

条件付確率場による日本語未知語処理

東 藍 浅原 正幸 松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{ai-a,masayu-a,matsu}@is.naist.jp

本稿では、日本語形態素解析において問題となる未知語処理に対して条件付確率場 (Conditional Random Fields, CRF) を適用する手法を提案する。提案手法では、形態素解析と同時に入力文中の部分文字列に対して未知語候補を追加することにより、形態素解析と未知語処理を同時に行う。また、従来最大エントロピーマルコフモデル (Maximum Entropy Markov Model, MEMM) などを用いた手法で指摘されていた label bias あるいは length bias の影響は、単に既知語の解析において問題になるだけではなく、未知語処理においても重要な問題となることを示し、CRF を適用することによりこれらの問題が解決されることを示す。そして大規模な正解タグ付コーパスを用いて実験し、本稿の提案手法の有効性を検証した。

Japanese Unknown Word Processing using Conditional Random Fields

Ai Azuma Masayuki Asahara Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
{ai-a,masayu-a,matsu}@is.naist.jp

This paper proposes a new method for Japanese morphological analysis with unknown word (i.e. out-of-vocabulary word) processing. The Japanese morphological analysis is based on conditional random fields (CRF) on a word trellis. In the word trellis, the analyzer expands not only known words (i.e. in-vocabulary word) but also substrings in a sentence as word candidates. Kudo (Kudo 2004) discussed an issue that maximum entropy Markov model (MEMM) has label as well as length bias problems in known word processing and CRFs have potential to cope with them. We discuss the same issue in unknown word processing. Evaluation experiments on large-scale corpora show the effectiveness and impact on the proposed method.

1. はじめに

日本語や中国語で書かれた文では、単語の境界に明示的な空白などが置かれることはない。従って、これらの言語で書かれた文を計算機で解析、処理するには、まず文を単語 (形態素) に分割する処理、すなわち形態素解析が必要となる。近年では隠れマルコフモデル (Hidden Markov Model, HMM)、最大エントロピーマルコフモデル (Maximum Entropy Markov Model, MEMM)⁵⁾、条件付確率場 (Conditional Random Fields, CRF)⁴⁾ などの確率モデルを仮定した上で、大規模な正解タグ付コーパス (e.g. 京都大学テキストコーパス^{*}, RWCP テキストコーパス, EDR コーパス^{**}, 日本語話し言葉コーパス^{***}) からの統計的学習によって統計モデルの推定を行う手法が主流である。日本語の形態素解析においては、浅原ら¹⁾、

内元ら⁹⁾¹⁰⁾、工藤ら³⁾ などが挙げられ、いずれも高い精度を達成している。

一方で、頑健な形態素解析システムを構築する上で、システムにとって既知でない単語、すなわち未知語をどう処理するかは非常に大きな問題となる。特に分かち書きされていない言語における未知語処理は解析のあいまい性が非常に高く、困難なタスクである。

形態素解析を要素技術とする応用分野として、特に近年では掲示板やブログなどといった媒体から情報を獲得するようなタスクの需要が急速に高まっている。このような媒体においては比較的くずれた言語表現が多用されるため、未知語が出現する頻度が極めて高く、また未知語の種類も非常に豊富である。このような言語資源をより頑健に解析するために、高精度な未知語処理の研究は急務である。ここで、掲示板やブログなどといった媒体では、通常漢字が用いられる表記に対してひらがなやカタカナを用いるなど多種多様な表記が現れることに注意していただきたい。このような曖昧な表記の存在は、単に既知語に対する解析を曖昧に

^{*} <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html>

^{**} <http://www2.nict.go.jp/kk/e416/EDR/>

^{***} <http://www2.kokken.go.jp/csj/public/>

するだけでなく、未知語にとって非常に大きな問題となることを後に指摘する。

以上の背景を踏まえ、本稿では日本語形態素解析における未知語処理の問題に取り組む。本稿で提案する手法は、通常の形態素解析と並行して、入力文中の各部分文字列に対して未知語の候補を常に生成するものである。このような未知語の候補の生成によって、未知語全体にわたる素性を柔軟に解析に利用できる。多種多様な未知語が現れるブログや掲示板での形態素解析には、このような素性の活用が重要な役割を果たすと思われる。

