点予測による形態素解析

森 信介[†] • 中田 陽介[††] • NEUBIG Graham[††] • 河原 達也[†]

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、既存手法である条件付き確率場と形態素 n-gram モデル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。

キーワード：形態素解析、単語分割、品詞付与、点予測、系列予測

Morphological Analysis with Pointwise Predictors

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

This paper proposes a pointwise approach to Japanese morphological analysis that decomposes the process into word segmentation and part-of-speech (POS) tagging. The pointwise approach refers, as features, only to the surface information of the input and not relies on any prediction results such as word boundaries or POS tags. This design allows us to use a variety of linguistic resources flexibly. This characteristic enables a fast and low-cost 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 and morpheme n-gram models. In addition, a domain adaptation experiment showed that the proposed method is able to achieve an effective domain adaptation with a smaller amount of annotations.

Key Words: Morphological analysis, Word segmentation, Part-of-speech tagging, Pointwise prediction, Sequence-based prediction

[†]京都大学 学術情報メディアセンター, Kyoto University, Academic Center for Computing and Media Studies
[††]京都大学情報学研究科, Kyoto University, School of Informatic

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などの後段の言語処理の入力となるばかりでなく、情報検索システムやテキストマイニング等の自然言語処理の応用の入力として直接利用される。そのため、形態素解析の精度は自然言語処理やその応用に大きな影響を与える。昨今、自然言語処理の応用は医療 (三浦，荒牧，大熊，外池，杉原，増市，大江 2010) や法律 (下畑，井佐原 2007) から Web 文書 (早藤，建石 2010) まで多岐に渡る。したがって、様々な分野のテキストに対して、高い形態素解析解析精度を短時間かつ低コストで実現する手法が望まれている。

現在の形態素解析器の主流は、コーパスに基づく方法である。この方法では、統計的なモデルを仮定し、そのパラメータをコーパスから推定する。代表的な手法は、品詞 n-gram モデル (永田 1999)、全ての品詞を語彙化した形態素 n-gram モデル (森，長尾 1998b)、条件付き確率場 (CRF)(工藤，山本，松本 2004) などを用いている。これらの統計的手法は、パラメータをコーパスから推定することで、際限なきコスト調整という規則に基づく方法の問題を解決し、コーパス作成の作業量に応じて精度が確実に向上するようになった。

一方、これらの既存の統計的手法による形態素解析器で、医療や法律などの学習コーパスに含まれない分野のテキストを解析すると実用に耐えない解析精度となる。この問題に対して、分野特有の単語を辞書に追加するという簡便な方法が採られるが、問題を軽減するに過ぎない。論文等で報告されている程度の精度を実現するには、解析対象の分野のフルアノテーションコーパスを準備しなければならない。すなわち、解析対象の分野のテキストを用意し、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する必要がある[1]。この結果、ある分野のテキストに自然言語処理を適用するのに要する時間は長くなり、コストは高くなる。

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実現すると同時に、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。

[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 (坪井，森，鹿島，小田，松本 2009) 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。

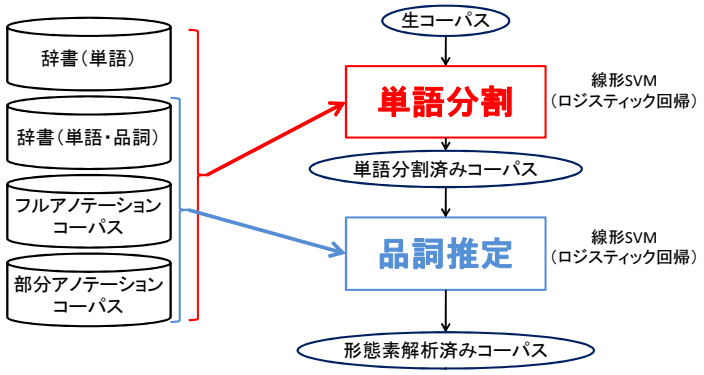

図 1　処理の流れ

x_{i-2} x_{i-1} x_i x_{i+1} x_{i+2} x_{i+3}
テキスト: 消　費　者　の　要　求　が　多　様　化
↑
t_i: 判定点

文字 (種)1-gram 素性: -3/消 (K), -2/費 (K), -1/者 (K), 1/の (H), 2/要 (K), 3/求 (K)
文字 (種)2-gram 素性: -3/消費 (KK), -2/費者 (KK), -1/者の (KH), 1/の要 (HK), 2/要求 (KK)
文字 (種)3-gram 素性: -3/消費者 (KKK), -2/費者の (KKH), -1/者の要 (KHK), 1/の要求 (HKK)

