Vol.2010-NL-198 No.8
2010/9/17

点予測による形態素解析

中田陽介[†1]　NEUBIG Graham [†1]
森信介[†1]　河原達也[†1]

本論文では、形態素解析の問題を単語分割と品詞推定に分解し、それぞれの処理で点予測を用いる手法を提案する。点予測とは、分類器の素性として、周囲の単語境界や品詞等の推定値を利用せずに、周囲の文字列の情報のみを利用する方法である。点予測を用いることで、柔軟に言語資源を利用することができる。特に分野適応において、低い人的コストで、高い分野適応性を実現できる。提案手法の評価として、言語資源が豊富な一般分野において、既存手法である CRF と形態素 n-gram モデルと品詞 2-gram モデル（HMM）との解析精度の比較を行い、同程度の精度を得た。さらに、提案手法の分野適応性を評価するための評価実験を行い、高い分野適応性を示す結果を得た。

Morphological Analysis with Pointwise Predictors

Yosuke NAKATA ,[†1] Graham NEUBIG ,[†1]
Shinsuke MORI [†1] and Tatsuya KAWAHARA[†1]

This paper proposes an approach to Japanese morphological analysis that divides the prediction process into word segmentation and part-of- speech estimation, then solves each step with pointwise predictors. Pointwise prediction uses as its feature set only surface information about the surrounding character strings, without relying on predicted information such as surrounding POS tags or word boundaries. This allows for the flexible use of a variety of linguistic resources, making it possible to achieve domain adaptation with a minimum amount of annotation. An evaluation was performed on a well-resourced general domain morphological task, and it was found that the proposed method achieved results comparable to those of existing methods such as CRFs, morpheme n-gram models, and POS 2-gram models (HMM). In addition, a domain adaptation experiment found that the proposed method was able to achieve effective domain adaptation with a smaller amount of annotation.

1. はじめに

形態素解析は、日本語における自然言語処理の基礎であり、非常に重要な処理である。形態素解析の入力は文字列であり、出力は単語と品詞の組（形態素）の列である。形態素解析の出力は、固有表現抽出や構文解析、あるいはテキストマイニング等の入力となる。そのため、形態素解析の精度は後続の処理に大きな影響を与える。したがって、様々な分野のテキストに対して高い解析精度が要求されている。

形態素解析の研究は非常に多くある[1)–5)]。これらの研究では、文を形態素列とみなし、形態素解析を単語分割と品詞推定を同時に行う系列ラベリングの問題として扱っている。これらの先行研究は、人手で記述した規則に基づく方法とコーパスから規則などを学習する方法に大別される。規則に基づく JUMAN[2)] では、ある品詞体系を設定し、品詞間での連接コストや単語の生起コストを人手で与えている。そのため、規則の作成時に想定していない分野のテキストに対して高い解析精度を実現するには、微妙なコスト調整などに膨大な人的コストがかかるという問題点がある。この問題を軽減するために、連接コストや単語生起コストなどを隠れマルコフモデル（HMM）[⋆1]のパラメータとみなし、これらをコーパスから学習する手法が提案されている[1),3)]。HMM(品詞 n-gram モデル) の改良として、すべての品詞を語彙化した形態素 n-gram モデルと形態素クラスタリングの結果を用いるクラス n-gram モデル[4)] や、柔軟な素性選択を可能とする条件付き確率場 (CRF; Conditional Random Fields) の利用[5)] が提案されている。これらのコーパスに基づく手法では、パラメータはコーパスから自動的に推定されるので、汎用性が高い。すなわち、当初想定していない分野のテキストに対して高い解析精度を実現するためには、その分野の例文に適切な単語境界情報と品詞情報を付与することで得られる学習コーパスを用意すれば十分である。これらの手法では、学習コーパスは、すべての文字間に単語境界情報を付与しすべての単語に品詞を付与したフルアノテーションコーパスである必要がある。したがって、コーパス作成の作業コストは小さくないが、規則の微調整に比べれば作業者を選ばないことや、作業量に応じて精度が確実に向上するなどの利点がある。