本稿では、CRFを学習モデルとして採用することを提案する。MEMMを初めとする、識別モデルを段階的に適用するモデルでは、label bias あるいは length bias の影響が避けられない。特に未知語を考慮した形態素解析では、上で述べたような曖昧な表記の問題と相まって、未知語解析における精度に対する label bias の影響が非常に大きいことを述べる。label bias, length bias に耐性のある CRF を学習モデルに採用することによって、既知語、未知語双方に対して高精度な解析ができる。

実験では、本稿で提案する手法を大規模な正解タグ付コーパスの解析に適用し、MEMM との結果を定量的、定性的に比較することにより、未知語、既知語双方の解析において精度が向上できることを検証した。

2. 日本語形態素解析

2.1 単語境界同定と品詞付与

日本語形態素解析は、入力文に対して単語境界同定を行い、同時に同定された単語境界各々に対し品詞を付与するタスクとなる。本稿では形式的に単語境界同定と品詞付与を同時に行うものとして考える。

単語境界同定及び品詞付与を行う最も単純な方法は、文字ごとに単語境界及び品詞をあらわすタグを付与する方法、すなわち文字タグ付け法である。しかしながら、この手法では文字単位の情報のみしか利用することができず、単語単位での情報を解析に活用しにくいという欠点を持つ。これはすなわち、豊富な辞書情報を解析に取り込みにくいという欠点につながる。辞書情報は、特に日本語において、既知語の形態素解析の精度に大きく貢献するため、文字単位でのタグ付けを日本語形態素解析に適用することは考えにくい。

そこで、入力文の各部分文字列に対して可能な全ての単語境界、品詞候補を、辞書を用いて生成する手法が考えられる。日本語形態素解析においてはこの手法が最も一般的である。この手法では、まず入力文（文字列） x が与えられたときに、 x の各部分文字列に対して可能な全ての単語境界、品詞候補を生成する。この手続きにより入力文に対して生成された、個々の単語境界、品詞候補を表現する頂点と、各候補間の連

接関係を表現する辺とで構成されたグラフ状の表現を形態素ラティスと呼ぶ。形態素ラティス中を通る全ての可能な単語境界、品詞列から、入力文に対して最も適切な単語境界、品詞列を推定し出力するのが形態素解析のタスクである。

より形式的に問題を定式化する。入力文 x に対して、辞書によって生成される可能な全ての単語境界、品詞列の集合を $\mathcal{Y}(x)$ とする。ここで、

$$\mathcal{Y}(x) = \{y_1, \dots, y_{|\mathcal{Y}(x)|}\}$$

であり、

$$y_n = (\langle w_{n1}, t_{n1} \rangle, \dots, \langle w_{n|y_n|}, t_{n|y_n|} \rangle)$$

である。 $w_{n1}, \dots, w_{n|y_n|}$ と $t_{n1}, \dots, t_{n|y_n|}$ は各々 y_n の単語境界列と品詞列を表し、 $|y|$ は出力系列 y の長さを表す。この表式が示すとおり、日本語、中国語など、分かち書きされていない言語における形態素解析では、一般に出力の単語境界、品詞列の長さは可変である。最も一般的な枠組みであるコスト最小法では、入力文 x が与えられたときに x に対して可能な単語境界、品詞列に各々コストを付与し、そのなかで最小のコストが割り当てられた単語境界、品詞列を正解とし、出力する。

近年では、入力文 x と単語境界、品詞列 $\mathcal{Y}(x)$ との間に何らかの確率モデルを仮定し、その確率モデルを大規模な正解タグ付コーパスから学習した上で、学習された確率モデルで推定される確率をコストの代用とする手法が盛んに研究、応用されている。最も一般的な形式では、入力文 x が与えられたときに、各単語境界、品詞列 $y \in \mathcal{Y}(x)$ が正解列である確率 $P(y|x)$ を用いて、 $\hat{y} = \operatorname{argmax}_y P(y|x)$ を正解として出力するタスクとして形式化される。 $P(y|x)$ のモデルとしては、HMM, MEMM⁵⁾, CRF⁴⁾ などが用いられる。

