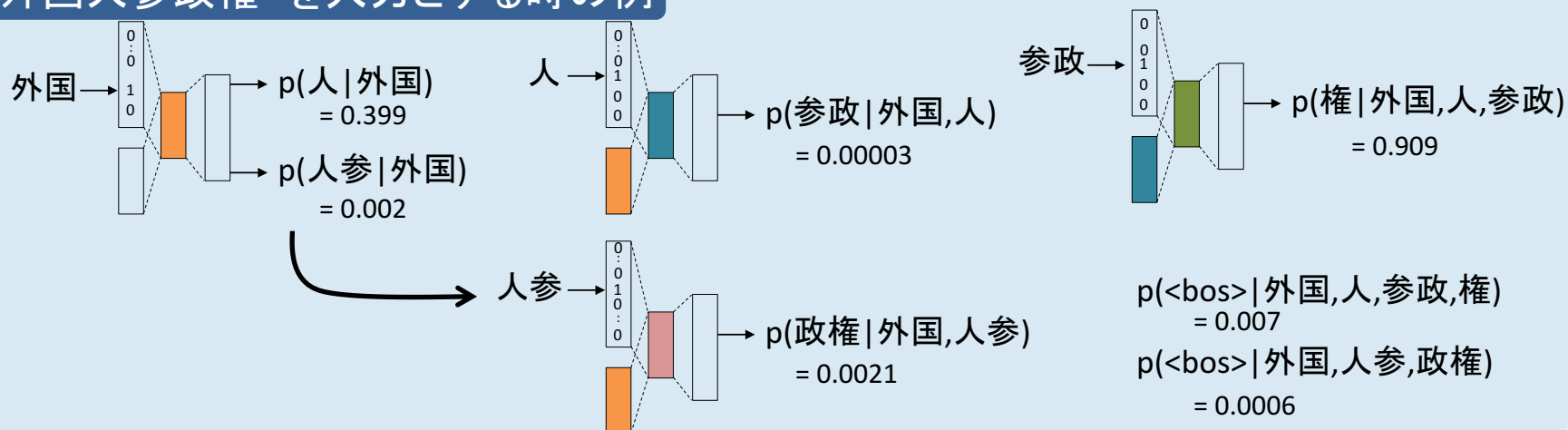


RNNLM



- コンテキ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

“外国人参政権”を入力とする時の例



JUMAN++による精度改善

[Morita+ EMNLP2015]

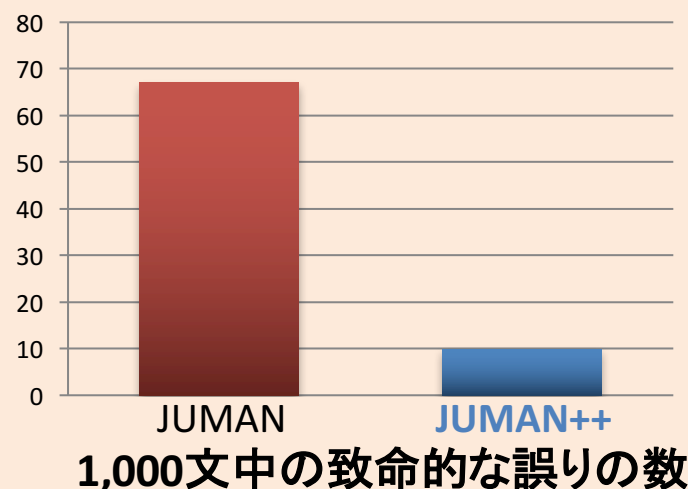
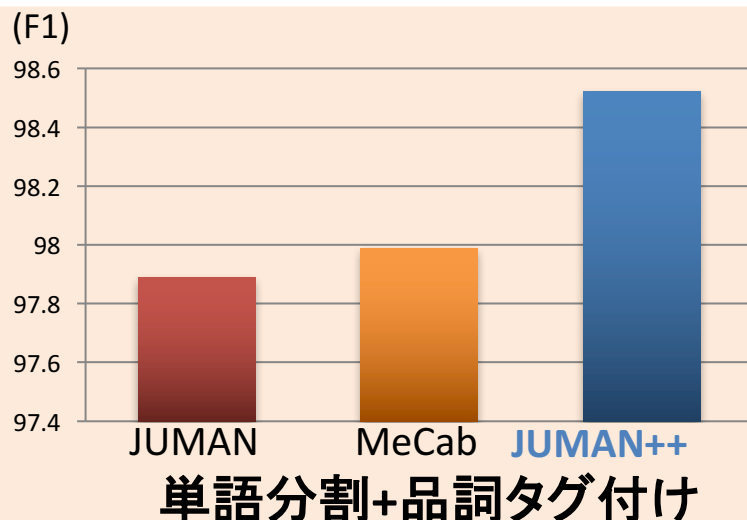
実験設定