近年、各分野での文書の電子化が進み、各分野で自然言語処理を用いた文書の自動処理の

†1 京都大学 情報学研究科
Kyoto University, School of Informatics

⋆1 品詞が付与されたコーパスからパラメータを推定する場合には、状態は明示されており隠れていないので、HMM と呼ぶのは不適切と考えられるが、ここでは慣例に従って HMM と呼ぶ。

Vol.2010-NL-198 No.8
2010/9/17

要求が高まっている。例えば、医療分野ではカルテや退院サマリの電子化が進み、テキストマイニングを用いてこれらの医療文書の薬物有害事象の自動抽出[6] といった要求がある。医療分野では、一般分野で学習した形態素解析器の解析精度は著しく低下し、テキストマイニングなどの自然言語処理全体の精度が著しく低下する。この例からも分かるように、大量の学習コーパスが存在しない様々な分野のテキストに対して、高い形態素解析精度を低コストかつ短時間で実現することが大きな課題となっている。新聞記事などの一般分野と比べるとフルアノテーションコーパスの量は少ないが、機械可読のテキストや専門用語辞書 (主に複合語) などの言語資源が様々な分野で増えている。したがって、高い分野適応性を実現するための鍵は、様々な言語資源を柔軟に利用可能とする設計である。前述の先行研究では、与えられた学習コーパスを用いて高い精度を実現することに主眼が置かれ、分野適応性を考慮した設計にはなっていない。結果として、既存の形態素解析器では、効率的な分野適応は実現されておらず、自然言語処理を様々な分野のテキストに適用する障害となっている[⋆1]。

本論文では、上述の形態素解析の現状と要求を背景として、大量の学習コーパスがある分野で既存手法と同程度の解析精度を保持しつつ、高い分野適応性を実現する形態素解析器の設計を提案する。具体的には、形態素解析を単語分割と品詞推定に分解し、それぞれを点予測を用いて解決することを提案する。点予測とは、推定時の素性として、周囲の単語境界や品詞情報等の推定値を参照せずに、周辺の文字列の情報のみを参照する方法である。提案する設計により、単語境界や品詞が文の一部にのみ付与された部分的アノテーションコーパスや、品詞が付与されていない単語や単語列からなる辞書などの言語資源を利用することが可能となる。この結果、従来手法に比して格段に高い分野適応性を実現できる。

2. 点予測を用いた形態素解析

本論文では、形態素解析を単語分割と品詞推定に分けて段階的に処理する手法を提案する (図 1 参照)。それぞれの処理において、単語境界や品詞の推定時に、推定結果しか存在しない動的な情報を用いず、周辺の文字列情報のみを素性とする点予測を用いる。

2.1 点予測を用いた単語分割

点予測による単語分割には先行研究[8)9)] があり、本研究ではこれを利用する。点予測による単語分割の入力は文字列 $\boldsymbol{x} = x_1 x_2 \cdots x_n$ であり、各文字間に単語境界の有無を示すタグ $\boldsymbol{t} = t_1 t_2 \cdots t_{n-1}$ を出力する。単語境界タグ t_i がとりうる値は、文字 x_i と x_{i+1} の間に単語境界が「存在する」か「存在しない」の 2 種類で、2 値分類問題として定式化される。点予測による単語分割では、以下の 3 種類の素性を参照する SVM による分類を行っている。

(1) 文字 n-gram: 判別するタグ位置 i の前後の部分文字列であり、窓幅 m と長さ n のパラメータがある。素性は、長さ $2m$ の文字列 $x_{i-m+1}, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_{i+m}$ の長さ n 以下のすべての部分文字列 (文字 n-gram) である (図 2 参照)。