2.2 豊富で柔軟な単語情報の利用

日本語における単語境界同定は曖昧性の高い困難なタスクである。例えばひらがなで表記された単語は単語境界の曖昧性が高い。我々が最終的に解析対象としたいのは、ブログなど多種多様な表記が用いられる言語資源であり、この種の曖昧性を避けることはできない。また同時に、日本語文においては助詞の品詞など品詞付与における曖昧性も高い。このように単語境界、品詞付与の曖昧性が高いことから、他の言語と比較して、日本語形態素解析における品詞タグの粒度は細かく、階層化など、品詞タグに対する構造化も複雑である。例えば、ChaSen^{*} に用いられている品詞体系においては、最大で4階層にまで階層化されている。最も大きな範疇で「名詞」、その下に「名詞-固有名詞」、さらに「名詞-固有名詞-人名」、「名詞-固有名詞-人名-姓」と続く、という具合である。また一部の助詞など、大きな曖昧性が生じる語については、語単位で品詞が割り振られているものも存在し（語彙化）、品詞タグ

^{*} <http://chasen.naist.jp/hiki/ChaSen/>

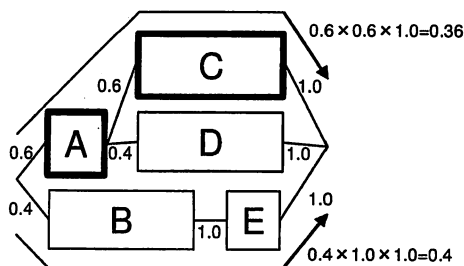


図1 label bias の例 (太枠は正解系列)

の数は 100 以上にも上る。また、品詞以外にも各語に対して読み、発音、活用型、活用形などといった属性も付与されている。このような曖昧で複雑な単語境界、品詞を同定するためには、学習および解析時に品詞の階層、活用形、活用型、またそれらの組み合わせなど、多種多様な素性が取り込める必要がある。

浅原¹⁾は HMM を拡張し、接続位置に基づく品詞のグループ化、語彙及び品詞情報、選択的な trigram の導入などを用いて高い精度を出している。しかしながら HMM は生成モデルであり、本質的に柔軟な単語情報の利用が困難であるという欠点を抱える。

単語候補の様々な情報を柔軟に解析に取り込むことは、単に既知語に対する解析精度だけではなく、未知語の解析においても非常に重要な役割を担うと考えられる。例えば、字種、単語内の字種の遷移、prefix, suffix, 文字数など様々な情報が未知語の同定に非常に有効に機能することは想像に難くない。

最大エントロピー法や Support Vector Machine (SVM)¹¹⁾といった識別モデルに基づく学習では、単語や単語間接続の情報を素性という形で柔軟に取り込むことができ、HMM に見られるような、素性同士がオーバーラップしてはならない(素性同士が互いに直交していない)といった制約も取り払われる。内元⁹⁾¹⁰⁾は、MEMM を拡張し、字種、字種遷移、語頭文字列、語末文字列、文字数などの素性を用いることにより高い精度で未知語解析を行っている。

2.3 label bias と length bias

MEMM など、識別モデルを段階的に適用するモデルでは、一般に label bias⁴⁾の影響を受ける。label bias とは、局所的に見て接続しやすい label を正解として選択しやすく、結果的に全体で見ると不自然な系列を選択してしまう現象を指す。

図1に label bias の例を示す。MEMM では、たとえば個々の接続の曖昧性を正しく解析できていたとしても、B に対する接続が E しか存在しないため、B からの接続の重みが全て E への接続に割り振られてしまう。入力文に対する正しい接続という観点から見れば B-E の接続は不正解であるにも関わらず、である。図1の例では、個々の接続に対して B に対する接続の重みが A に対する接続の重みと比較してそれ

ほど低くないこと、また B に後続する接続が E に続くもの1つしかない(より一般には後続の接続の曖昧性が低い)ことが label bias を現出させる要因になっている。また、B に対する局所的な接続の重みが特に高い場合には、たとえ B に後続する系列の曖昧性が比較的高い状況でも label bias が現出しえる。

label bias を学習時における問題として捉えた場合、識別モデルの段階的な適用では正解系列上の事象のみが考慮され、入力に対して可能な他の全ての系列上の事象が全く考慮されないことが問題である、と説明することもできよう。図1の事例を学習する場合を考えてみる。MEMM の学習では、「候補 A, B に対して A が選択される」、「A に対する C, D の接続の中から C が選択される」という事象は考慮されるが、B-E という接続がこの入力に対して適切ではないという事実は全く取り込まれないのである。