RNNLMの訓練に用いたデータ:

Web テキスト 1000万文

基本モデルの学習と評価に用いたデータ

京都大学テキストコーパス(news),
京都大学ウェブ文書リードコーパス(web)



✗ JUMAN

感想 | やこ | 要望

✓ JUMAN++

感想 | や | こ | 要望

副作用の軽減例

- JUMAN++による解析で, 副作用を軽減
 - 写真部だという|ことは|以前... (JUMAN)
 - 写真部だという|こと|は|以前... (JUMAN++)
- 人名として解析するほうが適切な文脈では, 人名として解析
 - |ことは|は以前写真部にいた (JUMAN++)

誤り例を集めています

部分アノテーションの共有

JUMAN++ が解析を誤った部分文字列に対する、部分アノテーションを収集しています。部分アノテーションを作成した際はページ下部のコメント欄で共有をお願いします。収集した部分アノテーションは、JUMAN++ 用の学習コーパスとして公開を予定しています。

部分アノテーションのフォーマット

明示的に形態素境界として指定したい箇所にタブ (\t) を挿入することで、挿入した箇所が必ず形態素境界として解析されます。また、タブに囲まれた範囲は必ず1形態素として解析されます。タブの入力が困難な場合、¥tもしくは\tと入力していただいても構いません。

例：（「実は」の解析を誤って1形態素としてしまう場合）
この\t実\tは発芽しない

部分アノテーションを共有する

- 注意事項:
 - ここで共有した部分アノテーションは Apache License, Version 2.0 で公開されます。
 - 個人情報を含む部分アノテーションの共有はご遠慮ください。
 - 他者の権利を侵害する可能性がある場合は、解析に影響がない範囲で共有された部分アノテーションの編集を行うことがあります。
 - 多量の部分アノテーションを共有する場合など当ページでの共有が困難な際は、下記連絡先までご相談ください。