図 2　単語分割に使用する素性 (窓幅が 3、n-gram 長の上限が 3 の場合)

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 (Graham，中田，森 2010) がある。提案手法での単語分割にはこれを採用する。以下では、この点予測による単語分割を概説する。

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単語境界タグ $\boldsymbol{t} = t_1 t_2 \cdots t_{n-1}$ 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 種類である。したがって、単語境界タグの推定

x_{-3} x_{-2} x_{-1} w x_1 x_2 x_3
テキスト: 消 費 者 の 要求 が 多 様 化
↑
注目単語

文字 (種)1-gram 素性: -3/費 (K), -2/者 (K), -1/の (H), 1/が (H), 2/多 (K), 3/様 (K)
文字 (種)2-gram 素性: -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 素性: -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 (窓幅が 3、n-gram 長の上限が 3 の場合)

は、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線形サポートベクトルマシン (Linear SVM) (Fan, Chang, Hsieh, Wang, and Lin 2008) による分類を行っている。参照する素性は以下の通りである (図 2 参照)。

(1) 文字 n-gram: 判別するタグ位置 i の周辺の部分文字列である。窓幅 m と長さ n のパラメータがあり、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である。

(2) 文字種 n-gram: 文字を文字種に変換した記号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。提案手法による品詞推定の入力は単語列であるが、品詞推定対象の単語以外の単語境界情報を参照しない。この設計により、一部の単語にのみ単語境界や品詞情報が付与された部分的アノテーションコーパスが容易に利用可能となる。この点が、英語などの単語に分かち書きされた言語に対する品詞推定の既存手法 (DeRose 1988) との相違点である。

提案手法における品詞推定は、注目する単語に応じて以下の 4 つの種類の処理を行う。

(1) 学習コーパスに出現し、複数の品詞が付与されている単語は、後述する単語毎の分類器で品詞を推定する。

(2) 学習コーパスに出現し、唯一の品詞が付与されている単語には、その品詞を付与する。

(3) 学習コーパスに出現せず、辞書に出現する単語には、辞書に最初に現れる品詞を付与する。

(4) 学習コーパスに出現せず、辞書にも出現しない単語は、その品詞を名詞とする。

分類器で品詞を推定する (1) の場合は、点予測を用いることとする。点予測による品詞推定は、

- **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

品詞を推定する単語 w とその直前の文字列 $\boldsymbol{x}_-$ と直後の文字列 $\boldsymbol{x}_+$ 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される。参照する文字列の窓幅を m' とすると、入力において参照される文脈情報は $\boldsymbol{x}_-, w, \boldsymbol{x}_+ = x_{-m'} \cdots x_{-2} x_{-1}, w, x_1 x_2 \cdots x_{m'}$ 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、推定対象の単語以外の単語境界情報や周囲の単語の品詞などの推定結果を一切参照しない。この設計により、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。

分類器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

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 1 対多方式 (one-versus-rest) を用いて多値分類を行う。

なお、全単語に対して 1 つの多値分類器を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。

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。この設計により、様々な言語資源を柔軟に利用することが可能となる。

系列ラベリングとして定式化する既存手法による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。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は、フルアノテーションコーパスと同様の単語の基準を満たし、品詞が付与されている。既存手法の分野適応に際しては、対象分野の形態素解析や文字 n-gram の統計結果 (森，長尾 1998a) から、頻度が高いと推測される単語から辞書に追加される。しかしながら、文脈情報が欠落するのでコーパスほど有効ではない。

