2007－NL－182 (3)
2007／11／19

係り先候補の相対的な距離を反映した統計的日本語係り受け解析

山本 悠二 † 増山 繁 †

† 豊橋技術科学大学 知識情報工学系

概要: 本稿では，係り先候補の相対的な距離を反映した統計的日本語係り受け解析手法を提案する．統計的係り受け解析手法は，文節間の係りやすさを訓練データから推定する．その際，従来手法では，文節間の距離はいくつかのカテゴリに分けられ，推定に用いられる素性として明示的に与えられる．しかし，複数の文節間候補が同一の距離カテゴリに属する場合，距離による弁別ができないため，最尤の係り先を決定することが困難である場合が多い．そこで提案モデルでは，文節候補集合中の二つの文節候補を逐次的に取り出し，どちらが係り元に近いかを明示させて係りやすさの推定を行う．京都大学コーパスを用いて実験を行った結果，係り受け正解率 91.60 %，文正解率 56.33 % となり，ベースライン手法と比べて有意に改善していることが確認された．

Statistical Japanese Dependency Analysis Reflecting Relative Distances Among Modifiee Candidates

Yuji YAMAMOTO † Shigeru MASUYAMA †

† Department of Knowledge-based Information Engineering, Toyohashi University of Technology

Abstract: We propose a novel method for statistical Japanese dependency analysis, which reflects relative distances among modifee candidates. Statistical Japanese dependency analizers estimate a dependency likelihood between a pair of *bunsetsu* chunks from training dataset. In conventional approaches, distances between pairs of *bunsetsu* chunks are divided into some feature categories, and the categories are embedded into training feature set explicitly. However, modifee candidates that belong to the same distance category, are possibly hard to be selected the most likely one, since they can not be distinguished by their distance categories. The proposed method selects two modifee candidates from all candidates sequentially. Each of the two modifee candidates and its modifier estimate the dependency likelihood after the modifee candidate appends extra information whether it is nearer from its modifier. The experimental results using Kyoto University Corpus achieved a dependency accuracy of 91.60% and a sentence accuracy of 56.33% respectively. We confirmed that the proposed method improved both dependency and sentence accuracy significantly, compared with the base-line method.

1 はじめに

係り受け解析は，文節間の関係についての基本的な情報を与えるため，自然言語処理における基本技術として認識されている．特に係り受け解析の精度向上は，解析後の応用タスクへの結果の改善に直接影響することが多いため，自然言語処理の研究課題において重要な位置付けとなっている．初期の研究では文節間の係りやすさをルールベースで定めていたが，網羅性や一貫性の点で問題がある．近年では，文節間の係りやすさを係り受け情報が付与された大規模コーパスから学習アルゴリズムを用いて統計的に推定する方法が提案されるようになった．

しかし，統計的係り受け解析において，連体修飾語の係り先の曖昧性や従属節の曖昧性といった依存構造の曖昧性がある場合の解析についてはいまだに困難な点がある．一般にこのような曖昧性に対して有用な素性は訓練データに存在する語彙情報や距離素性である．一方，初期の係り受け解析の研究における文節間の係りやすさをルールによって定める方法では，「日本語文内の文節は係り得る最も近い文節に係る」という優先規則がある．先の優先規則に見られるどの係り先候補が係り元に近いかという情報は「文節間の距離をいくつかのカテゴリにまとめた距離素性」とは異なる．距離素性を使用すると文節間の距離が異なる場合でも同一の素性が割り当てられるといった問題が生じ，係り先候補が係り元に近いかどうかを正確に表すことができない場合がある．そこで本稿では係り先候補の相対的な距離を反映した統計的係り受け解析手法を提案する．提案手法では，文節候補集合中の二つの文節候補を取り出

して，どちらが係り元に近いかという情報を明示的に与えて係りやすさの推定を行っている．

本稿の構成は以下の通りである．2節で統計的日本語係り受け解析について概説する．3節で一般に用いられている文節間素性にある距離素性についての問題点を示し，4節で提案手法について述べる．5節では提案手法のオンライン学習アルゴリズムでの定式化について示す．6節で実験とその結果，関連研究について述べ，7節でまとめを行う．

