社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (5)
2009－ＮＬ－189 (5)
2009／1／22

CRFによる係り受け解析の結果を反映させた日本語形態素解析

岸本 貴之 †， 高橋 治久 ††， 堀田 一弘 ††，

† 電気通信大学大学院電気通信学研究科情報通信工学専攻 †† 電気通信大学電気通信学部情報通信工学科

本稿では，日本語形態素解析の精度を，条件付確率場 (CRF) による係り受け解析を用いて，改善する方法を提案する．従来の確率モデルによる形態素解析は，一般的に，1個または2個前までの単語の品詞情報の相関関係によって，最適な候補を絞り込むというやり方を行っていた．しかし，それだけでは解析できない事例が存在しており，もっと広い範囲での単語の相関や，構文関係などを考慮に入れたモデルを考える必要がある．本稿では，形態素解析結果の候補に対し，係り受け解析を行い，その尤度を最大にする形態素解析結果により係り受け解析を選択する方法が，精度改善に有効であることを，従来法との比較実験により示す．

Japanese Morphological Analysis Reflecting Result of Dependency Analysis Using CRFs

Takayuki KISHIMOTO † Haruhisa TAKAHASHI †† Kazuhiro HOTTA ††

† Dept. of Informatin and Communication Engineering, Graduate school of Electro-Cmmunications

†† Dept. of Information and Communication Engineering, The Univ. of Electro-Communications

This paper presents a method of improving Japanese morphological analysis via Conditional Random Fields (CRFs) using the dependency analysis. Many existing probabilistic methods select a correct tokens by the correlation analysis between adjoining words and their part-of-speech. However, some instances cannot be correctly analyzed only with the correlation between adjoining words. In order to improve the accuracy, it would be needed to take into account correlation of words in wider range as well as syntactical features. We show that maximizing the likelihood of the dependency analysis for candidates of correct tokens improves the accuracy by computer experiments.

1 はじめに

形態素解析は日本語文章理解や情報検索において基礎となる技術である。日本語形態素解析の手法として，Hidden Markov Models (HMM) [1] や，Maximum Entropy Markov Models (MEMM) [3]，Conditional Random Fields (CRF) [4] などの学習モデルによる解析が成功している．特に CRF は，単語境界と品詞情報の同定が 95.9% の精度で完全正解しており，他の手法と比べても最高水準のモデルである．しかし，このモデルを用いても，誤った形態素を含んだ文がまだまだ多く存在しているのが現状である．日本語の文章を理解する上では，一つの文で正しい形態素列を得ているかどうかは重要であり，この意味で完全正解文の割合による精度 (文正解率) が一つの指標となる．

本稿では，文正解率を上げることを目指して，新たに CRF による係り受け解析の結果を反映した CRF による形態素解析手法を提案する．CRF による形態素解析では，隣接する2個までの単語間から構成される素性関数を基に，形態素系列の確率を求める [4]．しかし，実際にはもっと広い範囲での単語の相関に影響されており，隣接する2単語だけでは解析できない事例が存在している．一方，係り受け解析では，文中に含まれる全ての文節間から解析を行うため，隣接した2単語以外の相関についても比較する解析となる．したがって係り受け解析の結果を反映させることで，形態素解析の弱点を補完することができると考えられる．

文節の係り受け構造を決定するためには，正確な形態素を与える必要があるため，形態素の選択が適したものであるほど，係り受け解析の確かさもより大きくなると考えられる．このため，CRF による形態素解析で得られた尤度の高いいくつかの候補に対して，CRF による係り受け解析が最も良くなるような形態素系列を選べば，形態素解析の精度を上げることが出来ると期待される．

本稿では，2 章で CRF の一般式および日本語形態素解析と係り受け解析への適用法をそれぞれ説明する．次に 3 章で，係り受け解析の結果を反映した形態素解析モデルの適用法を説明する．4 章では京大コーパスを用いた提案モデルの実験結果および考察を述べ，最後に 5 章にて今後の課題について述べる．

2 条件付確率場

条件付確率場 (CRF)[5] は，HMM のように特徴の独立性を仮定する必要がなく，また MEMM のような label bias や length bias の影響を受けにくい識別モデルである．一般に，入力系列 $\mathbf{x}$ に対して，出力系列 $\mathbf{y}$ を得る条件付確率 $P(\mathbf{y}|\mathbf{x})$ を，次のように表す．

