

実テキストの情報分析のための 頑健な言語処理基盤

河原大輔 黒橋禎夫* 林部祐太*¹
森田一*² Arseny Tolmachev

京都大学 大学院情報学研究科
*科学技術振興機構 CREST

¹現在の所属はフェアリーデバイセズ株式会社

²現在の所属は株式会社富士通研究所

第11回テキストアナリティックス・シンポジウム (2017/9/7)

実テキスト

セブンの唐揚げ串美味しいし安いんですよ!! でも最近丸からっていう串に刺さってない方の唐揚げにハマってます!! 個人的には串の上をいくのでは!? って位大好きです(^▽`*)

LCCとかいう大陸間弾道夜行バス、(空港の滑走路で)渋滞はするしターミナルは(都市の端にある空港の更に端で)めんどいし、(機材遅れて)出発到着は軽率に数時間単位で...

今日は内枠の先行馬が勝つレースが多くて、メインも傾向を引き継いどった。中途半端な位置にいた馬が動くに動けん感じやったけど、そんなペースに持ってくあたりがルメール。

そして昨日は、我が家でバーベキューをしました。主人の後輩の方や普段お世話になっている方々を招いて、とても楽しい時間でした。私の父も来ました。

人が意思決定するときや、企業が自社の商品の評価を分析する上で重要な情報源

実テキストの情報分析の例



「渋滞」に関する不満の例

高速道路が渋滞しているときのトイレが大変。

平気で路駐してる自動車。通勤ラッシュの時はかなり迷惑です。

車線規制して大渋滞なのに、工事をしていない。

駐車場での渋滞さらにそれに料金とられた～！

鷺ノ宮地区の踏切、毎朝空かずに渋滞して、バスが遅れる。

...

因果関係グラフ

平気で路駐
してる自動車

車線規制

渋滞

バスが遅れる

トイレが大変

宅配便が指定時間
より30分遅れた

目次

- RNN言語モデルに基づく形態素解析器JUMAN++
 - Web上のリソースからの語彙獲得
 - 部分アノテーションの利用
 - JUMAN++の高速化
- 語彙知識に基づく形態素・構文統合解析

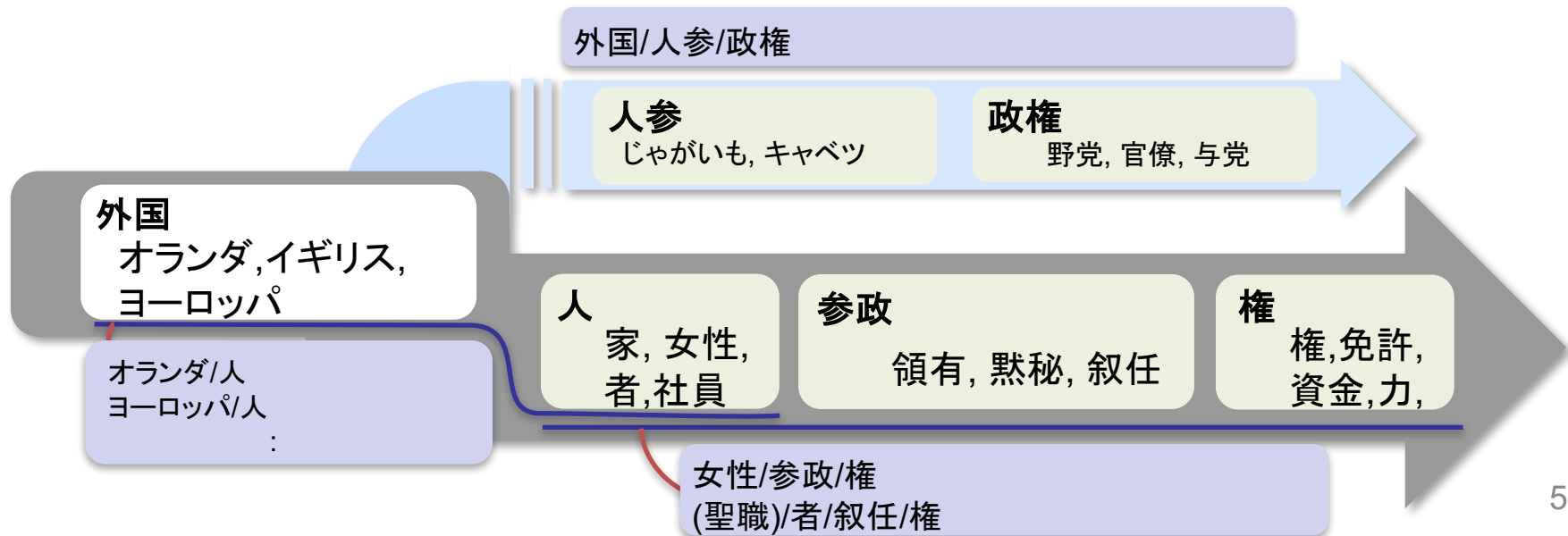
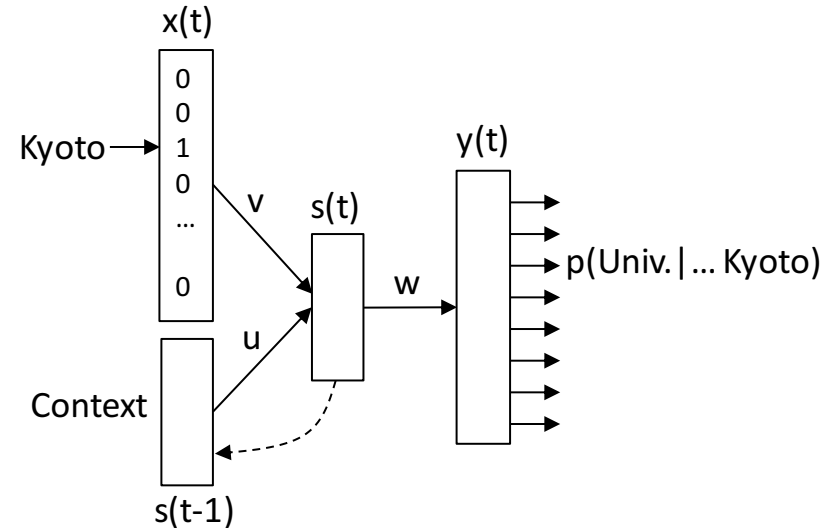
RNN言語モデルに基づく形態素解析器