送信

2017/02/23 13:20 品質 がめっちゃ いい んですね

収集された誤りの例

- まっくろくろすけでておいで

収集された誤りの例

- まっくろくろすけでておいで

まっ	接頭辞	"代表表記:真っ/まっ"
----	-----	--------------

くろく	形容詞	"代表表記:黒い/くろい"
-----	-----	---------------

ろすけ	名詞	"代表表記:露助/ろすけ"
-----	----	---------------

で	助詞	
---	----	--

ておい	名詞	"代表表記:手負い/ておい"
-----	----	----------------

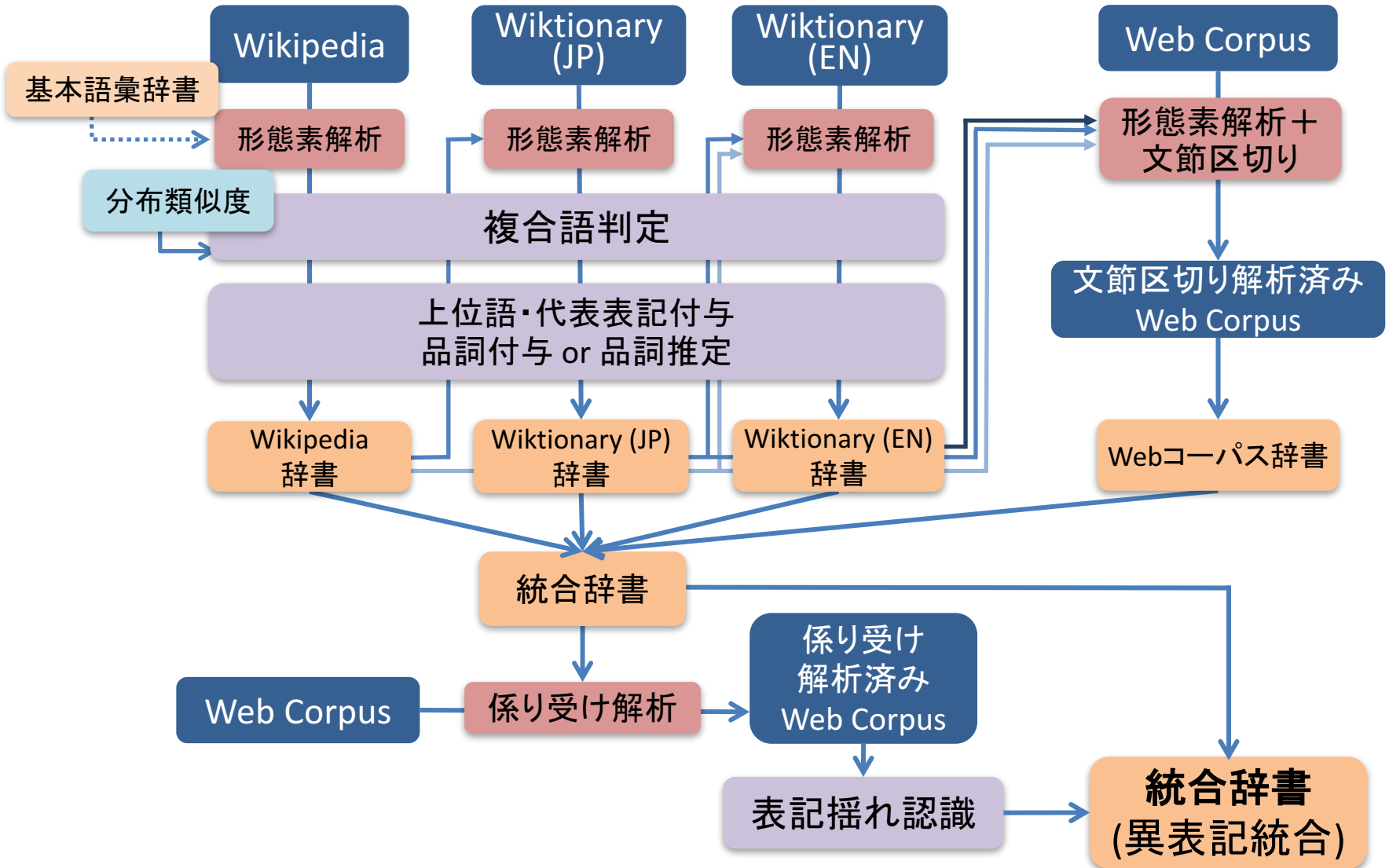
で	助詞	
---	----	--

語彙獲得フロー

語彙獲得の方針

- 基本的な語彙は人手で整備(基本語彙)
 - 語彙数を絞る代わりに, 語のドメインや反義語などの情報をリッチに付与する
- 基本語彙に含まれない語は自動的に獲得
 - 単語の候補(Wikipediaの記事タイトルなど)のうち, 複合語でないものを自動判定し辞書へ

JUMAN++辞書獲得フロ－



未獲得の語彙が見つければ

1. まずは自動獲得の見直し
 - Wikipedia, Wiktionary から語彙を獲得するルールや基準を見直し
2. 基本的な語彙であれば, 基本語彙として人手登録することも検討
3. 将来獲得したい候補として貯めておく
例) にがうるか

Wikipedia からの語彙獲得

- リソース
 - 約100万記事
 - 記事タイトル, リダイレクト, カテゴリ, テンプレート, 箇条書き等を利用
- 獲得する語彙
 - 幅広く名詞を獲得
 - 普通名詞: アベノミクス, GPGPU, ...
 - 組織名: Dentsu, ファーウェイ, ...
 - 人名: 山極, オバマ ...
 - 地名: 豊洲, 伊勢 ...

