Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

点予測と系列予測の 2 段階化による品詞推定の精度向上

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

本論文では、点予測による形態素解析の推定結果に対して、品詞連接の傾向を用いた系列予測による品詞のリランキング手法を提案する。点予測とは、分類器の素性として対象とその周辺の文字列情報のみを用いる手法であり、この手法により高い分野適応性を実現している。しかし、点予測では品詞推定に有用な品詞連接の傾向を利用することができない。品詞連接の傾向は分野依存性が低いと考えられ、異なる分野で学習した品詞連接の傾向を利用できると考えられる。この品詞連接の傾向を用い、点予測の品詞推定結果に対してリランキングすることにより解析精度の向上を実現する。

Improving Part-of-Speech Tagging by Combining Pointwise and Sequence-based Predictors

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

This paper proposes an approach to part-of-speech sequence reranking based on POS transition tendencies fot the result of morphological analysis 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. But pointwise prediction cannot use POS transition information that is important in POS prediction. It can be assumed that the transition tendencies of POSs are not highly domain dependent, transition information learned in one domain can be used in another domain. By applying POS sequence reranking that considers POS transition information to the result of pointwise predictors, we were able to achieve an improvement in POS tagging accuracy.

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、様々な分野で自然言語処理が用いられる近年、非常に重要な要素技術である。形態素解析は文字列に単語境界と品詞を付与する処理である。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、形態素解析には多種多様な分野のテキストに対して高い解析精度が求められる。

そこで、分野適応性の高い形態素解析手法として、点予測による形態素解析[1] が提案されている。この手法では、形態素解析を単語境界推定と品詞推定に分けて段階的に行い、各処理において点予測を用いる。点予測とは、系列予測の対義語であり、入力情報（文字列情報）のみを参照する手法である。つまりは、推定値（単語境界や品詞）を一切利用しない手法である。単語境界推定を行う場合で言えば、前後の文字列のみを参照し、すでに推定されている前方の単語境界を参照しない。同様に、品詞推定では、単語分割済みコーパスが入力であるが、品詞推定の際に、対象単語の単語境界と周辺の文字列のみを参照し、他の単語境界情報及び品詞情報を参照しない。そのように設計することで、既存手法では利用が困難であった部分的に品詞や単語境界が付与されたコーパス（部分的アノテーションコーパス）や単語表記のみの単語辞書（複合語辞書）等を利用することができる。このように言語資源を有効活用することで、少ない人的コストで効率のよい分野適応を実現している。

しかし、点予測では品詞推定において重要な情報源である品詞連接の傾向を利用することができない。品詞連接の傾向を利用することは、品詞解析精度の向上につながる。また、この品詞連接の傾向は分野依存性が低く、異なる分野で学習した品詞連接の傾向が利用可能であると考えられる。

このような背景のもと、本論文では点予測と系列予測の 2 段階化による品詞推定の精度向上手法を提案する。まず、点予測による形態素解析を行う。次に点予測で利用できない品詞連接の傾向を用い、系列予測でリランキング[⋆1]を行う。また、あらかじめ点予測の際に信頼度付きの品詞推定を行い、系列予測の素性として用いる。以上のような、点予測と系列予測による品詞推定の 2 段階化を提案し、言語資源を有効活用することで高い分野適応性を保ちながら、より高い解析精度を実現する。

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

⋆1 リランキングの先行研究は多数ある[2),3)]。しかし、品詞のリランキングは殆ど行われていない。

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

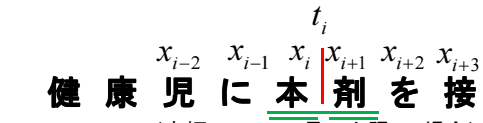

図 1 単語分割に使用する素性

2. 点予測による形態素解析

本研究では、系列予測による品詞推定の前処理として、点予測による形態素解析[1)] を用いる。点予測による形態素解析では、単語境界推定と品詞推定に分けて段階的に処理される。各処理において、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ている。なお、分類器には、精度と学習効率を考慮して線形 SVM[4)] を用いている。