JUMAN++ [Morita+, EMNLP2015]

外国|人参|政権
より

外国|人|参政|権
の方が自然だとわかりたい

意味を汎化したベクトル表現によって $p(w|\text{context})$ を計算する



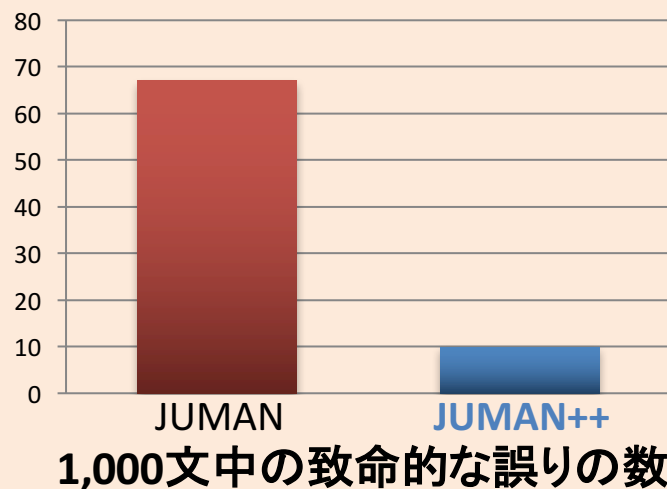
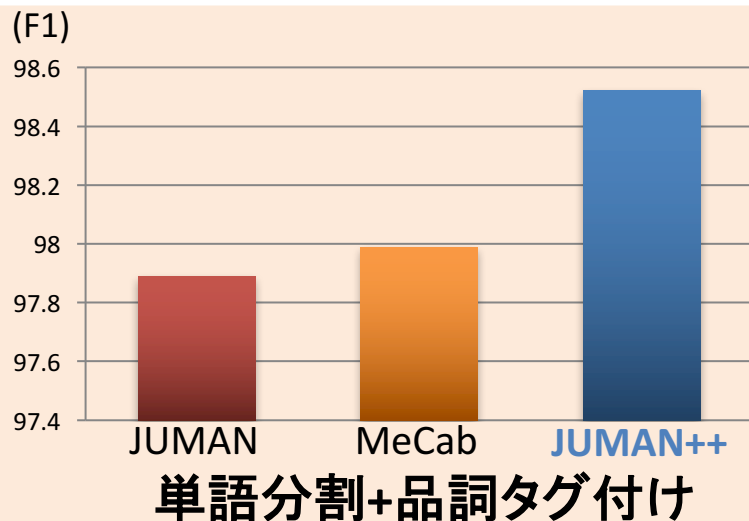
JUMAN++による精度改善