フルアノテーションコーパス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。このような人材の確保は困難であり、比較的短時間の訓練の後に実作業にあたるのが現実である。その結果、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かのいずれかしかない。

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、確信の持てる箇所で確信の持てる情報のみのアノテーションを許容する枠組みが渇望されている。提案する枠組みでは、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)。

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用意することができる。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。最も効果が高い分野適応の戦略は、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。

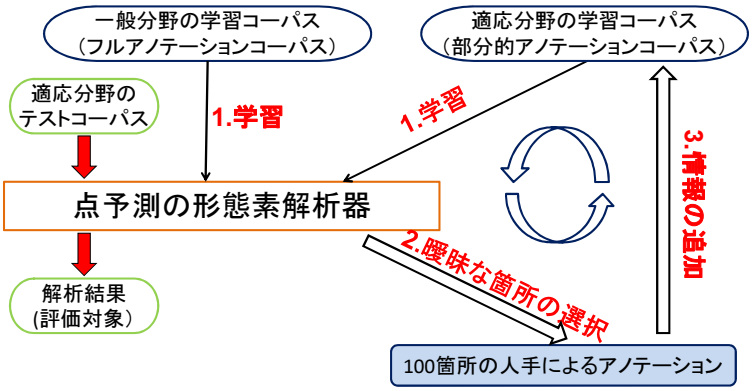

図 5　能動学習

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。手順は以下の通りである (図 5 参照)。

(1)　一般分野の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、後述する方法で推定の信頼度が低い 100 箇所を選択する[2]。

(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)　一般分野のフルアノテーションコーパスと適応分野の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。

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

[2]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。

表 1　コーパス

コーパス名	出典	用途	文数	形態素数	文字数
BCCWJ (日本語書き言葉均衡コーパス)	白書・書籍・新聞 (一般分野)	学習	27,338	782,584	1,131,317
		テスト	3,038	87,458	126,154
	Yahoo!知恵袋 (適応分野 1)	学習	5,800	114,265	158,000
		テスト	645	13,018	17,980
JAPIC	医薬品情報テキスト (適応分野 2)	テスト	1,236	45,303	67,828

報を付与する。

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 つ目の実験では、自然言語処理の適応対象を医薬品情報のテキストと想定し、言語資源が豊富な一般分野のコーパスで学習を行い、医薬品情報のテキストに対する形態素解析精度を既存手法と比較する。2 つ目の実験は、能動学習による提案手法の分野適応性の定量的評価である。比較的大きなコーパスが存在する分野のテキストを対象に、一部をテストコーパスとし、残りを能動学習を模擬するための学習コーパスとして利用し、アノテーション数と精度の関係を評価する。

3.1　コーパス

実験には「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) のコアデータ (以下 BCCWJ と呼ぶ)(前川 2009) と医薬品情報のテキスト (以下 JAPIC と呼ぶ) を用いた。これらのコーパスは、単語分割と品詞付与が人手で行われている。コーパスの諸元を表 1 に示す。また、219,583 形態素を収めた UniDic(伝，小木曽，小椋，山田，峯松，内元，小磯 2007) を辞書として用いた。

本論文で提案するのは、分野適応性の高い形態素解析器であり、1 つ目の実験では、一般分野と JAPIC(適応分野) をテストコーパスとする評価を行う。この実験では、コーパスと同じ基準の辞書がある場合とない場合も比較した。それぞれの場合のカバー率は表 2 の通りである。2 つ目の実験では、提案手法と既存手法の代表である CRF の能動学習を行ない、分野適応性を評価する。この実験でも JAPIC を適応分野とするのが理想的であるが、我々は能動学習の実験に必要なアノテーションを模擬する学習コーパスを有していない。したがって、性質に応じて BCCWJ を 2 つに分割し、能動学習の実験を行った。分割においては、文献 (Maekawa, Yamazaki, Maruyama, Yamaguchi, Ogura, Kashino, Ogiso, Koiso, and Den 2010) を参考に、他の出典のデータと大きく性質が異な