2 統計的日本語係り受け解析

この節では，依存文法に基づく係り受け解析でよく用いられているモデルについて説明する．表記法については，主に [4] に準拠する．

前提として日本語文における係り関係が以下のような制約を満たすものとする．

(1) 係り受けは前方から後方に向いている (後方修飾)．

(2) 係り関係は交差しない (非交差条件)．

(3) 係り関係は係り先を一つだけ持つ．

制約の (1) と (3) より，文末には係り先が存在しないことが導かれる．

次に，文節列とその係り先についての定義を示す．まず，日本語の文に対して，その文節列を順序付き集合 $B = \{b_1, b_2, \cdots, b_m\}$ とする．また，ある文節 b_p と，別の文節 b_q が与えられているとする．このとき，$p < q$，つまり，b_p が b_q よりも前に出現することを，順序関係 $\prec$ を用いて $b_p \prec b_q$ と表す．同様に，b_q が b_p より後に出現することを $b_q \succ b_p$ と表す．

そして，文節 b_i $(1 \leq i \leq m)$ を係り元の文節とする係り受けパターン列を順序付き集合 $D = \{d_1, d_2, \cdots, d_m\}$ とする．ただし，d_i は文節 b_i の係り先文節番号を示すものとする．例えば，文節 b_i が 文節 b_j に係る場合は $d_i = j$ となる．また，b_i が b_j に係るとき，$b_i \rightarrow b_j$ と表記する．なお，文末の文節の係り先 d_m は前提により存在しないため，便宜的に $d_m = -1$ と定義する．

統計的係り受け解析は，訓練データとして与えられた複数の文節列と係り受けパターン列を用いて，新たな入力文の文節列から最尤の係り受けパターン列を生成する問題であると捉えることができる．

ここで，個々の文節の係り関係は全て独立だと仮定[1]して係り先を決定することを考える．日本語係り関係の前提 (1) から，文節 b_i の係り先候補は，b_i より後方にある全文節である．b_i の係り先候補集合を $C_i = \{b_{i+1}, \cdots, b_m\}$ と表記する．また，b_i とその係り先候補 $b_j \in C_i$ を特徴付ける素性ベクトルを $F(\langle b_i, b_j \rangle) \in \mathbb{R}^n$ と表記する．一般に，素性ベクトルには，本稿で問題として取り上げる文節間の距離を含め，係り元と係り先候補の品詞や語彙などの情報，後方文脈を含む語彙情報を素性として組み入れる．この時，重みベクトル $\mathbf{w} \in \mathbb{R}^n$ を用いて，b_i と b_j の係りやすさの確信度を $\mathbf{w} \cdot F(\langle b_i, b_j \rangle)$ と定める[2]．

重みベクトル $\mathbf{w}$ は，訓練データから与えられた複数の文節列と係り受けパターン列を使用して学習を行う．その際，学習方法に関して，絶対モデルと相対モデル [4] の二つの戦略がある．前者は，素性のベクトル空間に対して，係ると判定する領域と係らないと判定する領域に分離超平面を用いて線形分離させるという戦略である．このとき，$\mathbf{w}$ は，分離超平面の法線ベクトルを表す．一方，後者は，係り先候補間での係りやすさの大小関係を学習するという戦略である．本稿では文節候補間の位置による選好について取り上げるため，主に相対モデルを中心に議論を進める．なお，絶対モデルでは文節候補間の比較を行うことで文節間の係りやすさの推定を行うことはできない．また，相対モデルは絶対モデルと比較して係り受け正解率，文正解率ともに有意に向上していることが実験的に知られている [4]．

3 文節間距離素性における問題点

まず，従来の統計的係り受け解析手法で一般的に使用されている素性の一覧を表 1 に示す．これらは多少の違いはあるものの，[1–8] 等で使われている素性である．

本稿では，これらの素性のうちで文節間距離素性についての問題点について指摘する．この素性は初期のルールに基づく係り受け解析における「日本語文内の文節は係りえるもっとも近い文節に係る」[9] という優先規則を統計的日本語係り受け解析でも適用できるように組み入れられたものであると解釈できる．一般に文節間距離素性はデータの過疎性を回避するために，文節間のそのままの距離を使用せ

