3Q-5

日本語推敲支援のための文の語順整序

田中 麻祐子 † 大野 誠寛 ‡ 加藤 芳秀 †† 松原 茂樹 ‡‡ 石川 佳治 ‡‡
† 名古屋大学 工学部電気電子・情報工学科 ‡ 名古屋大学 情報基盤センター
†† 名古屋大学 情報連携統括本部 ‡‡ 名古屋大学 大学院情報科学研究科

1 はじめに

日本語は，語順が比較的自由であるため，語順を強く意識しなくても，意味の通じる文を書くことができる．しかし，語順に関する制約がないわけではないため，文法的には間違っていないものの読みにくい文が生成されることがある．本論文では，日本語文に対して，より読みやすくなるように文節を並び替える手法を提案する．これまでにも語順整序を行う手法がいくつか提案されているが [1, 2]，これらはいずれも事前に係り受け解析を施すことを前提としている．これら従来手法に対し，本手法では，係り受け構造が付与されていない文を入力とし，係り受け解析と語順整序を同時に行う．係り受けと語順の適切さを同時に考慮することにより，読みやすい語順を同定することができる．新聞記事を用いた実験を行い，本手法の有用性を確認した．

2 日本語における語順と係り受け

文献 [3] では，語順を決定する基本的要因が詳細に整理されており，例えば，以下の例文に示すように，長い修飾句を持つ文節は前方に位置する傾向が強いといったことが指摘されている．

1. 鈴木さんが佐藤さんが何日かかってもどうしても解けなかった問題をすぐ解いてしまった。
2. 佐藤さんが何日かかってもどうしても解けなかった問題を鈴木さんがすぐ解いてしまった。

例文 1 は,「鈴木さんが」とその係り先「解いてしまった。」が遠く離れているため,「鈴木さんが」の係り先が分かりにくくなっており，例文 2 の方が自然であるといえる．この例は，例文 1 の係り受け構造が分かれば，例文 2 のように読みやすく語順を変更できる可能性があることを示唆している．一方，係り受け解析は一般に，係り元と係り先の距離が離れると，その同定精度は低下することが知られている．そのため，例文 1 は，例文 2 のように語順を変更した後に解析した方が高精度に解析できる可能性がある．このように語順整序と係り受け解析は互いに依存しているといえる．

3 語順整序手法

本手法では，文法的には間違っていないものの読みにくい文（形態素解析と文節まとめ上げは施された文）が入力されることを想定し，その文に対して，係り受け解析を行うと同時に，より読みやすくなるような語順を同定する．なお，本手法では，文節の言い換えなどは行わず，文節を並び替えることのみを行う．

3.1 語順整序のための確率モデル

本手法では，入力文の文節列を $B = b_1 \cdots b_n$ とするとき，$P(O, D|B)$ を最大にする語順整序結果 $O = \{o_{(1,2)}, \cdots, o_{(1,n)}, \cdots, o_{(i,j)}, \cdots, o_{(n-1,n)}\}$ と係り受け構造 $D = \{d_1, \cdots, d_{n-1}\}$ を求める．ここで，$o_{(i,j)}$（なお，$1 \leq i < j \leq n$）は，2 文節間 b_i と b_j の語順整序後の順序を表し，文節 b_i が先か（$o_{i,j} = 1$），後か（$o_{i,j} = 0$）のいずれかの値をとる．また d_i は，文節 b_i を係り元の文節とする係り受け関係とする．

2 文節間の語順 $o_{(i,j)}$ は他の 2 文節間の語順とは互いに独立であり，かつ，係り受け関係 d_i も他の係り受け関係とは互いに独立であると仮定すると，ある O と D に対する $P(O, D|B)$ は以下のように計算できる．

$$\begin{aligned} P(O, D|B) &= P(O|B) \times P(D|O, B) \\ &\cong \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} P(o_{(i,j)}|B) \times \prod_{i=1}^{n-1} P(d_i|O, B) \end{aligned}$$