表 2　カバー率

言語資源	BCCWJ 一般	Yahoo!知恵袋	JAPIC
BCCWJ 一般	98.37%	96.29%	88.56%
+ UniDic	99.95%	99.80%	95.99%

る 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した。分野適応性の評価のための実験で、UniDic を利用することも考えられるが、表 2 から分かるように、これを語彙に加えた場合の Yahoo!知恵袋のカバー率は 99.80%と JAPIC を対象とした場合の 95.99%に比べて非常に高く、実際の分野適応を模擬していることにはならない。したがって、分野適応性の評価実験においては、UniDic を使用しないこととした。なお、この場合のカバー率は 96.29%であり、この判断はおおむね妥当である。

3.2　評価基準

本論文で用いた評価基準は、文献 (永田 1995) で用いられている再現率と適合率であり、次のように定義される。正解コーパスに含まれる形態素数を N_{REF}、解析結果に含まれる形態素数を N_{SYS}、単語分割と品詞の両方が一致した形態素数を N_{COR} とすると、再現率は N_{COR}/N_{REF} と定義され、適合率は N_{COR}/N_{SYS} と定義される。例として、コーパスの内容と解析結果が以下のような場合を考える。

コーパス

外交/名詞 政策/名詞 で/助動詞 は/助詞 な/形容詞 い/語尾

解析結果

外交政策/名詞 で/助詞 は/助詞 な/形容詞 い/語尾

この場合、分割と品詞の両方が一致した形態素は「は/助詞」と「な/形容詞」と「い/形容詞語尾」であるので、$N_{COR} = 3$ となる。また、コーパスには 6 つの形態素が含まれ、解析結果には 5 つの形態素が含まれているので、$N_{REF} = 6$, $N_{SYS} = 5$ である。よって、再現率は $N_{COR}/N_{REF} = 3/6$ となり、適合率は $N_{COR}/N_{SYS} = 3/5$ となる。また、再現率と適合率の調和平均である F 値も評価の対象とした。

3.3　各手法の詳細

提案手法においては、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' をすべて 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 (Fan et al. 2008) を用いた。

比較対象とした既存手法は、品詞 2-gram モデル (HMM) (永田 1999) と、形態素 n-gram モデル

表 3　一般分野に対する単語分割精度および形態素解析精度 (UniDic なし)

手法	単語分割			形態素解析		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
品詞 2-gram モデル (HMM)	96.32	96.84	96.58	93.77	94.27	94.02
形態素 2-gram モデル	97.44	98.52	97.98	96.58	97.65	97.11
形態素 3-gram モデル	97.49	98.53	98.00	96.70	97.73	97.21
CRF(MeCab-0.98)	97.19	98.30	97.74	96.72	97.84	97.28
提案手法 (KyTea-0.1.1)	98.73	98.71	98.72	98.07	98.06	98.06

表 4　JAPIC に対する単語分割精度および形態素解析精度 (UniDic なし)

手法	単語分割			形態素解析		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
品詞 2-gram モデル (HMM)	86.79	89.54	88.14	82.96	85.58	84.25
形態素 2-gram モデル	89.42	91.58	90.49	86.21	88.30	87.24
形態素 3-gram モデル	89.46	91.71	90.57	86.40	88.57	87.47
CRF(MeCab-0.98)	84.63	91.00	87.70	82.55	88.76	85.54
提案手法 (KyTea-0.1.1)	94.16	95.28	94.71	93.02	94.12	93.57

($n = 2, 3$)(森，長尾 1998b) と、CRF に基づく方法 (MeCab-0.98)(工藤他 2004) である。予備実験の結果、CRF に基づく方法において素性とする語彙は、学習コーパスに出現する全単語のうちの低頻度語 500 語以外とした。また、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、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、品詞と表記を用い、右文脈 2-gram と左文脈 2-gram には、品詞 2-gram と語彙化された単語を用いた。

3.4　既存手法との比較

まず、一定量の言語資源がある状況での精度を既存手法と比較した。表 3 と表 4 は、各手法において学習コーパスのみを用いる場合の一般分野と適応分野のテストコーパスに対する精度である。また、表 5 と表 6 は、言語資源として辞書も用いる場合の結果である。