実験設定

RNN言語モデルの学習
Webコーパス 1000万文

学習・評価データ

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 | やご | 要望

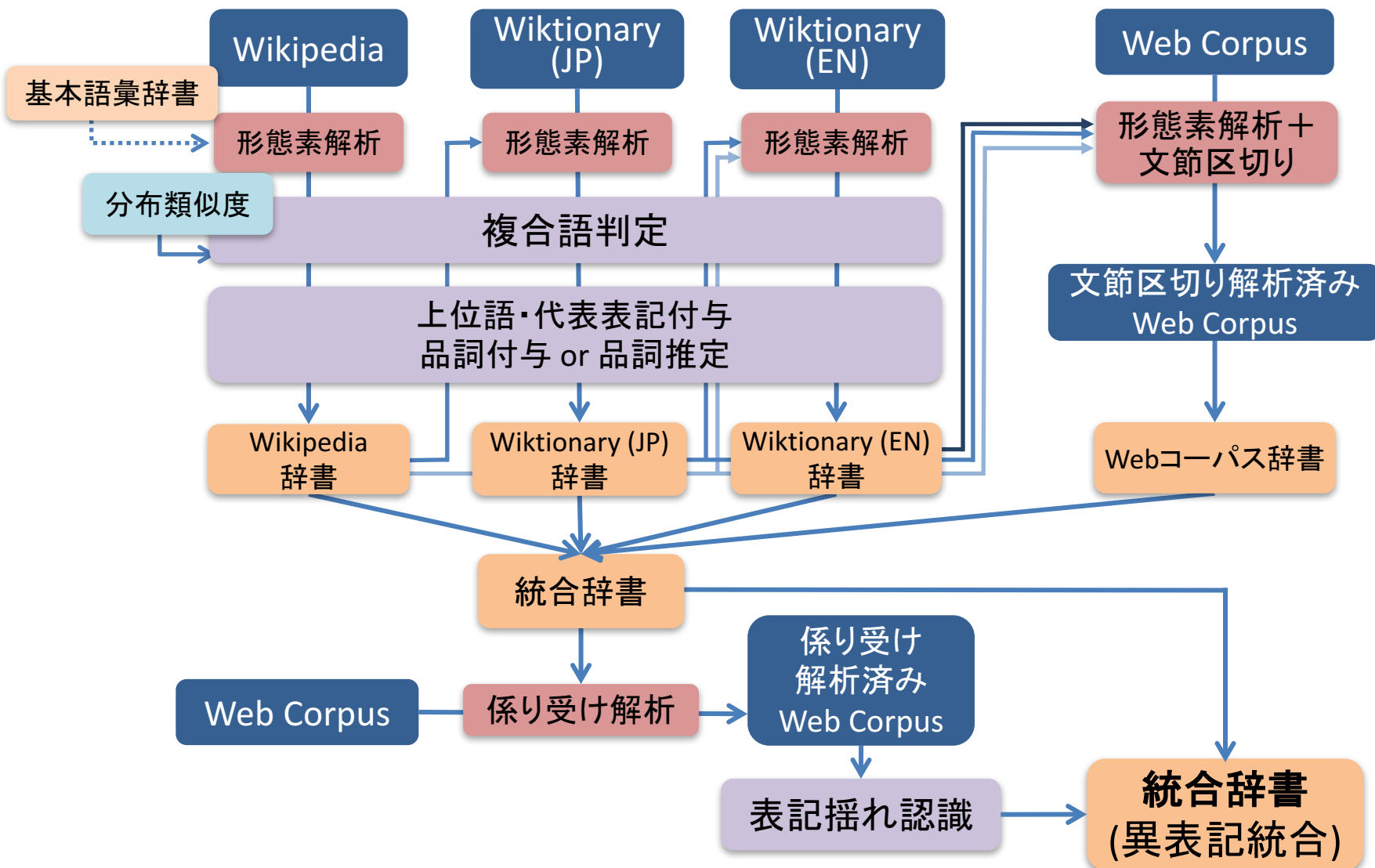
✓ JUMAN++(Proposed)

感想 | や | ご | 要望

語彙獲得の方針

- 基本的な語彙は人手で整備(基本語彙)
 - 語彙サイズを絞る代わりに、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
cf. mecab-ipadic-NEologd [佐藤+, 2017]