2.1 点予測による単語境界推定

点予測による単語境界推定[5)] 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。点予測による単語境界推定では、以下の 3 種類の素性を参照する。

（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 1 参照）。

（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。

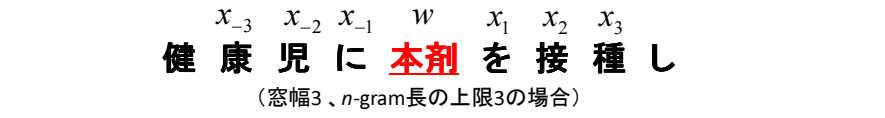

図 2 品詞推定に使用する素性

2.2 点予測による品詞推定

点予測による品詞推定では推定対象の単語によって、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、辞書の登録順で品詞を付与する。

（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。

（1）の場合は各単語の品詞候補毎の分類器を作り、one v.s. rest 法を用いて多値分類を行う。入力は単語列であるが、推定する際には対象単語 w とその前の文脈の文字列 $\boldsymbol{x}_-$ と後の文脈の文字列 $\boldsymbol{x}_+$ とみなし、これらのみを参照して単語 w の品詞を推定する。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 $x_{-m'} \cdots x_{-2} x_{-1}, w, x_1 x_2 \cdots x_{m'}$ 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。

品詞推定に利用する素性は以下の通りである（図 2 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

2.3 点予測による柔軟な言語資源利用

点予測を用いることで新たに利用可能となる言語資源があり、それらの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している。

（1） 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

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

語境界情報のみが付与された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。

(2) 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。

無論、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フルアノテーションコーパス、見出し語に品詞が付与されている形態素辞書も利用可能である。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。点予測による形態素解析では、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している。

3. 点予測と系列予測による品詞推定の 2 段階化

前節の点予測では、学習コーパスの品詞列の情報を利用することができなかったが、この情報は品詞推定において非常に重要である。本研究では、品詞連接の傾向は分野依存性が低いと仮定し、本節では以上に述べた仮定の下で、異なる分野から学習した品詞連接の傾向を利用して、点予測の形態素解析の結果得られる品詞列に対して、系列予測による品詞のリランキング手法を提案する。

3.1 提案手法の流れ

提案手法の全体の流れについて説明する。提案手法を含む形態素解析は「点予測による単語境界推定」、「点予測による品詞推定」、「系列予測による品詞のリランキング」からなる。まず文字列を入力として、点予測による単語境界推定を行う。次に単語列を入力として、点予測による品詞推定を行う。この際品詞推定は各単語に対して品詞候補とそれぞれの信頼度を出力する。そして最後に、点予測の信頼度付き品詞推定結果である品詞列に対して、系列予測による品詞のリランキングを行う。

3.2 点予測による信頼度付きの品詞推定

2.2 項で述べた点予測による品詞推定では、各単語に対して唯一の品詞を出力する。提案手法では、出力を可能なすべての品詞とそれぞれの信頼度とし、系列予測の素性として用いる（図 3 参照）。

点予測による品詞を信頼度は以下のように定義する。まず r（$r \geq 1$）番目の品詞候補の分離平面からの距離（マージン）を d_r とする。その上で r 番目の品詞候補の信頼度を $C_r = d_r - d_2$ とする。この結果、第 1 候補の信頼度のみが正の値（L2 正則化によりほぼ全てが 100 より十分に小さい値となる）となり、第 2 候補の信頼度は 0、第 3 候補以降の信頼度は負の値となる。品詞候補がない場合（2.2 項における 4 の場合）は名詞が付与され信頼度を 0 とする。品詞候補が 1 つの場合（2.2 項の 2、3 の場合）は信頼度を 100（特別な値）とする。品詞候補と品詞の信頼度の例を図 3 に示す。