まず、全体の傾向としては、多くの場合に表の上から順に精度が良くなっていく。品詞 2-gram モデルと形態素 2-gram モデルと形態素 3-gram モデルの精度は、いずれの場合もこの順に向上する。これは、文献 (森，長尾 1998b) に報告されている通りである。唯一の例外は、JAPIC に対す

表 5　一般分野に対する単語分割精度および形態素解析精度 (UniDic あり)

手法	単語分割			形態素解析		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
品詞 2-gram モデル (HMM)	98.18	97.74	97.96	95.76	95.32	95.54
形態素 2-gram モデル	99.21	99.42	99.32	98.52	98.73	98.63
形態素 3-gram モデル	99.23	99.42	99.33	98.59	98.78	98.68
CRF(MeCab-0.98)	99.60	99.55	99.57	99.25	99.21	99.23
提案手法 (KyTea-0.1.1)	99.42	99.33	99.37	98.91	98.82	98.86

表 6　JAPIC に対する単語分割精度および形態素解析精度 (UniDic あり)

手法	単語分割			形態素解析		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
品詞 2-gram モデル (HMM)	91.97	93.79	92.87	88.61	90.37	89.48
形態素 2-gram モデル	93.29	95.78	94.52	90.17	92.58	91.36
形態素 3-gram モデル	93.27	95.72	94.48	90.42	92.79	91.59
CRF(MeCab-0.98)	93.25	96.17	94.69	91.54	94.40	92.94
提案手法 (KyTea-0.1.1)	93.96	96.41	95.17	92.50	94.92	93.70

る単語分割精度である。これは、過学習が原因であると考えられる。

次に、CRF に基づく方法と品詞 2-gram モデルとの比較である。ある程度大きな辞書が利用可能でカバー率が高いという条件下では CRF に基づく方法は品詞 2-gram モデルより精度が高いことがわかる。これは、文献 (工藤他 2004) に述べられている結果と同じである。しかしながら、利用可能な辞書がなくカバー率が低い場合には、学習コーパスと異なる分野のテキストに対してほぼ同じ形態素解析精度になっている。この原因は、CRF に基づく方法の未知語処理が不十分で[3]、単語分割精度が著しく低いことである。

形態素 n-gram モデルは、いずれの条件でも品詞 2-gram モデルよりも高い精度となっている。これは、文献 (森，長尾 1998b) の結果を追認し、品詞列のみならず、表記列の情報をモデル化することの重要性を強く示唆する。形態素 n-gram モデルと CRF に基づく方法との比較では、単語分

[3]CRF に基づく方法を提案している文献 (工藤他 2004) には、「もし, 辞書にマッチする単語が存在せず, ラティスの構築に失敗した場合は, 別の未知語処理が起動される.」と記述されており、既知語列に分解できない場合にのみ文字種に対するヒューリスティクスに基づく未知語処理が起動されると考えられる。この結果、例えば「投与/名詞」を「投/動詞 与/接頭辞」と誤って解析することが頻繁に起こっている。これは MeCab-0.98 に固有の問題で、CRF に基づく方法一般の問題ではないかもしれない。しかしながら、我々の知る限り、適切な未知語モデルも含めた CRF に基づくモデルを提案し、その評価について日本語を対象として報告している論文ははない。

割においては形態素 n-gram モデルが CRF を用いる方法よりも優れているが、品詞の一致も評価に含めた場合、CRF に基づく方法がより優れている。唯一の例外は、カバー率が最も低い表 4 の場合で、CRF に基づく方法の単語分割精度が低すぎて、形態素解析精度においても形態素 n-gram モデルよりも低い精度となっている。