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp\left(\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y})\right) \tag{1}$$

ただし $Z_{\mathbf{x}}$ は全出力系列 $Y(\mathbf{x})$ を考慮したときに確率の和が 1 になるようにするための正規化項であり，次式で与えられる．

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}' \in Y(\mathbf{x})} \exp\left(\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}')\right) \tag{2}$$

また，$\Phi(\mathbf{x}, \mathbf{y})$ は $(\mathbf{x}, \mathbf{y})$ に含まれる言語的特徴を示すベクトルであり，$\mathbf{w}$ は $\Phi(\mathbf{x}, \mathbf{y})$ の各要素に対する重み (パラメータ) を並べたベクトルである．

入力 $\mathbf{x}$ に対する最適な出力 $\hat{\mathbf{y}}$ は

$$\hat{\mathbf{y}} = \underset{\mathbf{y}' \in Y(\mathbf{x})}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}' \in Y(\mathbf{x})}{\operatorname{argmax}} \mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}) \tag{3}$$

となる．

2.1 パラメータ推定

パラメータ $\mathbf{w}$ は最尤推定を用いて選択する．すなわち学習データ $T = \{(\mathbf{x}, \mathbf{y})\}_{j=1}^{N}$ に対する対数尤度 $\mathcal{L}_{\mathbf{w}}$ の最大化を行う．

$$\begin{aligned}
\mathcal{L}_{\mathbf{w}} &= \sum_j \log\left(P(\mathbf{y}_j|\mathbf{x}_j)\right) \\
&= \sum_j \Big[\log\big(\sum_{\mathbf{y} \in Y(\mathbf{x})} \exp\big(\mathbf{w} \cdot [\Phi(\mathbf{x}_j, \mathbf{y}_j) - \Phi(\mathbf{x}_j, \mathbf{y})]\big)\big)\Big] \\
&= \sum_j \left[\mathbf{w} \cdot \Phi(\mathbf{y}_j, \mathbf{x}_j) - \log(Z_{\mathbf{x}_j})\right] \\
\hat{\mathbf{w}} &= \underset{\mathbf{w}}{\operatorname{argmax}}\, \mathcal{L}_{\mathbf{w}}
\end{aligned}$$

目的関数の凸性から，最適点では以下が成立する．

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathbf{w}}}{\partial \mathbf{w}} &= \sum_j \Big(\Phi(\mathbf{x}_j, \mathbf{y}_j) - \sum_{\mathbf{y} \in Y(\mathbf{x})} \Phi(\mathbf{x}_j, \mathbf{y}) P(\mathbf{y}|\mathbf{x}_j)\Big) \\
&= 0
\end{aligned} \tag{4}$$

本研究では，パラメータ $\mathbf{w}$ の更新を勾配法によって行う．

$$\mathbf{w}^{\text{新}} = \mathbf{w}^{\text{旧}} + \eta \frac{\partial \mathcal{L}_{\mathbf{w}}}{\partial \mathbf{w}}\bigg|_{\mathbf{w}=\mathbf{w}^{\text{旧}}}$$

2.2 CRF による日本語形態素解析

CRF による形態素解析 [4] では，入力文字列 $\mathbf{x}$ に対して形態素列 $\mathbf{y} = (\langle w_1, t_1\rangle \ldots, \langle w_{\#\mathbf{y}}, t_{\#\mathbf{y}}\rangle)$($w_i$: 単語, t_i : 品詞) を得る条件付確率 $P(\mathbf{y}|\mathbf{x})$ を求める．

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp\left(\mathbf{\Lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x})\right) \tag{5}$$

素性ベクトル $\mathbf{F}(\mathbf{y}, \mathbf{x})$ は，1 単語 $\langle w_i, t_i\rangle$，または隣接する 2 単語 $\langle w_i, t_i\rangle, \langle w_{i-1}, t_{i-1}\rangle$ の出力ラベルのみによって構成される．

最適出力系列は Viterbi アルゴリズムを用いて効率よく探索できる．また，Viterbi アルゴリズムを適用後に，後向きに A^* アルゴリズムを適用することで，N-best 解 (確からしさ上位 N 件の出力系列) を効率的に求めることができる [2].