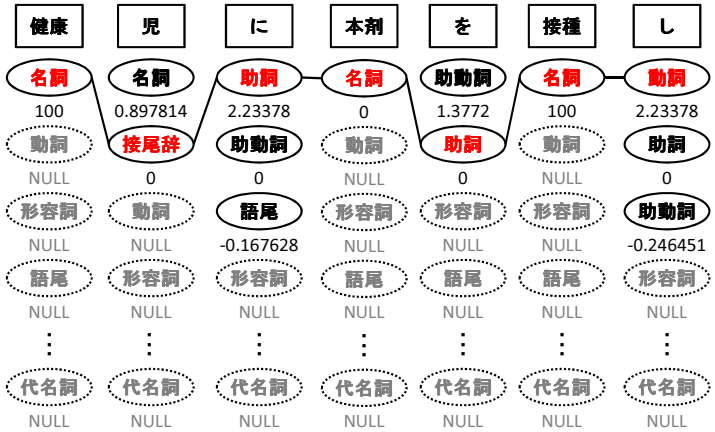

図 3 系列予測による品詞のリランキング

3.3 系列予測による品詞のリランキング

前項で説明した処理の結果、単語列とその各単語の可能なすべての品詞及びその信頼度が付与された文が得られる。それに含まれる全ての品詞系列から最適と考えられる品詞系列を、点予測の推定結果と学習コーパスにおける品詞連接の情報を用いて探索する。なお、単語境界のリランキングは行わないため、点予測による単語境界は変らない。

系列予測には、柔軟に素性を設計でき、文全体で最尤の品詞列を推定できる CRF[6] を用いる。CRF の学習時の入力は、単語境界と品詞がフルアノテーションされたコーパスから得ることができる品詞列である。素性は、点予測の結果得られる「文脈情報素性」と「信頼度情報素性」である（詳細は次項で述べる）。解析時には、点予測の推定結果と信頼度を素性として、文全体で最尤の品詞列を出力する。図 3 の例では、実線で結ばれた品詞が出力され、単語「児」の品詞が「名詞」から「接尾辞」に適切に変更されている。

3.4 系列予測の素性

前項で述べた CRF の素性は「信頼度素性」と「文脈情報素性」である。「信頼度素性」

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

は、点予測による推定結果である品詞の信頼度をもとに、以下の素性生成規則に従い作成される 3 種類の素性である。つまり T 種類の品詞がある場合、素性生成規則毎に T 個、合計 $3T$ 個の素性を作成する。

規則 1（素性 1〜素性 T）： 単語の品詞候補が複数ある場合は、品詞の信頼度を素性とする。

規則 2（素性 T+1〜素性 2T）： 品詞が単語の品詞候補ではない場合、素性を 1 とする。

規則 3（素性 2T+1〜素性 3T）： 単語の品詞候補が 1 つの場合、素性を 1 とする。

規則に当てはまらない場合、素性は NULL とする。また、規則 1 で信頼度が存在する品詞に関しての信頼度を与える。規則 2 では、品詞候補ではない品詞の情報を与えている。規則 3 では、点予測での信頼度が高いもの（修正が不要と推定される品詞）の情報を与えている。

もう 1 つは「文脈情報素性」である。これを以下に示す。

(1) 対象単語を含む窓幅 m'' に含まれる単語 n-gram を素性とする。

(2) 対象単語を含む窓幅 m'' に含まれる単語に対応する文字種集合 n-gram を素性とする。単語の文字種集合は、その単語の表記に含まれる文字種を要素とする集合である。文字種は 2.1 項と同様の 6 種であり、文字種集合は $2^6 - 1$ 通りとなる。文字種集合 n-gram は、連続する n 個の文字種集合の列である。

3.5 学習コーパスの作成方法

CRF の学習コーパスとして、未知の単語列に対する点予測による品詞推定結果と正解の組が必要である。つまりは、信頼度付与を行う点予測の品詞推定器の学習コーパスと、信頼度付与対象は異なる必要がある。そこで、単語境界と品詞のフルアノテーションコーパスから、以下の手順で作成する（図 4 参照）。