最後に、本論文で提案する点予測に基づく方法と既存手法の比較についてである。品詞 2-gram モデルや形態素 n-gram モデルとの比較においては、唯一の例外 (表 5 の単語分割の再現率) を除いて、提案手法が高い精度となっている。CRF に基づく方法との比較では、辞書を用いて学習コーパスと同一の分野のテストコーパスを解析対象とする表 5 の場合を除いて、提案手法が高い精度となっている。現実的な応用を想定した JAPIC を対象とする場合 (表 4 と表 6 参照) において、提案手法がいずれの既存手法よりも高い精度となっている点は注目に値する。特筆すべきは、コーパスと同じ基準で作成された辞書がない表 4 の場合に、提案手法が他の手法と比べて圧倒的に高い精度となっている点である。

以上の結果から、点予測に基づく方法は、ある単語分割および品詞付与の基準に基づく言語資源作成の初期や、同じ分野の学習コーパスの存在が望めない実際の言語処理において非常に有効であることがわかる。

3.5 分野適応性の評価

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コーパスの作成手順は 2.4 項の通りである。なお、前述の通り、カバー率の観点から初期の言語資源として一般分野の学習コーパスのみを用い、適応分野を Yahoo!知恵袋とする。

Pointwise:part 適応分野の部分的アノテーションコーパスから構築した提案手法: 一般分野コーパスで学習を行い、適応分野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 になるまで繰り返す。

Pointwise:full 適応分野のフルアノテーションコーパスから構築した提案手法: 適応分野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なう。

CRF:part 適応分野の部分的アノテーションコーパスから構築した CRF に基づく方法: 上述の **Pointwise:part** で得られる部分的アノテーションコーパスに含まれる単語を CRF に基づく方法の語彙として追加する。

CRF:full 適応分野のフルアノテーションコーパスから構築した CRF に基づく方法: 上述の

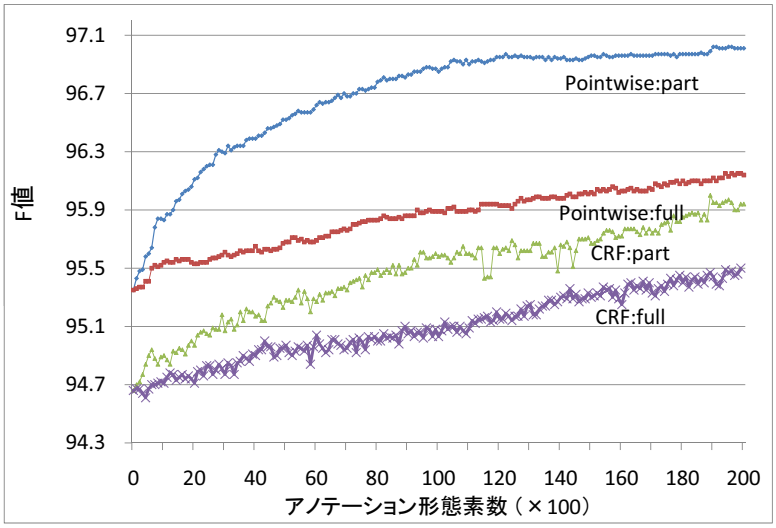

図 6　形態素解析精度と適応分野のアノテーション形態素数の関係

Pointwise:full でフルアノテーションした文に出現する単語を CRF に基づく方法の語彙として追加し、さらにそれらの文を学習コーパスに追加する。

以上のそれぞれで学習したモデルで適応分野の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 6 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、フルアノテーションによる解析精度向上に対する効果は、いずれの手法においてもほぼ同じであることがわかる。最後に、部分的アノテーションによる解析精度向上に対する効果は、提案手法においてより大きいことがわかる。このことから、点予測による形態素解析手法と部分的アノテーションによる能動学習は、非常に良い組み合わせであり、本論文の提案により既存手法に比べて高い分野適応性が実現できることが分かる。このことは、ある分野のテキストに対して言語処理がどの程度有効かを迅速に示す必要があるようなプロジェクトの初期や、形態素解析がプロジェクトの一部に過ぎず、投資額が限られるような実際の言語処理において非常に大きな意味を持つ。

4　おわりに

本論文では、点予測による形態素解析手法を提案した。言語資源が豊富な一般分野のコーパスで学習を行い、一般分野と適応分野において提案手法と既存手法の解析精度の比較を行った。その結果、提案手法を用いた形態素解析は、実際の言語処理において非常に有効であることが示された。