また、日本語形態素解析を初めとして、出力系列が可変長な系列であるタスクへの MEMM の適用は、length bias の影響も受ける。length bias とは、短い系列が長い系列よりも選ばれやすい問題である。MEMM では、個々の接続の確率の積で系列全体の確率を見積もるため、個々の接続確率が小さくても、系列が短い(積の回数が少ない)ほうが系列全体の確率が高くなりやすいのである。

これらの問題は入力文に対する全候補系列各々に対する確率を直接モデル化する CRF では現出しにくい。CRF では、学習において正解系列以外の全周辺系列がいわば不正解の事例として適切に学習に反映されるため、MEMM における label bias の問題が解消できる。図1の事例を CRF で学習する場合では、B-E の接続に対する重みが系列全体から見て低くなるように正しく学習されるのである。また系列全体で重みが正規化される CRF では、各出力系列の大きさに基づいて確率重みが正しくスケール化されると考えることができるため、length bias に対しても耐性がある。工藤³⁾では、MEMM の日本語形態素解析における label bias, length bias の2つの問題を論じ、また CRF を日本語形態素解析に適用することによって、既知語の精度を、HMM, MEMM などを用いる既存のシステムより向上させることに成功している。

2.4 未知語処理

2.4.1 日本語形態素解析における未知語処理

日本語形態素解析における未知語処理では、未知語の同定を形態素解析と同時に並行して行う手法と、形態素解析とは独立した前処理あるいは後処理として行う手法が考えられる。

例えば浅原²⁾は、N-best 解を出力する品詞タグ付解析と SVM を用いた未知語同定処理とを段階的に適用することにより高い未知語解析精度に到達している。しかしながら彼らの解析対象は新聞記事、特許文であり、未知語の種類やパターンは比較的限られたも

のであるといえよう。

未知語の出現頻度が高い言語資源における未知語の同定、あるいはひらがな表記による未知語など、単語境界の曖昧性をも内包するような未知語を多く含む言語資源における未知語の同定を想定する上で、既知語に対する単語境界、品詞同定と同時に、並行して未知語処理を行う発想はより自然であろう。形態素解析の精度向上が未知語同定の精度向上に貢献すると同時に、未知語同定の精度向上が形態素解析の精度向上に貢献すると考えられるからである。また、既知語に対する単語境界、品詞同定の曖昧性解消が、未知語の同定の決め手となる場合も多々ありうるためである。

未知語と既知語とを同時に処理するにあたり、大きくわけて2つの方法が考えられる。

1つは入力文中に存在する未知語を同定する問題を、入力文中の各文字に対する未知語としての位置、すなわちその文字が未知語を構成しているかしていないかのタグを付ける問題として解くものである。中川ら⁹⁾は、既知語の解析に対して、辞書を用いた単語境界、品詞候補を生成すると同時に、各文字に対する未知語タグ付けの候補を並行して生成し MEMM の枠組みで解析する手法を提案し、既知語、未知語双方で高い精度を実現している。この手法では、追加で生成されるノードの数が各文字位置毎に常に固定されたままであり、また解析できる未知語の長さにも上限が存在しない。一方で、未知語同定において文字単位の情報しか利用できず、未知語同定におけるモデルの表現力は不足すると思われる。

もう1つは、入力文字中の全ての部分文字列に対してその部分文字列が未知語として切り出される可能性を考慮して解析するものである。内元ら⁹⁾¹⁰⁾は、入力文の各部分文字列が形態素として切り出される可能性を全て考慮し、MEMM を拡張したモデルと、単語単位の様々な素性を利用して、それらの中から最適な推定を推定する手法を提案しており、未知語に対する高い精度を実現している。この手法では、最大で文長に等しい長さの未知語候補を生成する必要がある、また計算量は文長の2乗に比例してしまう。実用的には未知語候補の最大文字数に上限を設けることによってこの困難さは避けることができる。ただし、依然として未知語候補を追加することによる計算量の増大は無視できない。一方で、この手法では未知語の同定に未知語候補単位の情報が利用できるため、先頭文字列や未知語候補内での字種の遷移などといった複雑な素性も利用できる。ブログあるいは掲示板など、多種多様な表現が用いられる言語資源を解析対象とするにあたって、後者が持つ未知語候補単位で素性が獲得できるという利点は、大きく影響するものと思われる。