記事タイトルからの語彙獲得

記事タイトル

JUMAN

上位語

定義文

JUMANは、日本語の形態素解析システム。京都大学黒橋・河原研究室で開発されている。

人手で整備した辞書に基づいており、ChaSenの元となったシステム。

外部リンク [\[編集\]](#)

- 京都大学 黒橋・河原研究室
- 日本語形態素解析システム JUMAN (ウェブサイト)

 この項目は、コンピュータに関連した書きかけの項目です。この項目を加筆・訂正などして下さる協力者を求めています (PJ:コンピュータ/P:コンピュータ)。

カテゴリ: 自然言語処理

- 記事タイトル, リダイレクトから語彙を獲得
 - 記事中の定義文から上位語を抽出し, 細分類(人名,地名等)を判定
 - 複合語を除外して辞書に登録する

複合語の除外

- 単語候補を形態素解析した結果を利用
- 解析結果が誤っていると思われる例を辞書に追加
 - 辞書に無く、分割もできないカタカナ語, アルファベット
例) ゴスペル, ThinkPad
 - 一文字の語の並びとして解析される場合
例) 爽 | 健 | 美 | 茶, み | ん | ぱ | く
 - カタカナ語の誤った分割
 - カタカナ語の分割は難しく、単語候補が分割できても、あやまった分割であることが多い
 - 単語候補とその主辞の間で、WEBテキスト上に出現するコンテキストを比較し、類似していなければ単語として獲得する
例) フット+サル (≠ サル)

記事のカテゴリや Infoboxからの獲得

記事タイトル

坂本龍馬

坂本 龍馬（さかもと りょうま、[天保6年11月15日](#)（[新暦・1836年1月3日](#)） - [慶応3年11月15日](#)（[新暦・1867年12月10日](#)））は、江戸時代末期の志士、[土佐藩郷土](#)。

定義文

[諱](#)は直陰（な **上位語** ちに直柔（なおなり）。[通称](#)は龍馬^{[\[注 1\]](#)}。他に才谷 梅太郎（さ

：

坂本 龍馬



Infobox

カテゴリ: [坂本龍馬](#) | [土佐勤王党の人物](#) | [海援隊の人物](#) | [日本の貿易立国論者](#)
| [江戸時代の商人](#) | [暗殺された人物](#) | [革命家](#) | [剣客](#) | [北辰一刀流剣術](#) | [1836年生](#)
| [1867年没](#)

カテゴリ

- カテゴリ, Infobox を利用して人名や組織名を判定
 - 人名は姓と名を分割して, それぞれを登録

一覧ページからの獲得

- 一覧ページには、個別の記事を持たない語であっても記載されている場合がある
- 一覧から語彙を獲得
 - 複合語の除外
 - 助詞を含む単語列や記号等を除外
 - 一覧では、作品名など単語として不適当なものも存在するため
 - タイトルから上位語を抽出する
 - 上位語から細分類を判定

上位語

スター・ウォーズ登場人物一覧

A [編集]

一覧

- アイラ・セキュラ (Aayla Secura)
- アディ・ガリア (Adi Gallia)
- アクバー提督 (Admiral Ackbar)
- …
- アシュラ (Ashla)
- アスク・アック (Ask Aak)
- アソーカ・タノ (Ahsoka tano)
- オーラ・シング (Aurra Sing)
- アヌーンポンダーラ

Wiktionary 辞書構築

- Wiktionary 内のエントリを解析
 - 英語版, 日本語版を利用
 - 英語版: 8万記事
 - 日本語版: 4万記事
 - 見出し語と品詞, (動詞や形容詞の場合) 活用型を抽出
 - JUMAN と文法体系が異なるためルールにより品詞と活用型を変換
 - 複合語をWikipediaと同じ基準で除外する



[メインページ](#)
[コミュニティ・ポータル](#)
[カテゴリ](#)
[編集室](#)
[最近の更新](#)
[おまかせ表示](#)
[ヘルプ](#)
[寄付](#)

ツール

[リンク元](#)
[関連ページの更新状況](#)
[特別ページ](#)
[印刷用バージョン](#)
[この版への固定リンク](#)
[ページ情報](#)
[このページを引用](#)