JUMAN++辞書獲得フロー



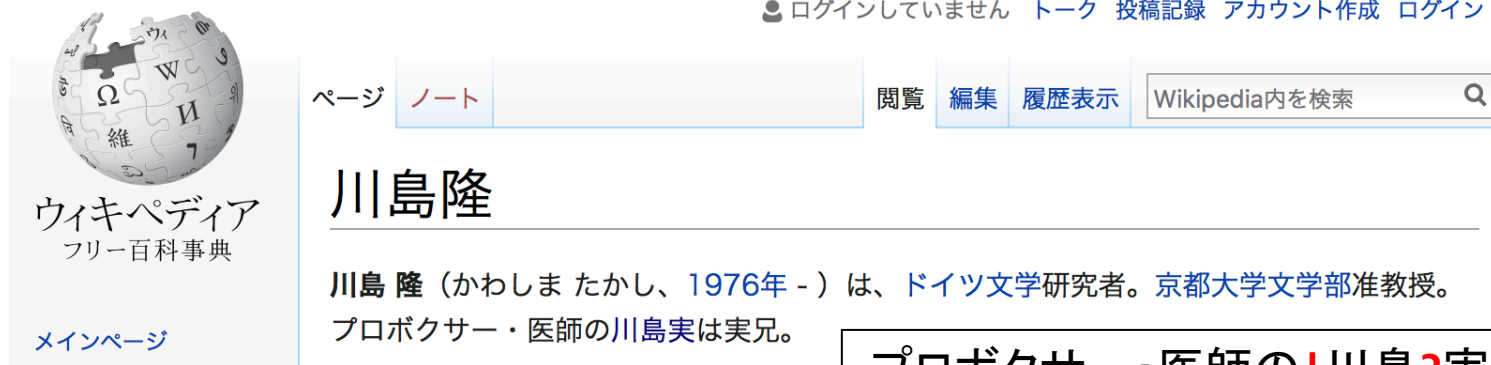
構築した辞書

	語彙サイズ	例
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	ググる, ねんどろいど
合計	90万語	

JUMAN++における部分アノテーション の利用

[林部, 2017]の部分アノテーションコーパス

- Fairy Morphological Annotated Corpus
<https://github.com/FairyDevicesRD/FairyMaCorpus>
- Wikipediaのハイパーリンクに基づく自然アノテーションと自動形態素解析結果が異なる箇所を人手でチェック