さらに、部分的アノテーションを用いる能動学習と提案手法を組み合わせることで、既存手法と比較して高い分野適応性が実現できることが示された。

謝辞

本研究の一部は、科学研究費補助金・若手 A(課題番号: 08090047) により行われました。

参考文献

DeRose, S. J. (1988). “Grammatical Category Disambiguation by Statistical Optimization.” *Computational Linguistics*, **14** (1), pp. 31–39.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). “LIBLINEAR: A Library for Large Linear Classification.” *Journal of Machine Learning Research*, **9**, pp. 1871–1874.

Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., and Den, Y. (2010). “Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Graham Neubig，中田陽介，森信介 (2010). “点推定と能動学習を用いた自動単語分割器の分野適応.” 言語処理学会年次大会.

下畑さより，井佐原均 (2007). “日英特許コーパスからの専門用語対訳辞書の自動獲得.” 自然言語処理, **14** (4), pp. 23–42.

前川喜久雄 (2009). “代表性を有する大規模日本語書き言葉コーパスの構築.” 人工知能学会誌, **24** (5), pp. 616–622.

早藤健，建石由佳 (2010). “２ちゃんねる解析用の形態素解析器の作成.” 言語処理学会年次大会.

三浦康秀，荒牧英治，大熊智子，外池昌嗣，杉原大悟，増市博，大江和彦 (2010). “電子カルテからの副作用関係の自動抽出.” 言語処理学会年次大会.

伝康晴，小木曽智信，小椋秀樹，山田篤，峯松信明，内元清貴，小磯花絵 (2007). “コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用.” 日本語科学, **22**, pp. 101–122.

永田昌明 (1995). “EDR コーパスを用いた確率的日本語形態素解析.” EDR 電子化辞書利用シンポジウム, pp. 49–56.

永田昌明 (1999). “統計的言語モデルと N-best 探索を用いた日本語形態素解析法.” 情報処理学会論文誌, **40** (9), pp. 3420–3431.

森信介，長尾眞 (1998a). “n グラム統計によるコーパスからの未知語抽出.” 情報処理学会論文誌,

39 (7), pp. 2093–2100.

森信介，長尾眞 (1998b). “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, **5** (2), pp. 75–103.

工藤拓，山本薫，松本裕治 (2004). “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, NL161 巻.

坪井祐太，森信介，鹿島久嗣，小田裕樹，松本裕治 (2009). “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, **50** (6), pp. 1622–1635.

略歴

森 信介：1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了. 同年日本アイ・ビー・エム (株) 入社. 2007 年より京都大学学術情報メディアセンター准教授. 京都大学博士 (工学). 1997 年情報処理学会山下記念研究賞受賞. 2010 年情報処理学会論文賞受賞. 情報処理学会会員.

中田 陽介：2009 年香川大学工学部信頼性情報システム工学科卒業. 2011 年京都大学大学院情報学研究科修士課程修了. 同年エヌ・ティ・ティ・コミュニケーションズ (株) 入社.

NEUBIG Graham： 2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業. 2010 年京都大学大学院情報学研究科修士課程修了. 同年同大学院博士後期課程に進学. 現在に至る. 自然言語処理に関する研究に従事.

河原 達也：1987 年京都大学工学部情報工学科卒業. 1989 年同大学院修士課程修了. 1990 年博士後期課程退学. 同年京都大学工学部助手. 1995 年同助教授. 1998 年同大学情報学研究科助教授. 2003 年同大学学術情報メディアセンター教授. 現在に至る. 音声言語処理, 特に音声認識及び対話システムに関する研究に従事. 京都大学博士（工学）. 1997 年度日本音響学会粟屋潔学術奨励賞受賞. 2000 年度情報処理学会坂井記念特別賞受賞. 2004 年から 2008 年まで言語処理学会理事. 日本音響学会, 情報処理学会 各代議員. 電子情報通信学会, 人工知能学会, 言語処理学会, IEEE 各会員.