[1] ただし，素性から係りやすさの確信度を求める場合，他の文節との係り関係を素性として明示的に組み込むことが多い．

[2] 上記は線形カーネルを用いた場合の確信度である．実験では多項式カーネルを用いている．

表 1　一般的に用いられる文節間素性

前/後文節	主辞見出し，主辞品詞，主辞品詞細分類， 主辞活用形，語形見出し，語形品詞， 語形品詞細分類，語形活用，語形活用形， 括弧の有無，句読点の有無 文節の位置 (文頭，文末)
文節間	距離 (1，2-5，6 以上) 括弧の有無，句読点の有無

ず，いくつかの距離のカテゴリに分割して使用される．[1] では，距離素性は係り受け解析において特に有効であることを個々の素性の抜き取った実験結果から確認している．

次に，文節間距離素性では文節候補間における距離の弁別ができない例を図 1 に示す．この場合，係り関係 {彼が → 読んだ} と {彼が → 買った} は同一の距離素性に属するため，どちらの係り先が近いか遠いかを弁別することが不可能である．そのため，係りやすさの確信度を距離素性をもとにして推定することできず，最尤の係り先を選択することが困難になることが考えられる．もちろん，このような場合においても，文節間素性，動的素性[3]の使用により，正しい係り先を同定できる可能性はある．しかし，それらの素性が有効である場合は，訓練データ中に係りやすさの推定をする文節間にある素性と類似した文節間素性や動的素性が存在する時である．このような素性のもととなる文節は異なり数が多く，係りやすさの推定にうまく貢献できない場合が考えられる．

この問題の単純な回避方法として，文節間距離素性の距離に関するカテゴリを増やすことで，それぞれの係り先候補における係り元との距離を弁別することが考えられる．しかし，文節間距離カテゴリの細分化は，データの過疎性を引き起こすため，係りやすさの推定に有効に働かない可能性がある．加えて，先に示した「係りえるもっとも近い文節に係る」という優先規則を適用するために必要な情報は，係り先文節のうちから二つの文節が選ばれたときにどちらが係り元に近いかであり，必ずしも係り元からの絶対的な距離を必要としない．そのため，二つの係り元文節のうちでどちらかが近いかを表す情報として文節間距離の絶対量を使用するのは特徴量としてあまり汎化しているものとはいえない．

[3] 係り元もしくは係り先において解析途中で既に得られている係り関係をもとにした素性

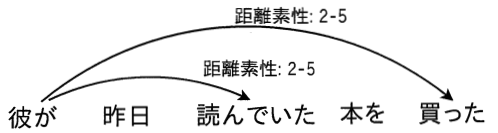

図 1　文節間距離素性では距離の弁別が不可能な例

4　提案手法

先に示した文節間距離素性の設定は，単一の文節間における距離をもとにしている．一方で複数の文節間の相対的な「近さ」「遠さ」を一種の距離素性として取ることも可能である．例えば，係り先候補集合から任意の異なる二つの文節を取り出した場合，どちらかの文節が係り元に近く，もう一方の文節が係り元から遠いという位置関係が得られる．

そこで提案手法では，係り先候補から任意の異なる二つの文節を取り出し，係り元に近いほうの文節間素性集合に “L”，遠いほうの文節間素性集合に “R” という素性を追加することで文節間の相対的な距離による弁別を行えるようにする．以下では，素性 “R”，“L” を位置素性と表記する．これらの係り先候補文節において，どちらが係り元から係りやすいかという順序付けが行えるならば，文節間距離素性からは弁別できなかった係りやすさが推定できることが期待できる．以上のことから，提案手法の係りやすさの推定の方針を以下に示す．

全文節 b_i と，その候補集合 C_i について，以下の制約を満たすようにベクトル $\mathbf{w} \in \mathbb{R}^n$ を導出せよ．

$\forall i$，$\forall c \in C_i \backslash \{b_{d_i}\}$ において，

- $b_{d_i} \prec c$ (b_{d_i} が c の前に出現) の場合:
  $\mathbf{w} \cdot F(\langle b_i, b_{d_i} \rangle, \mathrm{L}) > \mathbf{w} \cdot F(\langle b_i, c \rangle, \mathrm{R})$
- $b_{d_i} \succ c$ (b_{d_i} が c の後に出現) の場合:
  $\mathbf{w} \cdot F(\langle b_i, b_{d_i} \rangle, \mathrm{R}) > \mathbf{w} \cdot F(\langle b_i, c \rangle, \mathrm{L})$