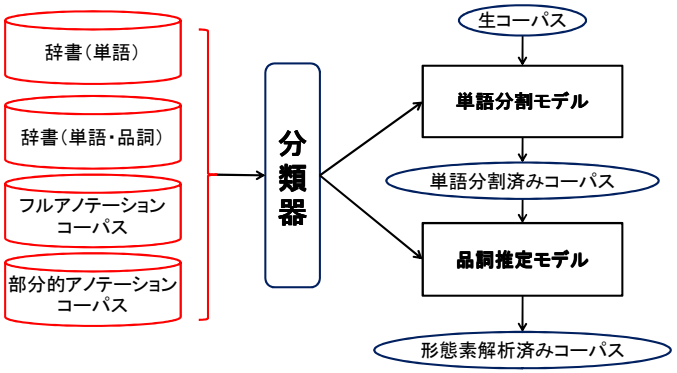

図 1 処理の流れ

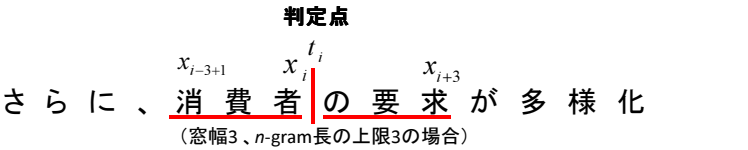

図 2 単語分割に使用する素性

⋆1 CRF のパラメータを部分的アノテーションコーパスから推定する研究[7] もあるが、能動学習などの際に生じる非常にスパースかつ大規模な部分的アノテーションコーパスからの学習の場合には、必要となる主記憶が膨大で、現実的ではない。

(2) 文字種 n-gram: 文字を文字種に変換した列を対象とする点以外は文字 n-gram と同じである。文字種は、漢字（K）、片仮名（k）、平仮名（H）、ローマ字（R）、数字（N）、その他（O）の 6 つである。

(3) 単語辞書素性: 判別するタグ位置 i を始点とする単語、終点とする単語、内包する単語が辞書にあるか否かのフラグと、その単語の長さである。

2.2 点予測を用いた品詞推定

様々な言語資源を有効活用するために、点予測による単語分割の考え方を拡張し、点予測を用いた品詞推定手法を提案する。品詞推定の入力は、一般的な単語列である[10]。しかしながら、点予測を用いた品詞推定では、推定する単語 w とその前の文脈の文字列 $\boldsymbol{x}_-$ と後の文脈の文字列 $\boldsymbol{x}_+$ を入力とし、これらのみを参照して単語 w の品詞を推定する多値分類問題として定式化する。参照する窓幅を m' とすると、入力において参照される情報は $x_{-m'}\cdots x_{-2}x_{-1}, w, x_1x_2\cdots x_{m'}$ となる。すなわち、この文字列と w の前後に単語境界があり内部には単語境界がないという情報のみから w の品詞を推定する。換言すれば、周囲の単語の品詞の推定結果や、推定対象の単語以外の単語境界情報を一切参照しない。この設計により、後述のように、パラメータ推定時に様々な言語資源の柔軟な活用が可能となる点に注意されたい。

品詞推定に利用する素性は以下の通りである（図 3 参照）。

(1) $\boldsymbol{x}_-\boldsymbol{x}_+$ に含まれる文字 n-gram

(2) $\boldsymbol{x}_-\boldsymbol{x}_+$ に含まれる文字種 n-gram

また、本手法では、単語によって異なる以下の 4 つの種類の処理を行う。

(1) 学習コーパスに品詞候補が複数出現する単語は、分類器で推定を行う。

(2) 学習コーパスに品詞候補が 1 つしか出現しない単語には、その品詞を付与する。

(3) 学習コーパスに出現しないが辞書に出現する単語には、辞書に含まれる初めの品詞を付与する。

(4) 学習コーパスにも辞書にも出現しない単語には、名詞を付与する。

単語分割とは異なり、品詞推定は多値分類である。したがって、各単語の品詞候補毎の分類器を作る。つまり、ある単語に品詞候補が 3 つ存在すれば分類器はその単語に対して 3 つ作り、推定には one v.s. rest 法を用いて多値分類を行う。