2.4.2 未知語処理と label bias

工藤ら³⁾は未知語処理にはほとんど関与しておらず、文字種などによる経験則的な手法に言及するのみ

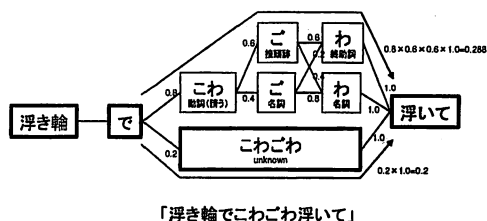


図2 未知語解析における label bias の例 (太枠は正解系列)

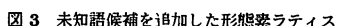
である。そして、学習及び解析時には入力文中の全単語の情報が既知であるとの仮定の上で実験、評価している。また彼らは、label bias, length bias の一般的な性質、あるいは既知語に対する label bias や length bias の具体的影響を論じているものの、それらが未知語を考慮した解析においてどう影響するかまでは論じていない。本稿では特に、未知語処理における label bias の影響に着目する。

先に述べたとおり、MEMM などの識別モデルを順次適用するモデルでは一般に label bias が存在する。この bias の存在は、日本語形態素解析における未知語処理では特に重要な問題である。

未知語処理において label bias が重要な問題となるのは、未知語に対する接続の重みが既知語に対する接続の重みに対して非常に小さいことに起因する。学習事例中に現れる既知語と未知語の頻度を比較した場合、一般に既知語が出現する頻度が未知語の頻度より圧倒的に高く、このような事例上で学習した結果として、局所的には既知語の接続に重みが振られやすい。そして段階的に識別モデルを適用する学習では、正解系列以外の系列が不正解であるということが学習に全く取り込まれない。この結果、未知語候補に対して既知語候補が対抗した場合に、たとえ既知語候補の後方の接続が不自然なものであっても既知語候補がより選択されやすくなる、という現象が発生する。

図2に未知語解析における label bias の典型的な例を挙げる。この例では「こわごわ」という副詞を未知語と想定した場合の解析例を挙げている。この例では、「で-こわ (動詞)」の接続は「で-こわごわ (未知語)」の接続の重みよりかなり大きい。さらに「こわ (動詞)」以降の接続はどれも非常に不自然なものばかりである。この「こわ (動詞)」以降の接続は学習事例中にはほとんど現れることがないと思われる接続であり、割り振られる重みは非常に不安定になる。結果的に、MEMM のような局所的に重みの正規化を行うモデルでは、たとえ「こわ (動詞)」以降の接続が非常に不自然なものであっても、局所的には非常に接続重みが大きい既知語「こわ (動詞)」への系列を取ってしまうのである。この現象は典型的な label bias の影響と捉えることができる。

一方で CRF の学習では、正解系列以外の全周辺



- 71 -

	RWCP (訓練時) / (開発用) / (評価時)
事例(文)数	23472 / 5868 / 7336
語数(コーパス中)	595980 / 148229 / 186594
異なり語数	23742 / 13309 / 14569
未知語数	10200 / 3142 / 3461
未知語率	1.71 % / 2.12 % / 1.85 %
語数(辞書中)	289554
素性数	300109

表 1 実験における RWCP コーパスの諸元

	CSJ (訓練時) / (開発用) / (評価時)
事例(文)数	252 / 64 / 80
語数(コーパス中)	659905 / 126995 / 144702
異なり語数	13579 / 6735 / 7397
未知語数	3950 / 1207 / 1371
未知語率	0.60 % / 0.95 % / 0.94 %
語数(辞書中)	22946
素性数	182427