ただし，

c: 候補文節中にある正しい係り先でない文節

b_{d_i}: 正しい係り先文節

である．また，

$F(\langle b_i, b_j \rangle, \mathrm{R})$: $F(\langle b_i, b_j \rangle)$ に距離素性 R を付加した素性ベクトル

$F(\langle b_i, b_j\rangle, \mathrm{L})$: $F(\langle b_i, b_j\rangle)$ に距離素性 L を付加した素性ベクトル

と定める．

ここで提案手法と従来研究での相対モデルとの違いについて述べる．従来研究での相対モデルの係りやすさの推定の方針を以下に示す．

> 全文節 b_i と，その候補集合 C_i について，以下の制約を満たすようにベクトル $\mathbf{w} \in \mathbb{R}^n$ を導出せよ．
>
> $\forall i$，$\forall c \in C_i \backslash \{b_{d_i}\}$ において，
> $\mathbf{w} \cdot F(\langle b_i, b_{d_i}\rangle) > \mathbf{w} \cdot F(\langle b_i, c\rangle)$

これより，提案手法は従来手法の相対モデルに対して，係り先候補と係り元の位置関係により素性ベクトルを変えるといった拡張を行っていることが分かる．

次に，係りやすさの推定により求めたベクトル $\mathbf{w}$ を用いて，係り元文節 b_i と係り先候補集合 C_i から最尤の係り先を探索する方法を考える．係り先候補集合 C_i の要素数を n としたときに，時間計算量 $O(n)$ で係り先を探索する戦略として以下の方法が考えられる．

> 係り元に最も近い二つの係り先候補文節を取り出す．そして，それらの二つの文節に対して位置素性を付与した文節間素性ベクトルを用いて，どちらが係り先として尤もな文節であるかを選出する．これを繰り返すことで勝ち抜き戦の形で最尤の係り先を同定する．

例を図 2 に示す．これは，図 1 の係り元と係り先候補集合から最尤の係り先を探索するという問題設定である．このときの探索過程は以下のようになる．

1. 係り先候補文節 b_2 と b_3 の比較
   $\mathbf{w} \cdot F(\langle b_1, b_2\rangle, \mathrm{L}) < \mathbf{w} \cdot F(\langle b_1, b_3\rangle, \mathrm{R})$ が成り立ち，文節 b_3 を選出．
2. b_3 と b_4 の比較
   $\mathbf{w} \cdot F(\langle b_1, b_3\rangle, \mathrm{L}) > \mathbf{w} \cdot F(\langle b_1, b_4\rangle, \mathrm{R})$ が成り立ち，文節 b_3 を選出．
3. b_3 と b_5 の比較
   $\mathbf{w} \cdot F(\langle b_1, b_3\rangle, \mathrm{L}) > \mathbf{w} \cdot F(\langle b_1, b_5\rangle, \mathrm{R})$ が成り立ち，文節 b_3 を選出．これにより，最尤の係り先を b_3 に決定する．

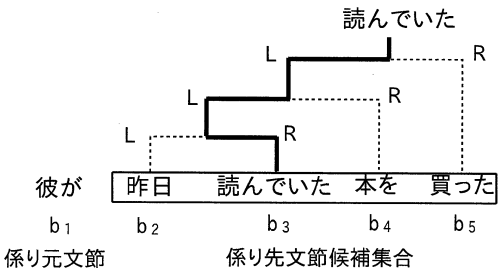

図 2　係り先文節の探索

このように対象候補集合から勝ち抜き戦の形で最尤の候補を探索する方法はトーナメントモデル [13] と呼ばれ，日本語ゼロ代名詞の先行詞同定などに使われている．

5 Passive-Aggressive アルゴリズムによる定式化

提案手法で定めた係りやすさの推定の方針は，二つの対象のうちでどちらが上位になるかをパラメータの更新則とする学習アルゴリズムならば適用可能である．このような学習アルゴリズムには，パーセプトロン [10] や最大エントロピー法，SVM の優先度学習への拡張 [11, 12] などといったものがある．本稿では，カーネルの適用，大量の学習事例への対応，外れ値への頑健性から，オンライン学習アルゴリズムの一種である Passive-Aggressive アルゴリズム [14] を用いて定式化を行う．