ここで，$P(o_{(i,j)}|B)$ は，文節列 B における文節 b_i と文節 b_j の語順が $o_{(i,j)}$ になる確率を，$P(d_i|O, B)$ は，文節列 B を語順整序結果 O に従って並び替えた後の文において，文節 b_i を係り元とする係り受け関係が d_i になる確率を表す．これらの確率はともに最大エントロピー法により推定する．$P(o_{(i,j)}|B)$ を推定する際は，文献 [1] で使われている素性のうち，係り受け情報を利用することなく取得可能な素性を用いた．$P(d_i|O, B)$ を推定する際は，文献 [4] と同じ素性を利用した．

3.2 探索アルゴリズム

入力文 B に対して考えられる O と D のパターンは膨大な数であり，O と D は互いに依存しているため，効率的に最適解を求めることは難しい．本手法では，入力文の語順と日本語の構文的制約（後方修飾性，非交差性，係り先の唯一性）[4] を利用し，$P(O, D|B)$ を最大とする O と D の近似解を効率よく探索する．

本研究では，入力文として文法的には間違っていない文を想定しており，入力文の係り受け構造は日本語の構文的制約を満たすものとしている．一方，語順整序後に文の意味が変わることは許されないため，語順整序後の文の係り受け構造は，語順整序後の語順で日本語の構文的制約を満たすだけでなく，入力文と同じものである必要がある．従って本研究では，D の探索空間を，入力文の語順と日本語の構文的制約から考えられる係り受け構造に絞り込むことができ，さらに，これら絞り込んだ係り受け構造から考えられる語順（係り受け構造は維持しつつ日本語の構文的制約を満

Word Re-Ordering for Japanese Revision Support
Mayuko Tanaka†, Tomohiro Ohno‡, Yoshihide Kato††, Shigeki Matsubara‡‡, Yoshiharu Ishikawa‡‡
†Dept. of Information Engineering, School of Engineering, Nagoya University ‡Information Technology Center, Nagoya University ††Information and Communications Headquarters, Nagoya University ‡‡Graduate School of Information Science, Nagoya University

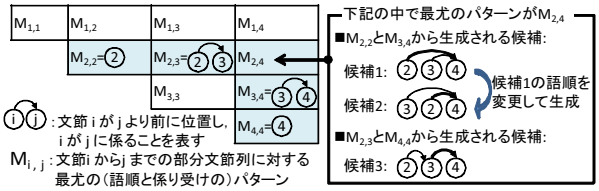

図 1: 探索アルゴリズムの実行例

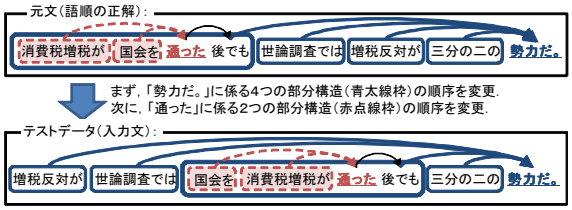

図 2: テストデータの作成例

たすように語順を変更したパターン）を探索すればよい．そこで本手法では，従来の係り受け解析で利用されてきた動的計画法を改良し，係り受け構造とその係り受け構造から考えられる語順をボトムアップに探索する．

本手法では，長さ n の入力文に対して，図 1 の左図のように（図は $n = 4$），$n \times n$ の三角行列を用意し，i 行 j 列に，部分文節列 $b_i \cdots b_j$ に対する最尤の語順 $O_{i,j}$ と係り受け構造 $D_{i,j}$ のパターン $M_{i,j}$ を書き込む（ただし，$M_{i,j}$ は最適解とは限らない）．左下から右上にかけて対角線に沿って三角行列を埋めていくことにより，最終的に，$P(O, D|B)$ を最大とする O と D の近似解 $M_{n,n}$ を求める．