2.3 CRF による日本語係り受け解析

入力となる文節列を $\mathbf{b} = \{b_1, b_2, \ldots, b_m\}$，係り受けパターン列を $\mathbf{d} = \{d_1, d_2, \ldots, d_m\}$ と定義する．ただし，d_iは，文節 b_i の係り先文節番号を表す．また制約条件として，$\mathbf{d}$ は，各文節はその文節より後方に必ず 1 つの係り先をもつものと仮定する．この制約により文節 b_i の係り先候補 $\mathcal{D}_i$ は $\mathcal{D}_i = \{i+1, \ldots, m\}$ となる．

CRF による係り受け解析では，文節列 $\mathbf{b}$ に対して，文節 b_i の係り先が j となる条件付確率 $P(d_i =$

$j|\mathbf{b}$) を求める．

$$P(d_i = j|\mathbf{b}) = \frac{1}{Z_{b_i}} \exp\left(\mathbf{\Theta} \cdot \mathbf{G}(b_i, b_j)\right) \quad (6)$$

素性ベクトル $\mathbf{G}(b_i, b_j)$ は，2 つの文節 b_i, b_j あるいはその文節間で観測される情報によって構成される．

それぞれの係り関係は独立であると仮定すると，文節列 $\mathbf{b}$ に対する係り受けパターン列 $\mathbf{d}$ を得る条件付確率 $P(\mathbf{d}|\mathbf{b})$ は，

$$P(\mathbf{d}|\mathbf{b}) = \prod_{i=1}^{m-1} P(d_i = j|\mathbf{b}) \quad (7)$$

となり，最適な係り受けパターン列 $\hat{\mathbf{d}}$ は，

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmax}} P(\mathbf{d}|\mathbf{b}) \quad (8)$$

となる．

2.3.1 係り受け解析の負例の学習

従来の係り受け解析は，正しい形態素解析 (正例) が与えられた場合に，正しい係り受け解析結果を最大確率とするように設計されている。しかし本研究では，誤った形態素解析結果に対しても適用されるため，本研究の係り受け解析 CRF には，誤った形態素 (負例) を与えられた場合での学習も同時に行う．

負例を含んだ文節について，その係り先候補全てが均等に係り先となりうるよう学習する．すなわち，式 (4) の $\Phi(\mathbf{x}_j, \mathbf{y}_j)$ を，次のように置き換える．

$$\frac{1}{\#Y(\mathbf{x}_j)} \sum_{\mathbf{y} \in Y(\mathbf{x}_j)} \Phi(\mathbf{x}_j, \mathbf{y}) \quad (9)$$

このようにすることで，負例を含んだ文節の係り受け解析について，エントロピーが大きくなり，得られる尤度を小さくさせることができる．

3 係り受け解析結果を反映させた形態素解析

一般に，従来の形態素解析のアプローチは，連続する 2 個または 3 個までの単語の品詞情報の相関関係によって，最適な候補を絞り込むという方法をとっており，それ以上の個数の単語あるいは離れた単語同士での相関関係，文全体での構文関係・意味関係などは一切考慮に入れない．これは，広い範囲での単語間の相関関係を考慮に入れれば，必要な計算時間や学習パラメータ数が，莫大に増えてしまうためであり，構文関係・意味関係は，形態素が不明では解析できないためである．しかし実際には，隣り合う数個の単語品詞間だけでは，適切な形態素が必ずしも決定できるわけではない．

実際に，2.2 節の CRF による従来型の形態素解析で実験したところ，正解率は表 1 のような結果になった．この結果から分かるように，単語ごとに見た正解率は非常に高いものの，文単位で見た正解率 (文正解率) は，十分な精度とは言い難いというのが現状である．

表 1: 従来型の CRF による解析精度

	単語正解率	文正解率
完全正解	95.91	44.34
品詞正解	97.14	56.80
区分正解	98.94	81.14

本稿では，入力された文字列に対して，係り受け解析の尤度 (確率値) が最も大きくなるような形態素を，最適な候補として決定する方法を提案する．すなわち，入力系列 $\mathbf{x}$ に対して，最適な出力系列 $\mathbf{y}$ を次のように決定する．

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\hat{\mathbf{d}^{\mathbf{y}}}|\mathbf{x}) \quad (10)$$