まず，L 個の学習事例が存在する訓練用のデータセットを $T = \{(b_{T,1}, b_{T,d_1}, C_{T,1}), \cdots, (b_{T,L}, b_{T,d_L}, C_{T,L})\}$ と定める．ただし，$b_{T,k}$ は，k 番目の学習事例における係り元文節，b_{T,d_k} は，k 番目の学習事例における正しい係り先文節，$C_{T,k}$ は，k 番目の学習事例における係り先文節候補集合とする．

今，$k-1$ 番目までの学習事例でパラメータの更新を行った重みベクトルを $\mathbf{w}_{k-1}$ とする．そのとき，$\mathbf{w}_{k-1}$ と k 番目の学習事例を用いて，以下の最適化

問題を解くことにより重みベクトルの更新を行う．

$\forall c \in C_{T,k} \backslash \{b_{T,d_k}\}$ において，

$$\mathbf{w}_k = \arg\min_{\mathbf{w}'} \frac{1}{2}||\mathbf{w}' - \mathbf{w}_{k-1}||^2 + D\sum_c \xi_c \qquad (1)$$

subject to

$$\mathbf{w}' \cdot \Delta F(b_{T,k}, b_{T,d_k}, c) \geq 1 - \xi_c$$

$$\xi_c \geq 0$$

ただし，$\Delta F(b_{T,k}, b_{T,d_k}, c)$ を従来手法での相対モデル，提案手法での相対モデルに関して以下のように定める．

- 従来手法での相対モデル
  $F(\langle b_{T,k}, b_{T,d_k}\rangle) - F(\langle b_{T,k}, c\rangle)$
- 提案手法での相対モデル
  - $b_{T,d_k} \prec c$ のとき:
    $F(\langle b_{T,k}, b_{T,d_k}\rangle, \mathtt{L}) - F(\langle b_{T,k}, c\rangle, \mathtt{R})$
  - $b_{T,d_k} \succ c$ のとき:
    $F(\langle b_{T,k}, b_{T,d_k}\rangle, \mathtt{R}) - F(\langle b_{T,k}, c\rangle, \mathtt{L})$

なお，式 1 中の D は，各学習事例におけるベクトル $\mathbf{w}$ へのパラメータ更新の影響を表す変数である．D が大きければより積極的にパラメータ更新を行う．以下，変数 D をコストパラメータと呼ぶ．

先に示した最適化問題のラグラジアン双対問題は以下のように表される．

$\forall c, c' \in C_{T,k} \backslash \{b_{T,d_k}\}$ において

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \sum_{c,c'} \alpha_c \alpha_{c'} \Delta F(b_{T,k}, b_{T,d_k}, c) \cdot \Delta F(b_{T,k}, b_{T,d_k}, c')$$

$$\max \sum_c \alpha_c \{1 - \mathbf{w}_{k-1} \cdot \Delta F(b_{T,k}, b_{T,d_k}, c)\} - \frac{1}{2}\sum_{c,c'} \alpha_c \alpha_{c'} \Delta F(b_{T,k}, b_{T,d_k}, c) \cdot \Delta F(b_{T,k}, b_{T,d_k}, c')$$

subject to

$$0 \leq \alpha_c \leq D$$

この双対問題は，勾配登り法 [15] などのアルゴリズムにより解くことができる．また，上式中の内積部分を任意のカーネル関数に置き換えることによりカーネル化が可能である．

なお，実験では [10] と同様に，訓練データセット単位で T 回 のパラメータ更新を行い，毎回更新される重みベクトルを平均化したものを係りやすさの推定に使用する重みベクトル $\mathbf{w}$ として使用する．

6 実験

6.1 設定

京都大学コーパス 3.0 を以下の 3 つに分けて実験を行った．

- 訓練データ: 一般記事 1 月 1，3-11 日，社説 1-8 月，合計 24,263 文，47,580 文節
- 開発データ: 一般記事 1 月 12-13 日，社説 9 月，合計 4,833 文，47,580 文節
- 評価データ: 一般記事 1 月 13-17 日，社説 10-12 月，合計 9,287 文，89,982 文節

これらの記事の使いかたは [4]，[8] と同じである．

係り受け解析手法として，ベースライン，提案手法ともに文末の文節から文頭に向かって逐次的に係り先を同定するアルゴリズム [2] を使用した．実験で使用したベースラインは，先行研究での相対モデル [4] を使用し，ビーム幅 1 で係り先を探索した．なお，提案手法は，それぞれの係り元に対して最尤の係り先を一つだけ決定する方法である．