$M_{i,j}$ を求める際の処理の例として，$M_{2,4}$ の場合を図 1 に示す．候補 1 と候補 3 のパターンは，従来の係り受け解析における動的計画法と同様に生成される．すなわち，候補 1 は $M_{2,2}$ と $M_{3,4}$ の最終文節同士を，候補 3 は $M_{2,3}$ と $M_{4,4}$ の最終文節同士を係り受け関係で結ぶことにより生成される．一方，候補 2 は，従来の係り受け解析では生成されないものであり，語順整序を同時に行うために生成されるパターンである．この語順整序のためのパターンは，従来の係り受け解析における動的計画法と同様に生成されたパターン（候補 1 や候補 3）から，その係り受け構造を維持したまま語順を変更することにより生成される．候補 1 からは語順が異なるパターンとして候補 2 が一つ生成され（b_4 に係る b_2 と b_3 の順序を入れ替えて生成），候補 3 からは係り受け構造の形から語順が異なるパターンは生成されない．このようにして生成された 3 つの候補のうち，$P(O_{2,4}, D_{2,4}|B)$ を最大とする候補パターンが $M_{2,4}$ となる．

4 評価実験

4.1 実験概要

本実験では，京大テキストコーパス [5] に収録されている新聞記事文に対して，係り受け構造を維持しつつ語順を変更した文をテストデータとして用いた．人間によって作成された文をテストデータとすることが考えられるが，本実験では，問題の焦点を語順に絞ることを考慮し，文意は取れるものの読みにくい文を擬似的に作成することとした．図 2 にテストデータの作成例を示す．文末から順に，複数の文節から係られる文節（「勢力だ。」や「通った」）を起点として，その文節に係る部分係り受け構造の順序をランダムに変更することを繰り返して作成した．なお，文中に含まれる読点は取り除いた．このようにして，京大テキストコーパスの 1 月 9 日分の新聞記事から，擬似的に作成した文（865 文，7,620 文節）をテストデータとした．なお，学習データには，7 日分（1 月 1 日，3〜8 日）の新聞記事（7,976 文）を用いた．

表 1: 実験結果（語順整序）

	2 文節単位の一致率	文単位の一致率
本手法	76.1% (29,541/38,838)	22.5% (195/865)
ベースライン 1	72.8% (28,265/38,838)	21.2% (183/865)
ベースライン 2	73.5% (28,541/38,838)	22.3% (193/865)
テストデータ	61.5% (23,886/38,838)	8.0% (69/865)

語順整序結果の評価では，文献 [1] と同様に，文単位一致率（語順整序後の語順が元の文と完全に一致している文の割合）と 2 文節単位一致率（2 文節ずつ取り上げた時の文節の順序関係が元の文のそれと一致しているものの割合）を測定した．

比較のために，2 つのベースラインを設けた．いずれも，まず係り受け解析を行い，その後に文献 [1] の手法により語順整序を行うものである．係り受け解析に文献 [4] の手法を用いる場合をベースライン 1，CaboCha[6] を用いる場合をベースライン 2 とする．

4.2 実験結果

本手法及び各ベースラインの語順整序結果を表 1 に示す．最下位行は，テストデータの語順（語順整序前の語順）で測定した語順一致率を示す．2 文節単位と文単位のいずれも，本手法が最も高い語順一致率を達成しており，本手法の有効性を確認した．

5 今後の課題

日本語文を書き慣れていない被験者が作成した文を収集し，それを用いて評価実験を行う予定である．

謝辞 本研究は一部，科研費 No.22300051，及び，No.22300034 により実施した．

参考文献

[1] 内元ら． コーパスからの語順の学習． 自然言語処理, Vol. 7, No. 4, pp. 163–180, 2000.

[2] 横林ら．係り受けの複雑さの指標に基づく文の書き換え候補の生成と推敲支援への応用．情処学論, Vol. 45, No. 5, pp. 1451–1459, 2004.

[3] 日本語記述文法研究会．現代日本語文法 7．くろしお出版, 2009.

[4] 内元ら．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． 情処学論, Vol. 40, No. 9, pp. 3397–3407, 1999.

[5] 黒橋, 長尾．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 回年次大会論文集, pp. 115–118, 1997.

[6] 工藤, 松本．チャンキングの段階適用による日本語係り受け解析． 情処学論, Vol. 43, No. 6, pp. 1834–1842, 2002.