表 2 実験における CSJ の諸元

ず, 1 講演を 1 文として学習, 解析した。フィラー, 言いよどみなどは完全に削除し, 言いかえなどは適宜適切な候補に修正した。

評価に使用したコーパスに現れる全単語は, そのままではシステムとして既知語ばかりとなるため, そのままでは学習及び評価時に未知語が全く現れない。従って, コーパス中の単語のいくつかを擬似的な未知語として取り扱った上でモデルを学習, 評価する。我々の実験における学習と評価に用いた擬似的な未知語の基準は, 全コーパス中においてその語の基本形が 1 回しか現れない単語である, とした。単純に活用を含んだ語の出現回数で未知語を定義した場合, ある語の 1 回しか現れない活用形が全て未知語として学習され, 評価される。これは解くべき問題を不必要に難しくする。また通常, ある活用語のある 1 つの活用形の情報が既知であれば, その語の他の全ての活用形の情報が既知である, と想定するのが自然と考えられるからである。上記のように定義された未知語のうちの一部は, システムが想定する未知語処理では対応できない。例えば 5 文字以下の部分文字列を未知語として学習する未知語処理を想定した場合, 6 文字以上の未知語は学習, 解析できない。このような未知語については学習においては既知語として取り扱い, 評価の際には未知語として解析した。未知語候補を展開する際の未知語の文字長は最大で 5 文字とした。

実験で用いた素性を表 3, 表 4 にまとめる。表 3 に示した素性は bigram 中の前後の unigram に全く同じものを用いた。表 4 中のカッコ内の数字は bigram 中の位置を表す。使用した素性はコーパスで共通である。素性は頻度などに対する閾値などを用いず, コーパス中に現れる全ての素性を実験に使用した。

語	
品詞大分類	
品詞細分類	
活用型	
活用形	
文字数	- 1, 2, 3, 4, 5, 6 以上
先頭文字列	1 文字, 2 文字
末尾文字列	同上
先頭文字種	漢字, ひらがな, カタカナ, アルファベット, 数字, その他
末尾文字種	同上
文字種遷移	漢字, ひらがな, カタカナ, アルファベット, 数字, その他の間の 0, 1 回の遷移

表 3 実験で用いた unigram 素性

<語 (-1), 品詞大分類 (0)>
<品詞大分類 (-1), 語 (0)>
<品詞大分類 (-1), 品詞大分類 (0)>
<品詞大分類 (-1), 品詞細分類 (0)>
<品詞細分類 (-1), 品詞大分類 (0)>
<品詞細分類 (-1), 品詞細分類 (0)>

表 4 実験で用いた bigram 素性

学習には CRF を用い, 比較対象として MEMM で全く同等の解析を行った。実験で用いた CRF と MEMM はともに, 3.2 節で述べたように正規分布を用いて正規化を行った。全素性に共通である事前分布の分散 C は, 実験の設定に関わるハイパーパラメータとなる。このハイパーパラメータは, 5 分割交差検定を用いて最適値を決定した。パラメータの推定には L-BFGS⁷⁾ を用い, パラメータの更新による対数尤度の相対変化が 10^{-4} 以下になった段階で学習を停止した。

4.2 定量的評価

まず実験結果の定量的な評価について述べる。定量的評価に用いる指標 P (精度), R (再現率), F (F-measure) は以下のように算出される。

$$P = \frac{\text{解析結果中の正解単語数}}{\text{解析結果中の総単語数}} \quad (6)$$

$$R = \frac{\text{解析結果中の正解単語数}}{\text{(正解データ中の単語数)}} \quad (7)$$

$$F = \frac{2PR}{P+R} \quad (8)$$

RWCP における実験の結果を表 5 に, CSJ における実験の結果を表 6 に示す。表において, seg. は未知語と既知語を含めた単語境界同定の数値, word は未知語と既知語を含めた語 (未知語に対しては未知語と推定できれば正解とする) に対する数値である。また unknown seg. は未知語だけに対する単語境界同定の数値である。

CRF と MEMM の解析精度を比較した場合, 既知語, 未知語双方の解析の結果において明らかに有意な精度差が見られる。既知語, 未知語双方における精度向上は, とともに CRF の適用による label bias, length bias の影響が解消された結果であると解釈できよう。

		seg.	word	unknown seg.
CRF ($C = 1.3$)	P	98.0	95.8	79.5
	R	97.9	95.7	69.3
	F	97.9	95.7	74.1
MEMM ($C = 0.6$)	P	96.4	93.1	66.2
	R	96.3	93.0	53.7
	F	96.3	93.1	59.3