ただし，$\hat{\mathbf{d}^{\mathbf{y}}}$ は，$\mathbf{y}$ によって与えられる文節列 $\mathbf{b}$ の最適な係り受けパターン列である．

$$\hat{\mathbf{d}^{\mathbf{y}}} = \underset{\mathbf{d}^{\mathbf{y}}}{\operatorname{argmax}} P(\mathbf{d}^{\mathbf{y}}|\mathbf{b}^{\mathbf{y}}) \quad (11)$$

3.1 定式化

本節では，モデル CRF による形態素解析・係り受け解析を用いて，係り受け解析結果を反映させた形態素解析モデルの定式化を考える．一般に，形態素解析と係り受け解析を，同一の学習ネットワークに組み込めば，全ての形態素候補列に対して係り受け解析を行うことになるため，処理に膨大な時間を要してしまう．そこで，処理の簡約化のために，以下の仮定を設ける．

1. 形態素解析の出力と係り受け解析の出力は，互いに独立であるとする
2. 形態素解析の確からしさ上位 N 件の出力系列候補についてのみ，係り受け解析を行う

仮定 1. により，$P(\mathbf{d}^{\mathbf{y}}|\mathbf{x})$ は，従来型の形態素解析によって得られる形態素列の尤度と，それに対する係り受けパターン列の最大尤度との積となる．

$$\begin{aligned} P(\mathbf{d}^{\mathbf{y}}|\mathbf{x}) &= \sum_{\mathbf{y}'}[P(\mathbf{y}'|\mathbf{x})P(\mathbf{d}^{\mathbf{y}}|\mathbf{y}')] \\ &= P(\mathbf{y}|\mathbf{x})P(\mathbf{d}^{\mathbf{y}}|\mathbf{y}) \\ &= P(\mathbf{y}|\mathbf{x})P(\mathbf{d}^{\mathbf{y}}|\mathbf{b}^{\mathbf{y}}) \qquad (12) \end{aligned}$$

これによって，形態素解析自体のネットワークの規模を大きくさせることなく，係り受け解析の結果を適用するだけで，精度改善が可能となる．

また，制約 2. により，形態素解析の出力上位 N 件についてのみ係り受け解析を行い，その N 件の中で最適な形態素列を決定する．よって，式 (5), (6), (7) より最適な形態素列 $\mathbf{y}$ は，

$$\begin{aligned} \hat{\mathbf{y}} &= \operatorname*{argmax}_{\mathbf{y}} P(\mathbf{d}^{\hat{\mathbf{y}}}|\mathbf{x}) \qquad (13) \\ &= \operatorname*{argmax}_{\mathbf{y}(\text{上位 } N \text{ 件})} \log P(\mathbf{y}|\mathbf{x})P(\mathbf{d}^{\hat{\mathbf{y}}}|\mathbf{b}^{\mathbf{y}}) \\ &= \operatorname*{argmax}_{\mathbf{y}(\text{上位 } N \text{ 件})} \Big[\mathbf{\Lambda}\cdot\mathbf{F}(\mathbf{y},\mathbf{x}) + \sum_j \Big(\mathbf{\Theta}\cdot\mathbf{G}(b_j^{\mathbf{y}}, b_{d_j^{\mathbf{y}}}^{\mathbf{y}}) \\ &\qquad - \log(Z_{b_j^{\mathbf{y}}})\Big)\Big] \end{aligned}$$

となる．

4 計算機実験

4.1 実験方法

係り受け解析のフィードバックの有用性を示すために，提案モデルと従来型モデルとの比較実験を行った．また，提案モデルについては，3.1 節で導入した N の大きさと正解率との関係についても合わせて比較を行った．

実験には京都大学テキストコーパスを用いた．実際に用いたデータの詳細を表 2 に示す．形態素解析，係り受け解析のどちらについてもモデルは CRF を用い，またどちらに対しても表 2 のデータで学習および評価を行った．

形態素解析のモデル構築で用いた素性パターンは [4] と同じものを用いた．ただし，これらのパターンのうち，学習コーパス中に 3 回以上観測された素性のみを用いた．

また，係り受け解析については，表 3 のような基本素性を用いた．これらの中から [8] を参考に，有効と思われる基本素性の組み合わせを選択し，実際の素性として利用した．このうち，学習コーパスで 3 回以上観測された素性のみを用いた．