提案手法では、単語毎に多値分類器を作成するが、全単語に対して 1 つの多値分類器を one v.s. rest 法を用いて作るという方法も考えられる。予備実験で、この手法を能動学習で用いたところ、能動学習に対して頑健性が低く、偏ったデータを学習データに利用すると解析精度が大幅に下がる現象が起きたので本論文では利用しないこととした。

$x_{-3}x_{-2}x_{-1}$ w x_1 x_2 x_3

さらに、消費者の 要求 が多様化

（窓幅3 、n-gram長の上限3の場合）

文字（種）1-gram： -3/費（K） -2/者（K） -1/の（H） 1/が（H） 2/多（K） 3/様（K）
文字（種）2-gram： -3/費者（KK） -2/者の（KH） -1/のが（HH） 1/が多（HK） 2/多様（KK）
文字（種）3-gram： -3/費者の（KKH） -2/者のが（KHH） -1/のが多（HHK） 1/が多様（HKK）

図 3 品詞推定に使用する素性

2.3 点予測による柔軟な言語資源利用

点予測を用いた単語分割、および品詞推定は、入力から計算される素性のみを参照し、周囲の推定値を参照しないので、様々な言語資源を柔軟に利用することができる。

既存手法による系列ラベリングとしての形態素解析器のパラメータ推定には、一般的に次の 2 つの言語資源のみが利用可能である。これらは提案手法でも利用可能である。

(1) フルアノテーションコーパス: すべての文字間に単語境界情報が付与され、すべての単語に品詞が付与されている。一般的に、形態素解析の学習コーパスには、これを用いる。既存手法の分野適応に際しては、適応対象の文に対して人手によりこれらの情報を付与する必要があるが、各文の大部分の箇所は、一般分野のコーパスにすでに出現している単語や表現であり、文のすべての箇所に情報を付与することは効率的ではない。

(2) 形態素辞書: この辞書の各見出し語には、品詞が付与されている。これを作成する作業者は、対象分野の知識に加えて、単語分割基準と品詞体系の両方を熟知している必要がある。

フルアノテーションコーパスを作成する作業者は、不明な箇所や判断に自信のない箇所が含まれる文に対しては、その文すべてを棄却するか、確信の持てない情報付与をすることとなる。また、形態素辞書を作成する際にも、単語であることのみに確信があり、品詞の判断に自信がない場合、その単語を辞書に加えないか、確信の持てない品詞付与をすることになる。

このような問題は、言語資源作成の現場では非常に深刻であり、確信の持てる箇所のみへのアノテーションを許容する枠組みが渇望されている。提案手法では以下のような部分的な情報付与の結果得られる言語資源も、有効に活用することができる（図 4 参照）。

- **部分的単語分割コーパス**
  例)川の 流れ に任せて流れる
- **部分的品詞付与コーパス**
  例)川の 流れ/名詞 に任せて 流れ/動詞 る
- **単語列**
  例)香川 大学，鴨川
- **単語と品詞の組(形態素)の列**
  例)川/名詞 流れ/名詞，流れ/動詞，受入れ/名詞，受入れ/動詞

図 4 提案手法で利用可能な言語資源

(3) 部分的アノテーションコーパス: 文の一部の文字間の単語境界情報や一部の単語の品詞情報のみがアノテーションされたコーパスである。形態素解析という観点では、単語境界情報のみが付与された単語分割済みコーパスも部分的アノテーションコーパスの一種である。ほかに、部分的単語分割コーパスや部分的品詞付与コーパスなどがある。