(1) 学習コーパスを k 個に分割し、$C_1, C_2, ..., C_k$ を得る。

(2) ある i に対して C_i 以外の $k-1$ 個のコーパスから点予測による形態素解析器を学習し、C_i に対して単語境界推定及び信頼度付き品詞推定を行う。これをすべての $i \in 1, 2, ..., k$ に対して行う。

以上の結果、品詞候補と信頼度が付与された $C'_1, C'_2, ..., C'_k$ が得られる。これらをコーパスに正解の品詞を付与したものを CRF の学習コーパスとする。こうして得られる学習コーパスから文脈情報に応じた品詞連接の傾向を得ることができる。

3.6 提案手法と言語資源

最後に提案手法とコーパスの関係について述べる（図 5 参照）。一般分野に関しては様々

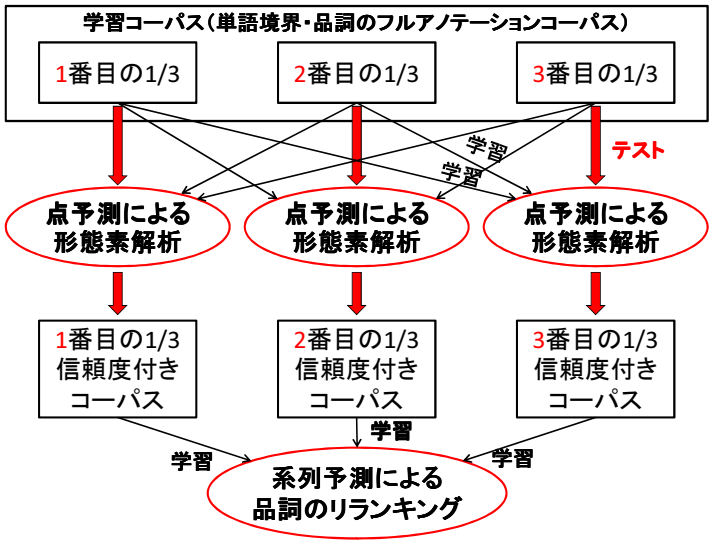

図 4 系列予測による品詞推定の学習コーパス作成方法（$k = 3$）

な単語境界と品詞の基準に対してフルアノテーションコーパス（GTF）[7),8)] がある。適応分野のコーパスは一般分野と異なり、単語境界と品詞の基準だけでなく、専門分野の知識が必要となり一般分野のコーパスと比べ、作成により多くの人的コストがかかる。その中でも、フルアノテーションと比べて部分的アノテーションは比較的簡単に用意できる。コーパスの例を図 6 に示す。図 6 では拡張 3 値表現を用いている。拡張 3 値表現は単語境界の情報を示す 3 値表現[9)] を拡張し、品詞情報を表現出来るようにした手法である。3 値表現では以下の 3 値が定義されている。