表 5 RWCP 形態素解析実験結果 (単位は%)

		seg.	word	unknown seg.
CRF ($C = 1.0$)	P	98.8	94.8	70.9
	R	98.9	94.8	62.9
	F	98.8	94.8	66.7
MEMM ($C = 0.8$)	P	97.4	91.7	48.3
	R	97.8	92.0	36.3
	F	97.6	91.8	41.4

表 6 CSJ 形態素解析実験結果 (単位は%)

CRF による未知語の解析における精度の改善は特に注目される。先に述べたとおり、未知語の解析に対する label bias の影響が大きいために、CRF の採用による精度の改善もまた顕著に現れているものと解釈できる。特に CSJ においては CRF による未知語の解析精度の改善が著しい。これは CSJ での実験における未知語の割合が低いことと関連するものと思われる。先に述べたとおり、未知語に対する label bias は学習事例中の未知語と既知語の頻度差が主要な要因となっていると考えられる。このため、未知語の割合が小さいコーパスでは label bias が特に顕著に現れるものと推定される。一方 CRF による解析では label bias に対する耐性があるため、本実験における CSJ のように、未知語の割合が小さいコーパスにおいても MEMM と比較して高い未知語解析能力を発揮できるものと思われる。

4.3 定性的評価

次に定性的な観点から実験結果を考察してみる。以下で引用する事例は全て RWCP コーパスの実験の解析結果から引用したものである。

まず先に述べたとおり、MEMM の解析誤り例として顕著であるのが length bias の影響を受けたと思われる誤り事例である。図 4 に length bias の影響が顕著に示されている例を挙げる。図 4 では最下の系列が正解であるが、非常に曖昧性が高い。MEMM の解析結果は下方の非常に曖昧な系列を選択することを避け、短く、曖昧性の小さい経路である「やってくれ」を選択している。しかしながら、これは全体から見れば誤った系列を選択してしまっている。典型的な length bias の影響と捉えることができよう。

また、未知語の解析結果についても触れる。未知語が絡んだ解析において MEMM が誤る事例に顕著なのが、他の既知語と先頭の文字列を共有する未知語に対する解析誤りである。例を図 5、図 6 に示す。図中の例文における未知語を下線で示している。MEMM

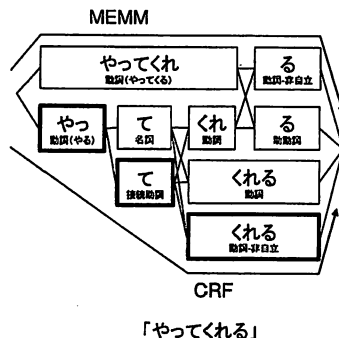


図 4 MEMM の解析誤り例 (正解系列は太枠)

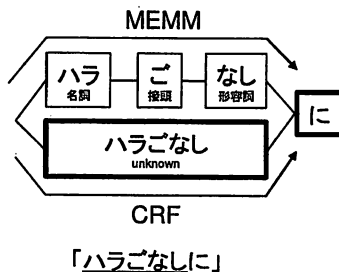


図 5 未知語に対する MEMM の解析誤り例 (正解系列は太枠)

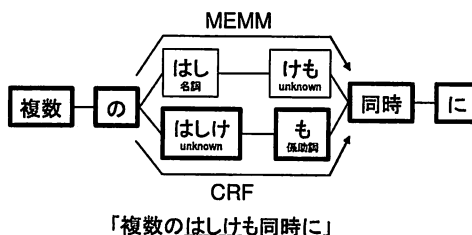


図 6 未知語に対する MEMM の解析誤り例 (正解系列は太枠)

におけるこのタイプの誤りは実験結果の誤り事例解析において非常に顕著であった。このタイプの誤り事例の原因は以下のように説明されると思われる。

訓練事例中においては通常、既知語の頻度が未知語の頻度より圧倒的に高い。これから、局所的には未知語に対する接続の重みは、既知語に対する接続の重みに比べて非常に小さいと考えられる。また、図 5、図 6 において MEMM が選択している系列の後方にある接続「ハラ-ご-なし」や「はし-けも」は、そもそも MEMM の学習中にはほとんど現れないような接続である。このため MEMM の学習ではこれらの接続の重みが非常に不安定になる。この結果、MEMM による解析では既知語を先頭に含んだ非常に不自然な系列が、未知語を含んだ正しい系列の重みより勝って