(4) 単語辞書: 単語の表記のみからなる辞書であり、比較的容易に入手可能である。自動単語分割の際に単語境界情報として利用できる。

フルアノテーションコーパスは、各分野で十分な量を確保することは難しいが、上記の言語資源は比較的簡単に容易することができる。本手法では、これらの様々な言語資源を有効活用することにより、高い分野適応性を実現する。

2.4 分野適応戦略

本項は、分野適応戦略について述べる。最も一般的な分野適応の戦略は、適応分野のフルアノテーションコーパスを用意することであるが、作成に必要な人的コストが膨大であるという問題がある。低い人的コストで高い効果を得るためには、推定の信頼度が低い箇所に優先的にアノテーションを行うことが望ましい。単語境界や品詞の推定の信頼度は、文内の各箇所で異なるので、アノテーションは文単位ではなく、推定対象となる最小の単位であるべきである。このようなアノテーションの結果、部分的アノテーションコーパスが得られる。既存手法の形態素解析器では、部分的アノテーションコーパスの利用は困難であるが、提案手法では周囲の文字列の情報のみを用いて形態素解析を行うので、部分的アノテーションコーパスの利用が容易である。

そこで、分野適応戦略として、形態素解析器の学習と部分的アノテーションを交互に繰り返し行う能動学習を採択する。アノテーション箇所の候補は、分類器の判断の信頼度が低い単語分割箇所と品詞推定対象の単語である。信頼度の尺度は、SVM の分離平面からの距離[8] であり、単語分割箇所と品詞推定の単語を一括して比較する。実際のアノテーションは、選択された箇所 (選択箇所) に応じて以下のように行う。

(1) 選択箇所が単語分割箇所 (文字間) の場合: 以下の 2 通りに分類する。
  (a) 選択箇所が単語内の場合: その単語の内部と前後の単語境界情報および品詞情報を付与する。
  (b) 選択箇所が単語境界の場合: その前後の単語の内部と前後の単語境界情報および品詞情報を付与する。

(2) 選択箇所が品詞推定箇所 (単語) の場合: その単語の内部と前後の単語境界情報および品詞情報を付与する。

提案する分野適応の手順を以下に示す。

(1) 一般分野の学習コーパス（フルアノテーションコーパス）で分類器の学習を行う。
(2) 適応分野の学習コーパス（初期状態は生コーパス）に対して形態素解析を行い、推定の信頼度が低い 100 箇所を選択する[⋆1]。
(3) 選択した箇所を作業者に提示し、単語境界と品詞を付与してもらう。その結果、適応分野の学習コーパス（部分的アノテーションコーパス）が得られる。
(4) 一般分野の学習コーパスと適応分野の学習コーパス（部分的アノテーションコーパス）を用いて分類器の再学習を行う。
(5) 上記の（２）〜（４）の手順を繰り返す。

3. 評　　価

提案手法の評価を行うために２つの評価実験を行った。１つは、言語資源が豊富な一般分野のコーパスで学習を行い、提案手法と既存手法の解析精度を比較するもので、もう１つは、提案手法を用いた分野適応性の評価である。それぞれの実験で、提案手法と既存手法の形態素解析の解析精度（F 値）を比較した。なお、学習コーパスのみを用いた予備実験により文字 n-gram 長の n の上限値、文字種 n-gram 長の n の上限値、窓幅 m, m' はすべて 3 とした。なお、分類器には、精度と学習効率を考慮して線形 SVM[11] を用いた。

⋆1 理論的には、1 箇所のアノテーション毎に分類器の再学習を行うべきであるが、それでは作業者の待ち時間の合計が非常に長くなる。また、予備実験で 1 箇所を選んだ場合の精度は 100 箇所を選んだ場合の精度と有意な差とならなかった。