他言語版

[English](#)

項目 [議論](#)

あきらか

目次 [\[非表示\]](#)

- 1 日本語
 - 1.1 形容動詞
 - 1.1.1 活用
 - 1.1.2 翻訳

日本語 [\[編集\]](#)

形容動詞 [\[編集\]](#)

あきらか [【明らか】](#)

- 物事がはっきりにしている様子。
- 明るい様子。

活用 [\[編集\]](#)

ダ型活用

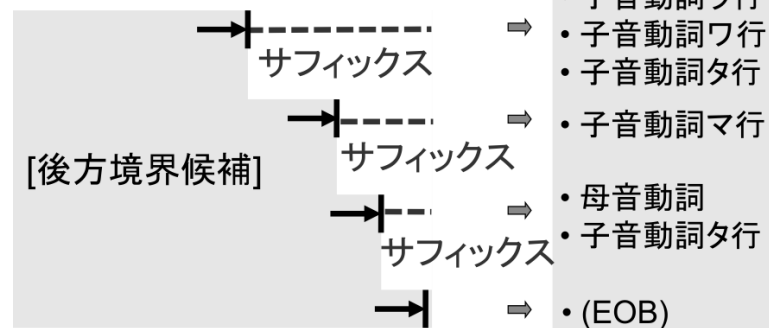
あきらか-だ

Webコーパスからの語彙獲得

• オンライン未知語獲得 (村脇, 2010) 中の品詞推定の手法を利用

- ▶ 語幹の候補に対して、後続する語尾と付属語列(サフィックス)から品詞の候補を列挙
- ▶ 語幹候補ごとに、用例をコーパス中から集め以下の条件を満たした時に品詞を確定する
 - 一定の割合で共通した品詞候補がある
 - 活用形の異なり数が一定以上

何となく **ググ** ってみた。



用例: **ググ** ってみた。

子音動詞ラ行 e.g. 走ってみた
子音動詞ワ行 e.g. 笑ってみた
子音動詞タ行 e.g. 立ってみた
子音動詞力行 e.g. 行ってみた
⋮

用例: **ググ** りたいキーワード

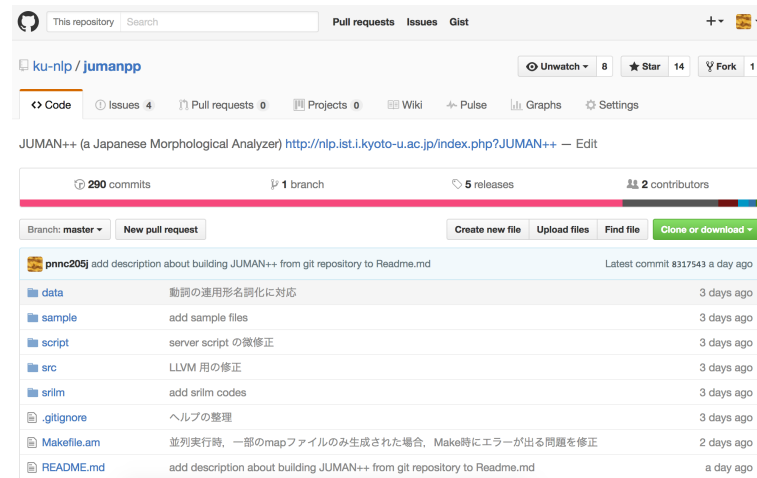
子音動詞ラ行 e.g. 走りたい
母音動詞 e.g. 降りたい

構築した辞書

	語彙数	Examples
基本語彙辞書	3万語	走る, 行く, 明日
Wikipedia	85万語	アベノミクス, Dentsu, 山極, 豊洲
Wiktionary	8千語	インセンティブ, 糾す
Web コーパス	1万語	逆進税, 政独委
合計	90万語	

JUMAN++

- 黒橋研Webページ, GitHubで公開



- 次バージョンを近日公開予定
 - 従来の4倍以上の解析速度
 - 実装から辞書依存を排除

まとめ

- JUMAN++と語彙獲得のフローを紹介
- ぜひJUMAN++を使ってみてください！