| ：単語境界がある。

\- ：単語境界がない。

␣ ：単語境界の有無は不明である。

新たに各単語の品詞情報表現を加え、「|単-語/品詞|」と表している。また、これらのコーパスから学習できる情報を以下に示す。

- GWF、AWF:「単語列」や「単語境界と前後の文字列」が学習できる。

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

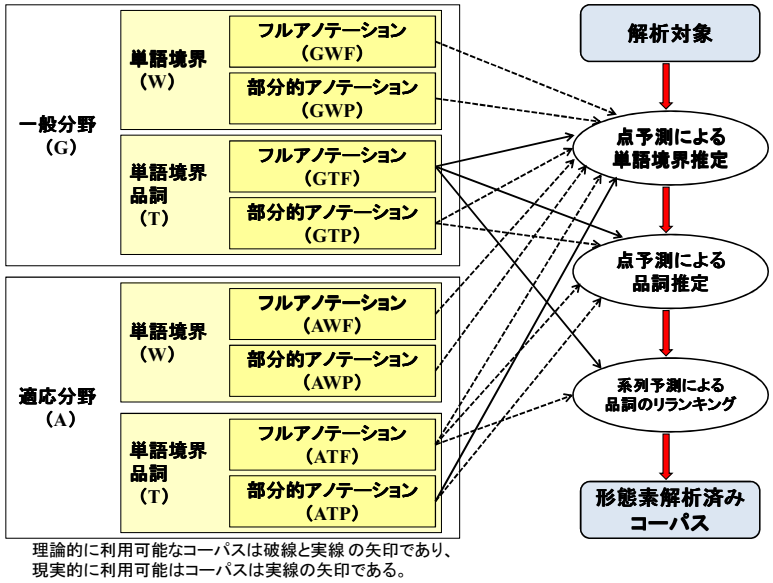

図 5 提案手法とコーパスの関係

- GWP、AWP:「単語境界と前後の文字列」が学習できる。
- GTF、ATF:「単語列」、「単語境界と前後の文字列」、「品詞列」、「形態素列」や「形態素と前後の文字列」が学習できる。
- GTP、ATP:「単語境界と前後の文字列」や「形態素と前後の文字列」が学習できる。

理論上では点予測のよる単語境界推定では、「単語境界と前後の文字列」が得られるコーパスであれば利用可能なため、全コーパスから学習可能である。点予測による品詞推定では、「形態素（単語と品詞）と前後の文字列」が得られるコーパスであれば利用可能なため、品詞が付与されている GWP、GWF、AWP、AWF のコーパスから学習可能である。系列予測による品詞推定では、「品詞列」の情報を学習するため、一文に単語境界と品詞が付与されているコーパスのみが利用可能であり、GTF、ATF のコーパスから学習が可能である。

実際の分野適応では、学習コーパスは、GTF に加えて比較的作成が容易な適応分野の部分的アノテーションコーパス（AWP または ATP）である。適応分野のフルアノテーションコーパス（AWF や ATF）は、準備するのに非常に高い人的コストがかかり現実的ではない。現実的な状況で利用可能なコーパスは図 5 の実線の矢印で示したコーパスであり、提案する形態素解析の枠組みでは、その言語資源を最大限活用している。

一般分野（G）	
単語境界（W）	フル（F）： |文-化|交-流|使|事-業|を|
	部分的（P）： ␣文␣化|交-流|使␣事␣業␣を␣
単語境界/品詞（T）	フル（F）： |文-化/名詞|交-流/名詞|使/接尾辞|事-業/名詞|を/助詞|
	部分的（P）： ␣文␣化|交-流/名詞|使␣事␣業␣を␣
適応分野（A）	
単語境界（W）	フル（F）： |血|小-板|の|減-少|が|
	部分的（P）： ␣血|小-板|の␣減␣少␣が␣
単語境界/品詞（T）	フル（F）： |血/接頭辞|小-板/名詞|の/助詞|減-少/名詞|が/助詞|
	部分的（P）： ␣血|小-板/名詞|の␣減␣少␣が␣

図 6 コーパス例

4. 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、既存手法との比較による提案手法の評価、もう 1 つは、点予測による分野適応時における提案手法の評価である。なお、予備実験により n-gram 長の n の上限を 2、窓幅 m'' はすべて 5 とした。また、系列予測に用いる学習コーパスの分割数は 9 とした。系列予測には CRFsuite[10] を用いた。

4.1 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）[8]⋆1を用いた。コーパスは単語境界と品詞情報が人手で付与されている。品詞は、大分類の 21 種類のみ利用した。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[11] 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。

4.2 評価基準

本論文で用いた評価基準は、文献 12) で用いられた再現率と適合率であり、次のように定

⋆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。

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

表 1 実験に用いるコーパスの詳細

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
	（適応分野）	テスト	645	13,018	17,980

義される。BCCWJ 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}、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/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。また本論文で用いる F 値は（2 × 適合率 × 再現率）/（適合率 + 再現率）とする。

4.3 評価実験 1 —既存手法との比較—

既存手法との比較による提案手法の解析精度の評価を行った。比較対象である既存手法は文献 1) と同様の CRF（MeCab-0.98）[13] と、形態素 n-gram モデル（n=2,3）[14]、品詞 2-gram モデル（HMM）[15]、さらに点予測である。この実験では、比較を行う既存手法と同じ言語資源を利用するために、図 4 で示した一般分野のフルアノテーションコーパス（図 5 の GTF）のみが利用可能であるとの条件を設定した。