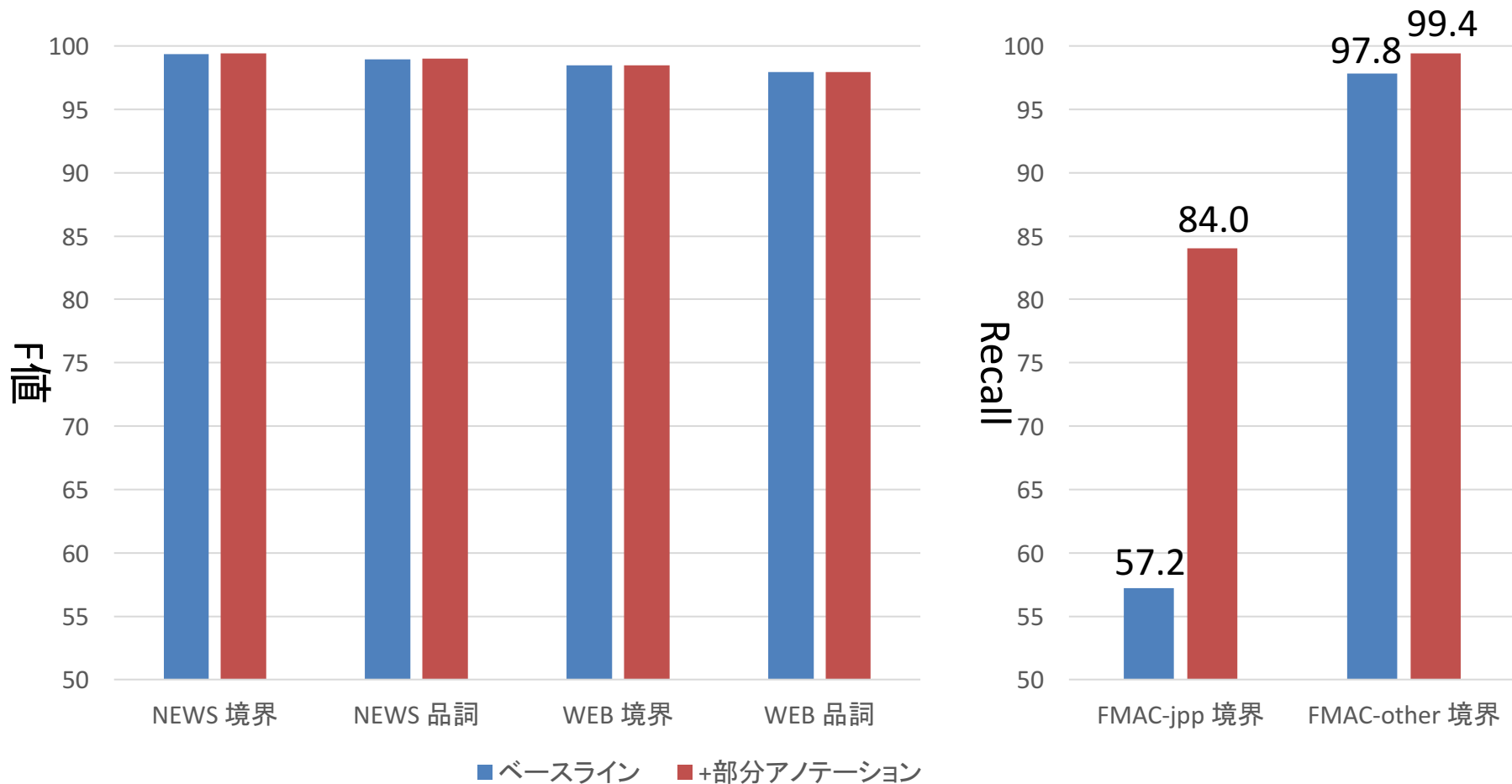


- 2,000文のうち、次の1,400文を利用
 - FMAC-jpp: JUMAN++による単語区切りと異なる箇所
 - FMAC-other: 機能表現を中心としたその他の部分アノテーション
- 例: ...遺伝子のうちのほんの|わず|か|だけ|が|機能する...

部分アノテーションを利用した JUMAN++の訓練

1. フルアノテーションの学習データを用いてJUMAN++を訓練
 - 学習データ: NEWS (3.7万文) + WEB (1.4万文)
2. 部分アノテーションの各文について、与えられた単語区切りに違反しないように、1のモデルで形態素解析
3. フルアノテーションの学習データに2のデータをマージし、これを用いてJUMAN++を再訓練

部分アノテーションを利用した JUMAN++の精度



改善例

- |越石|優|に|1-2|判定|負け|。
- 同|時期|に|は|細川|たかし|ら|が|いた|。

改善しなかった例

- 1956|年|に|退職|して|、|三|共に|入社|。
- かつて|は|ミッキー|形どら|焼き|「|ミッキー|スマイル|」|を|取り扱って|いた|。

JUMAN++の高速化

JUMAN++の高速化

- JUMAN++は高精度だが、解析速度の遅さが問題
 - 高速化版を開発
- 全体的な方針
 - CPUキャッシュの利用効率を高める
 - 冗長な計算の削減
- 詳細
 - 線形モデル
 - 辞書構造の改善により、string->idのID化のためのハッシュマップを削除
 - 素性計算の効率化
 - RNNモデル
 - スコア計算のバッチ化、ベクトル化

形態素解析の解析時間 (秒)

	1文	10文	100文	1000文	2000文
MeCab	0.013	0.014	0.017	0.047	0.080
JUMAN	0.058	0.060	0.084	0.220	0.369
KyTea	6.010	6.068	6.047	6.377	6.737
JUMAN++V1	0.157	0.592	5.794	60.460	141.730
JUMAN++V2	0.085	0.120	0.371	2.318	5.106

JUMAN++V1: 従来版, JUMAN++V2: 高速化版

- 高速化に関する今後の課題
 - 線形モデル
 - Unigram・Bigramの素性計算の効率化
 - RNNモデル
 - 完全に同じ状態・スコアのRNN計算の削減
(ひらがな表記の語など、同一の表層形、品詞を持つ語)

語彙知識に基づく 形態素・構文統合解析

[Kawahara+, IWPT2017]

逆転する可能性があるかないかを確認する

- 逆転する可能性が有るかないかを確認する
- 逆転する可能性が歩くかないか確認する

格フレーム

有る:3	
ガ	可能性 :121867
ニ	価格:23, 自分:20, ...
デ	段階:4, 影響:4, 確率:4, ...