Vol.2010-NL-198 No.8
2010/9/17

表 1 コーパス

コーパス名	出典	用途	文数	形態素数	文字数
日本語書き言葉均衡コーパス	白書・書籍・新聞	学習	27,338	782,584	1,131,317
	（一般分野）	テスト	3,038	87,458	126,154
(BCCWJ)	Yahoo!知恵袋	学習	5,800	114,265	158,000
		テスト	645	13,018	17,980

表 2 UniDic 使用時の一般分野および Yahoo!知恵袋に対する形態素解析精度（カバレージが 100% に近い場合の参考データ）

手法	分野（カバレージ）	F 値 ($\beta = 1$)
CRF(MeCab-0.98)	一般	99.23
提案手法	(99.95)	98.86
CRF(MeCab-0.98)	Yahoo!知恵袋	97.54
提案手法	(99.80)	96.83

3.1 コーパス

実験には「日本語書き言葉均衡コーパス」コアデータ（BCCWJ）[12)⋆1]を用いた。コーパスは単語分割と品詞付与が人手で行われている。出典は、白書と書籍と新聞と Yahoo!知恵袋である。Yahoo!知恵袋は、他の出典のデータと大きく性質が異なる[13)]ので Yahoo!知恵袋を適応分野とし、白書と書籍と新聞を一般分野とする。コーパスの詳細を表 1 に示す。

本論文の目的は、1 節で述べた通り、分野適応性が高い形態素解析の枠組みの提案である。一般分野に出現する単語を語彙とした場合、医療分野のコーパス (単語境界情報のみ付与) のカバレージは 97.18% であった。一般分野に出現する形態素を語彙とした場合の Yahoo!知恵袋のカバレージは 96.29% であり、Yahoo!知恵袋を分野適応の対象とすることはおおむね妥当と考えられる。

なお、BCCWJ と同じ基準の約 22 万形態素を収めた辞書（UniDic）を利用して実験することも考えられるが、これを語彙に加えた場合の Yahoo!知恵袋のカバレージは 99.80%と非常に高くなる。これは、ほぼ未知語が存在しない状況であり、分野適応の実験としては、現実的な設定でない。したがって、提案手法の評価実験においては基本的に UniDic を使用しないことするが、UniDic を使用した場合の形態素解析精度を参考として表 2 に提示する。

⋆1 正確には、「現代日本語書き言葉均衡コーパス」モニター公開データ（2009 年度版）である。

表 3 一般分野に対する形態素解析精度

手法	適合率 [%]	再現率 [%]	F 値 ($\beta = 1$)
品詞 2-gram モデル（HMM）	93.77	94.27	94.02
形態素 2-gram モデル	96.58	97.65	97.11
形態素 3-gram モデル	96.70	97.73	97.21
CRF（MeCab-0.98）	96.72	97.84	97.28
提案手法（KyTea-0.1.1）	98.07	98.06	98.06

表 4 Yahoo!知恵袋に対する形態素解析精度

手法	適合率 [%]	再現率 [%]	F 値 ($\beta = 1$)
品詞 2-gram モデル（HMM）	86.78	87.96	87.36
形態素 2-gram モデル	92.01	94.09	93.04
形態素 3-gram モデル	92.10	94.24	93.16
CRF（MeCab-0.98）	93.69	95.65	94.66
提案手法（KyTea-0.1.1）	95.19	95.51	95.35

3.2 既存手法の詳細

比較対象とした既存手法は、先行研究で高い精度が報告されている CRF（MeCab-0.98）[5)]と、形態素 n-gram モデル[4)](n=2,3)、品詞 2-gram モデル（HMM）[1)] である。CRF の素性とする語彙は、予備実験の結果、学習コーパスに出現する全単語から低頻度語 500 語を取り除いたものとした。予備実験の結果を踏まえて、学習コーパスの出現頻度上位 5,000 語を語彙化した。素性は、品詞、文字種、表記 2-gram と品詞 2-gram、形態素 2-gram である。素性列から内部状態素性列に変換するマッピング定義の 1-gram には、品詞、表記を用い、右文脈 2-gram、左文脈 2-gram には、品詞 2-gram と、語彙化された単語を用いた。