提案手法で用いる CRF の学習には、一般分野の学習コーパスから作成手順（3.5 項）に従い作成したコーパスを用いた。解析対象は一般分野のテストコーパスと、適応分野のテストコーパスである。まず、解析対象に対して点予測による形態素解析を行った。さらにの点予測の結果に対して系列予測による品詞のリランキングを行なった。

一般分野の結果を表 2 に、適応分野の結果を表 3 に示す。表の点予測とは図 5 の点予測による品詞推定の結果であり、提案手法は系列予測による品詞のリランキングの結果である。

表 2 一般分野に対する単語境界推定精度および形態素解析精度

手法	単語境界推定			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
点予測（KyTea-0.1.1）	98.73	98.71	98.72	98.07	98.06	98.06
提案手法	98.73	98.71	98.72	98.38	98.37	98.38

表 3 適応分野（Yahoo!知恵袋）に対する単語境界推定精度および形態素解析精度

手法	単語境界推定			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）	93.17	94.44	93.80	86.78	87.96	87.36
形態素 2-gram モデル	94.52	96.65	95.57	92.01	94.09	93.04
形態素 3-gram モデル	94.52	96.71	95.60	92.10	94.24	93.16
CRF（MeCab-0.98）	94.89	96.87	95.87	93.69	95.65	94.66
点予測（KyTea-0.1.1）	96.93	97.26	97.09	95.19	95.51	95.35
提案手法	96.93	97.26	97.09	95.86	96.18	96.02

提案手法は両分野で解析精度が向上した。特に適応分野で一般分野より大幅に解析精度が向上してることが分かる。このことより、3 節で述べた「品詞連接の傾向は分野依存性が低い」という仮定に妥当性があることが分かる。以上の結果より、提案手法の有用性が示された。

4.4 評価実験 2 —分野適応時における提案手法の評価—

点予測による形態素解析の分野適応時における提案手法の評価を行った。先行研究[1] で最も高い分野適応性を示した手法として部分的アノテーション手法（Pointwise:part）がある。部分的アノテーション手法は、能動学習に部分的アノテーションコーパスを追加していく手法である。本実験では、部分的アノテーション手法時の結果に対して提案手法を適用し解析精度の評価を行った。本実験でも CRF の学習には前項と同様のものを用いる。分野適応の際にも同じ学習コーパスを用いる。

具体的な手順を以下に示す（図 7 参照）。

(1) 一般分野の学習コーパス（図 5 の GTF）を用い、点予測による形態素解析の学習を行う。
(2) 適応分野の学習コーパスに対して、信頼度付きの推定を行う。
(3) 適応分野の学習コーパスの信頼度が低い 100 箇所（単語境界・品詞推定を同時に考