学習に用いた基本素性は表 1 の通りである．ただし，主辞とは文節内で品詞が特殊，助詞，接尾辞となるものを除いた文末に一番近い形態素，語形とは文節内で品詞が特殊となるものを除き，文末に一番近い形態素のことを指す．ただし，基本素性中の文節間距離素性の影響について調べるために，1) 距離素性を付加しない場合，2) 距離素性 (1，2 以上) を付与する場合，3) 距離素性 (1，2-5，6 以上) を付与する場合の三通りについて確認を行った．

また実験では基本素性の他に動的素性 [5] も使用した．動的素性は係り元もしくは係り先において解析途中で既に得られている係り関係を素性として使用するものである．具体的には以下のような文節を素性として使用している．

- 着目している係り先に既に係っている文節
- 着目している係り先が係る文節

ベースライン，提案手法のモデルはともに [6] に合わせて三次の多項式カーネルを使用した．学習パラメータについてはコストパラメータを 1 に固定し，反復回数 T を開発データから 4 回と定めた．

6.2 結果と考察

提案手法とベースライン手法の結果を表 2 に示す．ここでの係り受け正解率は文末の一文節を除くすべ

表 2　結果 - 係り受け正解率，文正解率

手法	加えた距離素性	係り受け正解率 (%)	文正解率 (%)
提案	1，2-5，6 以上	91.60 (73913/80695)	56.33 (5231/9287)
	1,2 以上	91.59 (73907/80695)	56.30 (5229/9287)
	なし	91.53 (73863/80695)	55.98 (5199/9287)
ベースライン	1，2-5，6 以上	91.30 (73671/80695)	55.24 (5130/9287)
	1,2 以上	91.29 (73667/80695)	55.25 (5131/9287)
	なし	91.08 (73494/80695)	54.66 (5076/9287)

ての文節に対して，正しく係り先が同定できたものの割合，文正解率は文単位で全体の文節の係り先が正しく同定できたものの割合を示す．

まず，提案手法とベースライン手法の比較について行う．距離素性のいずれの場合においても係り受け正解率，文正解率ともに向上していることが分かる．ここで「母比率に差がない」という帰無仮説を立ててマクネマー検定を行う．結果を表 3 に示す．なお，表中の係り受け，文の列にある値は，マクネマー検定での P 値である．また，括弧内の値は提案手法のみが正解した数，ベースライン手法のみが正解した数をそれぞれ表す．結果からいずれも P 値が 0.01 未満となり，有意水準 1%未満で有意差があることが確認できる．

表 4 では，係り先距離ごとの F 値を示す．ただし，距離 n の精度は，係り受け解析器が出力した距離 n の係り受け解析結果のうちで正解だった割合を表す．また，距離 n の再現率は，正解データ中の距離 n の係り受けのうちで係り受け解析が正しい係り受けを出力した割合を表す．加えた距離素性ごとに提案手法，ベースラインを比較すると，距離 1 を除いては精度，再現率ともに向上が見られる．提案手法は二つの文節候補が取り出された場合にどちらが係り元に近いか遠いかを弁別することができる．そのため，距離素性では複数の距離をまとめあげてカテゴリ化していた 2-5，6 以上の距離での係り受け正解率が向上したと考える．

提案手法において，表 2 と表 4 を見る限りでは距離素性を加えることは係り受け正解率，文正解率を向上させるように考えられる．そこで，提案手法における加えた距離素性ごとについてマクネマー検定で比較を行った．結果を表 5 に示す．なお，表中の係り受け，文の列にある括弧内の値は，加えた距離素性の左側の素性を加えたときのみ正解した数，右側の素性を加えたときのみ正解した数をそれぞれ表す．表から，有意水準 5%未満であっても「母比率に差がない」という帰無仮説を棄却できない．この理由は，距離素性を加えることで正解する係り受けもしくは文は存在するが，逆に素性を加える前に正解していたものが不正解になる数も無視できないぐらいあるためである．したがって，提案手法において距離素性を付加することは検討を要するであろうと結論付ける．