3.3 評価実験 1 —既存手法との比較—

言語資源が豊富な一般分野のコーパスで学習を行い、提案手法と既存手法の解析精度の比較実験を行った。解析対象は、一般分野のテストコーパスと、Yahoo!知恵袋のテストコーパスとした。4 つの既存手法と提案手法で、テストコーパスに対して形態素解析を行い、解析精度を測定した。一般分野の結果を表 3 に、Yahoo!知恵袋の結果を表 4 に示す。

まず、品詞 2-gram モデルは、最も精度が低いことがわかる。CRF との比較は文献[5)] に述べられている結果と同じで、CRF の精度のほうが高い。また、形態素 n-gram モデルとの比較は、文献[4)] に述べられている予測力の比較から想定される結果と同じで、形態素 n-gram モデルの精度のほうが高い。次に、CRF と形態素 n-gram モデルとの比較についてである。

両分野においても CRF がより高い精度となっている。最後に、提案手法と既存手法を比較すると提案手法は、両分野において、既存手法で一番高い精度を示した CRF よりも高い解析精度を示した。

3.4 評価実験 2 —分野適応実験—

提案手法の分野適応性を評価するために以下の 4 つの手法を比較した。部分的アノテーションの手順は 2.4 に従う。初期の学習には、一般分野の学習コーパスを利用する。適応分野を Yahoo!知恵袋とする。

Pointwise:part 提案手法の形態素解析器を用いた部分的アノテーション手法：一般分野コーパスで学習を行い、Yahoo!知恵袋の学習コーパスを生コーパスとみなして形態素解析を行う。単語境界推定または品詞推定の信頼度の低い 100 箇所に対して、単語アノテーションを行い、部分的アノテーションコーパスを作成する。部分的アノテーションコーパスを一般分野の学習コーパスに加えて、分類器の再学習を行う。同様の手順を、単語アノテーション箇所が 20,000 となるまで繰り返した。

Pointwise:full 提案手法の形態素解析器を用いたフルアノテーション手法：Yahoo!知恵袋の学習コーパスに文単位でフルアノテーションを行う。この際、文の内容が偏らないように、ランダムに文を選択し、能動学習で単語アノテーションした単語数とほぼ同じになるようにアノテーションを行った。

CRF:part CRF を用いた部分的アノテーション手法：Pointwise:part で部分的アノテーションした単語を CRF の語彙として追加したものを CRF での単語アノテーションとした。

CRF:full CRF を用いたフルアノテーション手法：Pointwise:full でフルアノテーションした文に出現する単語を CRF の語彙として追加し、その文を CRF の学習コーパスに追加したものを CRF のフルアノテーションとした。

それぞれで学習したモデルで Yahoo!知恵袋テストコーパスに対して形態素解析を行い、その精度を測定した。その結果を図 5 に示す。

まず各形態素解析器において、フルアノテーションと部分的アノテーションでは、部分的アノテーションの方が解析精度の向上に貢献していることがわかる。また、CRF のフルアノテーションと提案手法のフルアノテーションの解析精度の上昇率はほぼ同じであることがわかる。最後に、CRF の部分的アノテーションと提案手法の部分的アノテーションでは提案手法の部分的アノテーションの方が解析精度の上昇が大きい。すなわち、提案手法の方が、分野適応性が高いことがわかる。