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

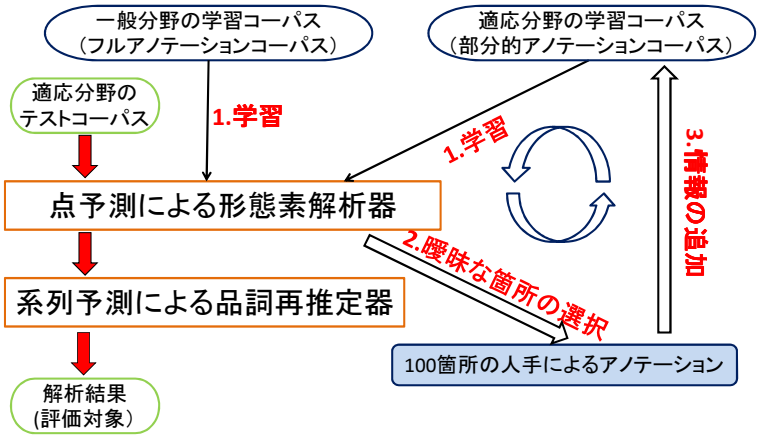

図 7 部分的アノテーションを用いた能動学習による分野適応

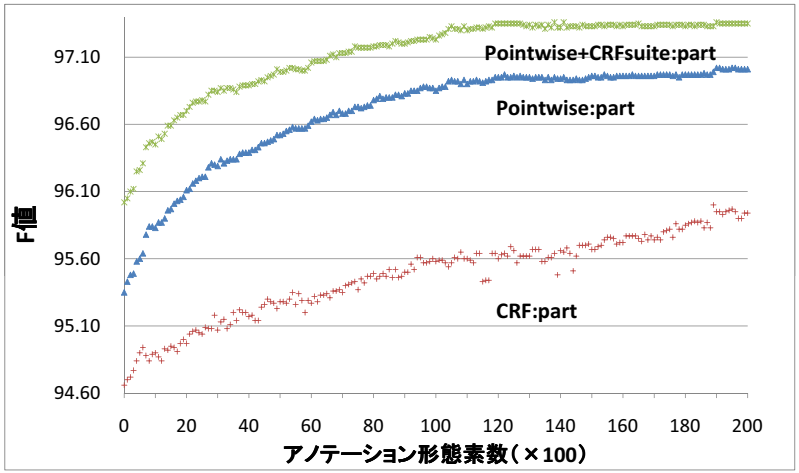

図 8 分野適応時における系列予測の形態素解析精度

慮）にアノテーションを行う（適応分野の部分的アノテーションコーパス（図 5 の ATP）を作成）。適応分野の部分的アノテーションコーパスを学習コーパスとして追加する。

この手順（1）～（3）を 200 回繰り返し、各回での信頼度付き適応分野のテストコーパスを用い、適応分野の CRF コーパスを作成し提案手法を行い、その適応分野のテストコーパスに対する解析精度を比較した。ベースラインとして点予測のみの結果と、CRF（MeCab-0.98）における、部分的アノテーションコーパス手法で作成した部分的アノテーションコーパスの語彙を追加した場合（CRF：part）を比較した。その結果を図 8 に示す。

図 8 から常にほぼ一定の解析精度の向上が確認できた。提案手法は分野適応時においても、解析精度が向上することが分かる。つまり、提案手法は点予測による形態素解析での分野適応性を保ちながら、解析精度の向上を実現している。このことより提案手法の有用性が確認できた。

5. おわりに

本論文では点予測と系列予測による品詞推定の 2 段階化手法を提案した。提案手法では、言語資源を有効活用することにより、点予測の高い分野適応性を保ちながら、品詞連接の傾向と点予測の推定結果を用いた系列予測による品詞のリランキングにより、解析精度向上を実現した。

参 考 文 献

1) 中田陽介，Neubig, G.，　森信介，河原達也：点予測による形態素解析，第 198 回自然言語研究会 (NL198)，東京 (2010).
2) 大庭隆伸，　堀貴明，　中村篤：ラウンドロビンデュエル識別法の提案と誤り訂正言語モデルによる評価，音響学会講演論文集，pp.29–30 (2010).
3) 　越川満，内山将夫，梅谷俊治，松井知己，山本幹雄：統計的機械翻訳におけるフレーズ対応最適化を利用した N-best 翻訳候補のリランキング，情報処理学会論文誌，Vol.51, No.8, pp.1443–1451 (2010).
4) 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).
5) 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).
6) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the Eighteenth ICML* (2001).

Vol.2011-IFAT-101 No.2
Vol.2011-NL-200 No.2
2011/1/28

7) 黒橋禎夫：京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会発表論文集，pp.115–118 (1997).
8) 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).
9) Mori, S. and Oda, H.: Automatic Word Segmentation using Three Types of Dictionaries, *Proceedings of the Eighth International Conference Pacific Association for Computational Linguistics* (2009).
10) Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007).
11) 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).
12) 永田昌明：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 (1995).
13) 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).
14) 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).
15) 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).