6.3　先行研究との比較

同じデータセットを使用している先行研究に関しての比較を表 6 に示す．

表中の相対モデルとベースラインは同一の学習方法である．ベースライン手法と相対モデルを比較すると相対モデルの方が係り受け正解率，文正解率ともに高い．この理由は相対モデルはオフライン学習である最大エントロピーモデルを使用しているのに対して，ベースライン手法はオンライン学習を使用しているためであると考える．オフライン学習はデータセット中に存在するすべての学習事例から最適な分類器を生成できるのに対して，オンラインアルゴリズムでは逐次的に与えられた学習事例を用いて分類器のパラメータを更新するところが異なる．

チャンキングモデルは，直後の文節に係るか係らないかという観点のみで決定的に係り受け解析を行うアルゴリズムである．このような係り受け解析手法は，局所的な探索空間で係り関係を同定するため，一般に確信度に基づく手法に比べ高速に解析することが可能である．なお，決定的な係り受け解析手法には文節数に対して線形時間で係り先を同定するアルゴリズムが存在することが知られている [7]．これらのアルゴリズムは日本語係り受けは近い文節に係りやすいという特徴から，探索範囲について後方の文脈を考慮しない近傍に限定して決定的に係り先を同定する．このため，特に係り先が短距離における解析精度が高い．一方，提案手法のモデルでは係り先候補のどの文節が係り元と近いかを探索の過程で

表 3　提案手法とベースライン手法の比較

加えた距離素性	係り受け		文	
1，2-5，6 以上	3.782×10^{-14}	(628/386)	5.113×10^{-8}	(219/118)
1,2 以上	3.056×10^{-14}	(615/375)	1.320×10^{-7}	(218/120)
なし	2.2×10^{-16} 未満	(962/593)	4.722×10^{-8}	(311/188)

表 4　係り先距離ごとの比較: F 値 (精度/再現率) (%)

手法	加えた距離素性	1	2-5	6-9	10 以上
提案	1，2-5，6 以上	97.3 (96.8/97.8)	82.9 (84.4/81.4)	74.7 (74.6/74.9)	74.5 (71.3/78.1)
	1,2 以上	97.3 (96.9/97.8)	82.9 (84.4/81.4)	74.5 (74.1/75.0)	74.1 (71.1/77.3)
	なし	97.3 (96.9/97.7)	82.8 (84.2/81.5)	74.3 (73.9/74.6)	74.1 (70.9/77.7)
ベースライン	1，2-5，6 以上	97.2 (96.7/97.8)	82.1 (83.8/80.4)	74.1 (73.9/74.4)	73.4 (70.0/77.3)
	1,2 以上	97.3 (96.7/97.8)	82.1 (83.8/80.5)	73.9 (73.4/74.4)	73.3 (70.2/76.6)
	なし	97.1 (96.8/97.4)	81.9 (83.0/80.7)	73.4 (73.0/73.8)	73.0 (69.8/76.6)

明示的に与えているため，係り先文節の短距離での係り先の選好を反映していると考える．加えて，提案手法のモデルは相対モデルの拡張であるため，後方文脈による係りやすさも考慮することができる．提案手法がチャンキングモデルよりも結果が良いのはそのためであると考える．

組合せモデルは相対モデルとチャンキングモデルを併用し，係り先距離に応じてどちらのモデルを係り先を使用するかを決めて係り受け解析を行うものである．組み合わせモデルと提案手法の結果はほぼ同等であるが，本手法は一つの係り受け解析モデルしか使用していない．モデルを二つ併用すると，それぞれのモデルで素性の設定やパラメータの調節が繁雑となり，より良い結果を出すのが困難である．その点で，本手法はパラメータ設定が比較的容易であるところが利点であるといえる．

n-best リランキングモデルはあらかじめ後方文脈モデル [3] で求めた n-best の係り受け結果に対して，大規模コーパスから得た格要素間の従属関係，格要素・用言間の共起関係の統計情報を用いてリランキングを行うものである．このモデルは係りやすさの推定を係り受けタグ付きの訓練データからのみではなく，他のタグのないコーパスも併用しているという点で興味深い．一方で，提案手法は n-best リランキングモデルよりもよい結果となった[4]．これは，提案手法が相対モデルの拡張であるため，後方文脈モデルのもととなっている絶対モデルよりもよく係りやすさの推定が行えたと考える．