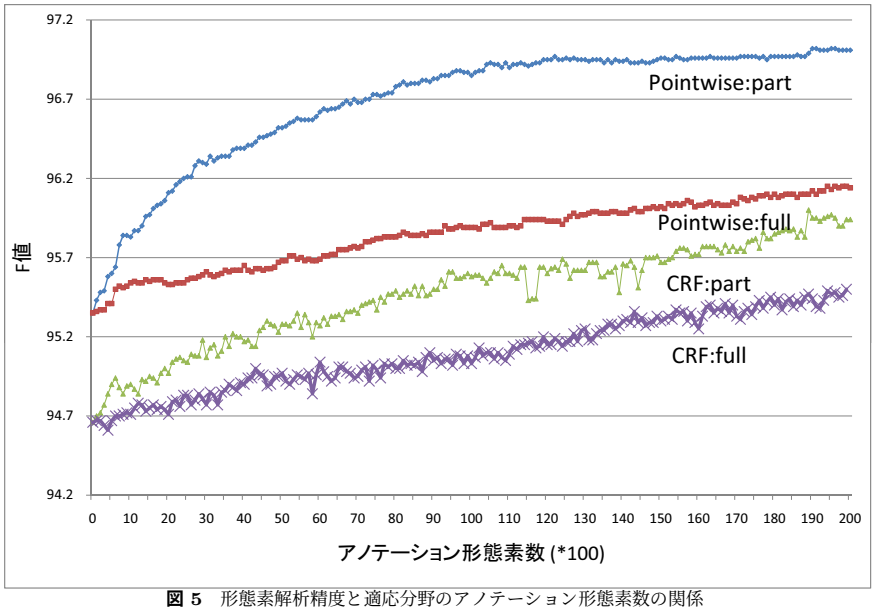

図 5 形態素解析精度と適応分野のアノテーション形態素数の関係

4. おわりに

本論文では、点予測による形態素解析手法の提案を行った。言語資源が豊富な一般分野のコーパスで学習を行い、提案手法と既存手法の解析精度の比較を行った。さらに、提案手法の分野適応性の評価を行った。提案手法を用いた形態素解析は、カバレージが低い場合、既存手法より高い解析精度を示した。カバレージが高い場合は CRF よりも低い精度となったがほぼ同等の精度を示した。分野適応の評価実験でも、フルアノテーション手法、既存手法と比較して高い分野適応性を示す結果となった。

参 考 文 献

1) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proceedings of the 15th International Conference on Computational Linguistics*, pp.201–207 (1994).
2) 松本裕治：日本語形態素解析システム JUMAN 使用説明書 version 2.0，日本語形態素解析システム JUMAN 使用説明書 version 2.0 (1994).

3) 松本祐治：形態素解析システム「茶筌」，情報処理学会誌， Vol.41, No.11, pp.1208–1214 (1997).
4) 森信介，長尾眞：形態素クラスタリングによる形態素解析精度の向上，自然言語処理，Vol.5, No.2, pp.75–103 (1998).
5) 工藤拓，山本薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告. 自然言語処理研究会報告， Vol.2004, No.47, pp.89–96 (2004).
6) 金子周司：テキストマイニングによる薬物有害事象の自動抽出を目的としたオントロジー構築とシステム開発に関する研究 (2009).
7) 坪井祐太，森信介，鹿島久嗣，小田裕樹，松本裕治：日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習，情報処理学会論文誌，Vol.50, No.6, pp.1622–1635 (2009).
8) Neubig, G.，中田陽介，森信介：点推定と能動学習を用いた自動単語分割器の分野適応，言語処理学会第 16 回年次大会，東京 (2010).
9) 颯々野学：日本語単語分割を題材としたサポートベクタマシンの能動学習の実験的研究，自然言語処理， Vol.13, No.2, pp.27–41 (2006).
10) DeRose, S.J.: Grammatical Category Disambiguation by Statistical Optimization, *Computational Linguistics*, Vol.14, No.1, pp.31–39 (1988).
11) Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classication, *Journal of Machine Learning Research*, Vol.9, pp.1871–1874 (2008).
12) 前川喜久雄：KOTONOHA『現代日本語書き言葉均衡コーパス』の開発，日本語の研究，Vol.4, No.1, pp.82–95 (2008).
13) Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H. and Den, Y.: Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese, *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (2010).