表 2: 実験に用いるデータ

データベース	京大コーパス ver.3	
ソース	毎日新聞 (95 年)	
学習	記事	1 月 1 日〜8 日
	文数	7,896
	単語数	189,804
	文節数	74,191
評価	記事	1 月 9 日
	文数	1,268
	単語数	30,607
	文節数	12,214

素性に用いられる用語について，以下に述べる．

主辞 文節内で，品詞大分類が特殊・助詞・接尾辞となるものを除き，最も文末にある形態素．

語形 文節内で，特殊を除き最も文末に近い形態素．

見出し 語の原形．ただし活用のない語の場合は語の表記とする．

Major 活用する語の場合は活用型，活用のない語の場合は品詞大分類．

Minor 活用する語の場合は活用形，活用のない語の場合は品詞細分類．

表 3: 係り受け解析に用いた基本素性

前/後文節
主辞見出し, 主辞品詞大分類,
主辞品詞細分類, 主辞活用型, 主辞活用形,
語形見出し, 語形 (Major) , 語形 (Minor)
句読点の有無（なし，読点，句点）
括弧の有無（なし，開，閉）
文節の位置（文頭，文末，それ以外）
文節間
文節間距離（1, 2-5, 6 以上）
文節間括弧の有無（なし，開，閉，開閉）
文節間読点の有無（なし，あり）

評価は，*all*(単語全情報が正解)，*top*(単語区切りと品詞大分類が正解)，*seg*(単語区切りが正解) の3つの基準で，単語正解率と文正解率を比較する．ただし，単語正解率については，システムの出力する単語数が，コーパスの単語数と異なる場合があるため，式(14)で示すF値によって評価を示す．また，文正解率は，文全体の解析が正しいものの割合を示す．

$$
\begin{aligned}
F\text{値} &= \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \qquad (14)\\
Recall &= \frac{(\text{正解した単語数})}{(\text{コーパスに含まれる単語数})}\\
Precision &= \frac{(\text{正解した単語数})}{(\text{システムが出力した単語数})}
\end{aligned}
$$

解析システムはC++で実装した．全ての実験において，CPU: Pentium IV 3.4GHz，主記憶: 2.0Gbyteの Windows XP の環境で行った．

4.2 実験結果

$N = 5, 10$ の場合での提案モデルと，従来モデルについて，正解率は表4のようになった．

表4: 単語正解率 (F値) と文正解率 (%)

単語正解率	従来	$N=5$	$N=10$
all	95.91	96.02	96.00
top	97.14	97.23	97.21
seg	98.94	98.98	99.00
文正解率	従来	$N=5$	$N=10$
all	44.34	45.41	45.49
top	56.80	57.46	57.62
seg	81.14	81.64	82.13

結果より，N がいずれの場合においても，提案モデルは，従来型より精度が優れていることが分かる．したがって，係り受け解析の結果を反映させることで，形態素解析の精度の改善が実現可能であると考えられる．

従来型のCRFによる形態素解析では，隣接する2単語の情報のみを素性として用いていたため，3単語以上での相関や，離れた単語同士での相関を考慮に入れることができなかった．これに対して，係り受け解析では，素性として，2つの文節間の情報を用いるため，複数個の単語間の相関を同時に比較することになり，さらに，離れた文節間の相関も比較するので，離れた単語間の相関も比較することになる．このように，隣接した2個の単語以外の相関を比較することになるため，係り受け解析を考慮に入れることで，形態素解析の問題点を改善させることができると考えられる．

4.3 文長ごとの性能評価

係り受け解析を反映させることによって，離れた形態素同士の相関を考慮できるようになるため，文長（一文に含まれる単語の数）が長くなるほど，精度が大きく向上できると期待される．これを確かめるために，コーパス内で一文に含まれる単語数が19以下，20〜34，35以上の場合の3段階に分けて文正解率を比較した．各結果を表5に示す．