しまうため、未知語の解析に失敗する。これは label bias の典型的な影響と解釈することができよう。一方で CRF はパス全体の確率を最大化しつつ、学習事例に対して可能な全ての不正解系列を考慮して学習できるように、正しく正解系列である未知語への接続を選択できると考えられる。結果、CRF ではたとえ未知語が他の既知語と共通な先頭文字列を含んでいたとしても、適切に未知語を切り出しうるわけである。

上に述べたような未知語解析に対する CRF の利点は、特に日本語のブログなど比較的小ざれた文体の日本語文に対する形態素解析において非常に重要な性質である。なぜならば、ブログなどでは未知語の頻度が高いと同時に、ひらがな表記など、多種多様な曖昧な表記が好んで用いられるからである。MEMM では、単に未知語の存在が問題になるだけではなく、ひらがな表記の既知語といった他の未知語と曖昧になる既知語の存在が、未知語解析の精度を著しく減少させることが上の定性的な説明から明らかである。一方で CRF は、たとえ既知語と未知語との間に曖昧性が存在しても、正しく未知語を含んだ正解系列を解析しえるのである。未知語を考慮した日本語形態素解析を行う上で、MEMM に対する CRF の優位性は明らかであろう。

5. 結 論

本稿では、未知語処理を形態素解析と同時に並行して行うことにより、未知語処理を行う手法を提案した。同時に、従来用いられてきた MEMM などでは特に未知語処理における label bias が大きな問題になることを示し、CRF を適用することによりこの問題を解消できることを示した。そして、大規模な正解タグ付コーパスの未知語処理付形態素解析に本稿で提案する手法を適用することにより、その有効性が実際に検証できた。特に、MEMM と比較して未知語に対する精度の向上は著しいものであった。

今後は、素性の取捨選択、あるいは trigram 言語モデルの導入などによるより豊かな言語情報を用いて既知語、未知語双方の解析精度向上を図るとともに、未知語に対してより頑健かつ高精度なモデルを模索したい。特に Web などの、非常に大規模だが正解タグ付けがなされていないような言語資源から、何らかの「単語らしさ」の尺度を抽出し、日本語形態素解析における未知語解析に組み込むことは非常に有効と考えられる。また Peng ら⁸⁾ が提案するように、確信度の高い未知語同定結果を辞書情報として半自動で取り込むことにより、単語境界同定と未知語同定の精度を双方向で向上するような枠組みもまた興味深いと思われる。

参 考 文 献

- 1) Asahara, M. and Matsumoto, Y.: Extended Models and Tools for High-performance Part-of-speech Tagger, *Proceedings of COLING*, pp. 21–27 (2000).
- 2) Asahara, M. and Matsumoto, Y.: Japanese Unknown Word Identification by Character-based Chunking, *In Proceedings of COLING*, pp. 459–465 (2004).
- 3) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *In Proceedings of EMNLP*, pp. 230–237 (2004).
- 4) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *In Proceedings of 18th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 282–289 (2001).
- 5) McCallum, A., Freitag, D. and Pereira, F.: Maximum Entropy Markov Models for Information Extraction and Segmentation, *In Proceedings of 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 591–598 (2000).
- 6) Nakagawa, T.: Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information, *In Proceedings of COLING*, pp. 466–472 (2004).
- 7) Nocedal, J. and Wright, S. J.: *Numerical Optimization*, Springer-Verlag New York, Inc. (1999).
- 8) Peng, F., Feng, F. and McCallum, A.: Chinese Segmentation and New Word Detection using Conditional Random Fields, *In Proceedings of COLING*, pp. 562–568 (2004).
- 9) Uchimoto, K., Nobata, C., Yamada, A., Sekine, S. and Isahara, H.: Morphological Analysis of The Spontaneous Speech Corpus, *In Proceedings of COLING*, pp. 1298–1302 (2002).
- 10) Uchimoto, K., Nobata, C., Yamada, A., Sekine, S. and Isahara, H.: Morphological Analysis of a Large Spontaneous Speech Corpus in Japanese, *In Proceedings of ACL*, pp. 479–488 (2003).
- 11) Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer (1995).