[4] 連体修飾節の解析を統合したモデル [8] では，係り受け正解率 91.25 %，文正解率 55.24 %である．

7　まとめ

本稿では，係り先候補の相対的な距離を反映した係り受け解析手法の提案した．具体的には，文節候補集合中の二つの文節候補を取り出し，どちらが係り元に近いかを明示させて係りやすさの推定を逐次的に行うことで係り先候補の相対的な距離を反映させている．また，提案手法は従来手法の相対モデルの拡張であることを示し，モデルの定式化，京都大学コーパスを用いた実験を行った．実験結果から，従来の相対モデルよりも係り受け解析，文正解率ともに有意に改善されていることが確認された．

今後の課題として，提案手法の n-best 係り受け解析結果を用いた半教師付き学習による性能評価を試みたい．

参考文献

[1] 春野 雅彦，白井 諭，大山 芳史．“決定木を用いた日本語係り受け解析”．情報処理学会論文誌，Vol. 39, No. 12，pp. 3177-3186，1998.

[2] 関根 聡，内元 清貴，井佐原 均．“文末から解析する統計的係り受け解析アルゴリズム”．自然言語処理，Vol. 6，No. 3，pp. 59-73，1999.

[3] 内元 清貴，村田 真樹，関根 聡．“後方文脈を考慮した係り受けモデル”．自然言語処理，Vol. 7，No. 5，pp. 3-17，2000.

[4] 工藤 拓，松本 裕治．“相対的な係りやすさを考慮した日本語係り受け解析モデル”．情報処理学会論文誌，Vol. 46，No. 4，pp. 1082-1092，2005.

表 5　提案手法での加えた距離素性における比較

加えた距離素性	係り受け		文	
1，2-5，6 以上 vs. 1,2 以上	0.8375	(300/294)	0.9462	(111/109)
1，2-5，6 以上 vs. なし	0.1224	(528/478)	0.1032	(197/165)
1,2 以上 vs. なし	0.1251	(415/371)	0.07758	(150/120)

表 6　先行研究との実検結果の比較

手法	係り受け正解率 (%)	文正解率 (%)
提案手法 (距離素性: 1，2-5，6 以上)	91.60 (73913/80695)	56.33 (5231/9287)
ベースライン手法 (距離素性: 1，2-5，6 以上)	91.30 (73671/80695)	55.24 (5130/9287)
相対モデル [4]	91.37 (73733/80695)	56.00 (5201/9287)
チャンキングモデル [4]	91.23 (73624/80695)	55.59 (5163/9287)
組合せモデル [4]	91.66 (73969/80695)	56.30 (5229/9287)
n-best リランキングモデル [8]	90.95 (73390/80695)	54.40 (5052/9287)

[5] 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. 1832-1842，2002.

[6] Taku kudo，Yuji Matsumoto．“Japanese Dependency Analysis Based on Support Vector Machines”．*Proc.* EMNLP/VLC，pp.18-25，2000.

[7] 颯々野 学．“日本語係り受け解析の線形時間アルゴリズム”．自然言語処理，Vol. 14，No. 1，pp. 1-18，2007.

[8] 阿部川 武，奥村 学．“共起情報及び複数格の組み合せを考慮した係り受け解析”．自然言語処理，Vol. 12，No. 1，pp. 107-123，2005.

[9] 長尾 眞，佐藤 理史，黒橋 禎夫，角田達彦．“自然言語処理 (岩波講座 ソフトウェア科学 15)”．岩波書店，1996.

[10] Michael Collins．“Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms”．*Proc.* EMNLP，pp.1-8，2002.

[11] Ralf Herbrich，Thore Graepel，Peter Bollmann-Sdorra，Klaus Obermayer．“Learning Preference Relations for Information Retrieval”．ICML-98 *Workshop*: *Text Categorization and Machine Learning*，pp. 80-84，1998.

[12] Thorsten Joachims．“Optimizing Search Engines using Clickthrough Data”．*Proc.* SIGKDD，pp.133-142，2002.

[13] 飯田 龍，乾 健太郎，松本 裕治．“文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定”．情報処理学会論文誌，Vol. 45，No. 3，pp. 906-918，2004.

[14] Koby Crammer，Ofer Dekel，Joseph Keshet，Shai Shalev-Shwartz，Yoram Singer．“Online Passive-Aggressive Algorithms”．Journal of Machine Learning Research，Vol. 7，pp. 551-585，2006.

[15] Nello Cristianini，John Shawe-Taylor．“An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods”．Cambridge University Press，2000.