表5: 文長ごとの文正解率

形態素数		従来	$N=5$	$N=10$
19以下	*all*	64.55%	64.75%	64.75%
	top	72.34%	72.13%	72.13%
	seg	90.37%	90.78%	90.78%
20-34	*all*	37.55%	39.63%	39.63%
	top	53.11%	54.56%	54.56%
	seg	80.50%	81.33%	81.54%
35以上	*all*	18.00%	18.80%	19.20%
	top	33.60%	34.40%	35.20%
	seg	64.40%	64.40%	66.40%

この結果は，全て期待通りの結果になった．短い文における精度向上があまり見られないのに対して，文長が中程度以上の文においては1〜2%程度の精度向上が見られた．

4.4 CRFによる係り受け解析の性能評価

提案モデルの性能評価とは別に，CRFによる係り受け解析の性能評価も行った．また，比較のために，チャンキングの段階適用法による係り受け解析 [6] (以下チャンキングモデル) の性能評価も行った．

ただし，判定には全て正しい形態素による文を与えるため，本実験に限り，CRFによる係り受け解析モデルにも正しい形態素のみを学習で与えた．

チャンキングモデルは，近い文節に係りやすいというヒューリスティックに沿って，効率的にSupport Vector Machine (SVM)[7] で学習，識別を行う手

法である．本実験では，[6] に合わせて，SVM のパラメータ C を $C = 1$ とした．また，Kernel 関数には多項式 Kernel を用い，次元数 d は $d = 3$ とした．

それぞれのモデルについて，解析時間および正解率は表 6 のようになった．

表 6: 係り受け解析の実験結果

CRF	係り受け正解率 [%]	88.61
	文正解率 [%]	42.46
	解析時間 [秒/文]	0.0011
チャンキング	係り受け正解率 [%]	89.29
	文正解率 [%]	47.53
	解析時間 [秒/文]	0.18

チャンキングモデルと比較して，CRF は，精度の面でやや劣るものの，解析時間の面では大幅に上回っていることが分かる．

提案手法による形態素解析では，1 つの文に対して N 回の係り受け解析を行うため，解析時間の短い係り受け解析モデルを選択することが重要となる．この点において，CRF は，提案手法に適用する係り受け解析の学習モデルとして適していることが分かる．

また，CRF の形態素解析を単独に用いた場合においても，HMM や MEMM といった既存手法の中で，最も高い精度を持っており [4]，提案手法に適用する学習モデルとして，CRF の選択は妥当であると考えられる．

5 おわりに

本稿では，CRF による係り受け解析の結果を反映させた日本語形態素解析の手法を提案し，計算機による実験結果を示した．従来の形態素解析では，隣接する 2 個までの単語からなる素性関数のみで解析していたのに対し，係り受け解析を取り入れることで，離れた複数個の単語の相関を比較することができる．そのため，一文ごとの精度，特に語数が多い文に対して大きな精度改善が可能となる．また，形態素解析自体のネットワークの規模を大きくさせることなく，精度改善を実現できることが本手法の利点である．

より高い精度の解析を実現するためには，誤った形態素に対する係り受け解析の学習方法，形態素解析に有効な係り受け解析の素性選択の検討が必要になる．また，フィードバック処理を効率化させることで，係り受け解析のモデルを大規模化させても，実用に十分な時間で高い精度の解析が可能であり，今後は，これらの点について検討し，さらに改善されたモデル構築を実現したい．

参考文献

[1] 村上 仁一, 嵯峨山 茂樹. HMM を用いた形態素解析. 情報処理学会第 45 回全国大会, pp.3-161-162, 1992.

[2] 永田 昌明. 前向き DP 後向き A^* アルゴリズムを用いた確率的日本語形態素解析システム. 情報処理学会 研究報告 NL-101-10, pp.73-80, 1994.

[3] 内元 清貴, 関根 聡, 井佐原 均. 最大エントロピーモデルに基づく形態素解析 — 未知語の問題の解決策 — . 自然言語処理 Vol.8 No.1, pp.127-141, 2001.

[4] 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 自然言語処理研究会 SIGNL-161, pp. 89-96, 2004.

[5] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of ICML, pages 282-289, 2001.

[6] 工藤 拓, 松本 裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌 Vol.43 No.6, pp.1834-1842, 2002.

[7] Vladimir N. Vapnik. Statistical Learning Theory. Wiley-Interscience, 1998.

[8] 内元 清貴, 関根 聡, 井佐原 均. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌 Vol.40 No.9, pp.3397-3407, 1999.