歩く:1	
ガ	人:57, 男性:15, 私:13, ...
ヲ	道:24236, 山道:4066, ...
カラ	駐車場:175, 駅:88, ...

可									
	能								
		性							
			が						
				あ					
					る				
						か			
							な		
								い	
									か

Input: 可能性があるかないか

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

Input: 可能性があるかないか

1. 形態素解析のN-best解に含まれる単語を
CKYテーブルに貼り付け

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

Input: 可能性があるかないか

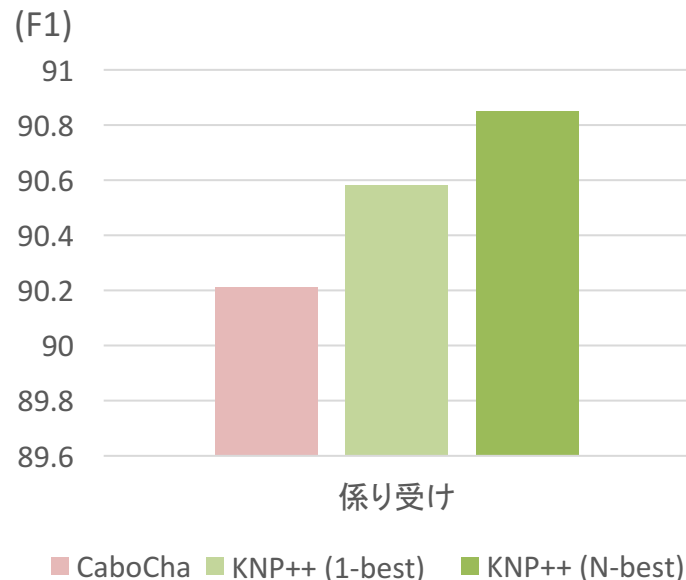
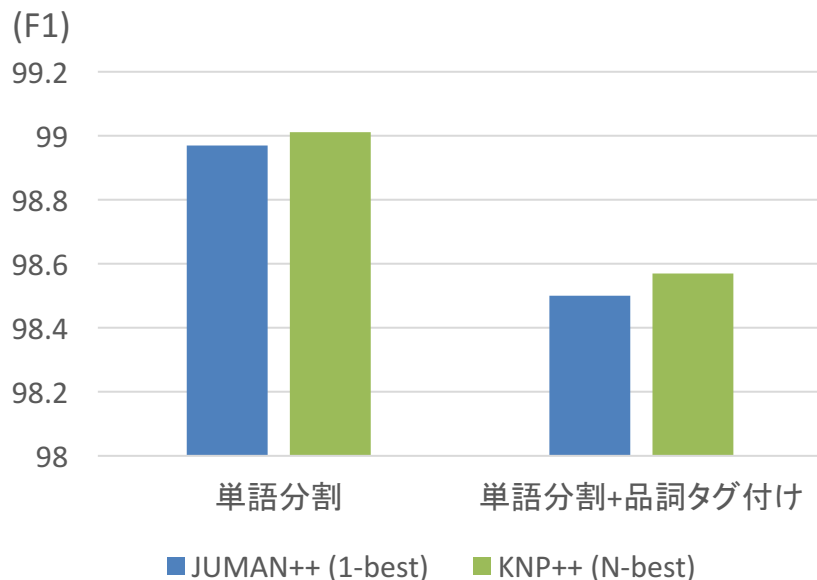
- 形態素解析のN-best解に含まれる単語をCKYテーブルに貼り付け
- 単語を基に、可能な句を作る
 - 句 = 自立語1個 + 付属語0個以上



1. 形態素解析のN-best解に含まれる単語をCKYテーブルに貼り付け
2. 単語を基に、可能な句を作る
 - 句 = 自立語1個 + 付属語0個以上
3. 可能な係り受けを作り，格フレームなどに基づくスコアを計算する

形態素・構文統合解析の評価結果

評価データ: NEWS+WEB



改善例

- お|いや|めい|と|別れた
- 祭り|で|は|で|もの|が|
み|つ|か|り|ま|す
- 泥|を|落として|から|部屋|に|
は|い|る



- おい|や|めい|と|別れた
- 祭り|で|は|でもの|が|
み|つ|か|り|ま|す
- 泥|を|落として|から|部屋|に|
は|い|る

まとめ

- **実テキスト**の情報分析に向けた頑健な言語処理基盤を開発
 - RNN言語モデルに基づく形態素解析器JUMAN++
 - 語彙知識に基づく形態素・構文統合解析器KNP++ (仮称)
- 近日公開予定！