係り受けを考慮したトポロジカルソートに基づく日本語文の語順整序

吉田 嵐 [1]　大野 誠寛 [1]　松原 茂樹 [2]
[1] 東京電機大学 大学院未来科学研究科　[2] 名古屋大学 情報基盤センター
25fmi35@ms.dendai.ac.jp　ohno@mail.dendai.ac.jp
matsubara.shigeki.z8@f.mail.nagoya-u.ac.jp

概要

日本語は語順が自由であるが，選好が存在するため，文法的には間違っていないものの読みにくい語順となっている文が作成されることがある．本稿ではそのような読みにくい文に対して，トポロジカルソートを用いた語順整序手法を提案する．提案手法では，2 文節間の前後関係の推定時にその 2 文節が係り受け木における祖先関係か否かを同時学習することにより，モデルの精緻化を図る．また，トポロジカルソートの適用後に複数の語順整序結果を得られるようにし，それらの出力文候補から最適な文を選択するモデルを導入する．

1　はじめに

日本語は語順が比較的自由であるが，実際には選好が存在しているため [1]，文法的には正しいものの，読みにくい語順を持った文が無意識に作成されることがある．そのような読みにくい文（例文 I）に対して，意味を変えずに読みやすい語順（例文 O）に整える語順整序手法が提案されている．

例文 I：子どもを昔のことを思い出し褒めた。
例文 O：昔のことを思い出し子どもを褒めた。

これまでの語順整序手法の多く（例えば，[2, 3, 4, 5, 6]）は，解析済みの係り受け情報，あるいは，同時的に解析して得られる係り受け情報を用いて，日本語の構文的制約[1)] を満たすように語順整序するアプローチを採用している．これらの手法は，係り受け解析が失敗すると，その影響を受けて，語順整序の精度も低下するという問題がある．

それに対して孫ら [8] は，係り受け情報を陽に用いることなく，1 文内のあらゆる 2 文節間の前後関係を推定し，その推定した前後関係をエッジ，各文節をノードとするグラフに対して，トポロジカルソートを実行することにより，文節を並び替える手法を提案している．必ずしも正しいとは限らない係り受け情報に縛られないという利点があるものの，その語順整序の精度は，係り受け情報を陽に用いる手法 [6] に及んでいない．2 文節間の前後関係の推定モデルの精緻化や，2 文節間の前後関係の誤りに対する頑健性が課題となっていた [8]．

そこで本稿では，孫らの手法 [8] を拡張し，トポロジカルソートを用いた語順整序手法を新たに提案する．具体的には，2 文節間の前後関係の推定において，係り受け情報の一部を同時に推定することにより，モデルの精緻化を図る．また，2 文節間の前後関係の誤りに対する頑健性を向上させるため，その推定確率が閾値以下の場合は両向きのエッジをもたせることにより，複数の語順整序結果を生成し，それらの出力文候補から，最も適切な文を選択するモデルを導入する．

2　日本語文の語順整序

日本語文の素朴な語順整序手法としては，入力文が n 個の文節を持つ場合，$n!$ 通りの語順を全て生成し，その中から最も読みやすい文を選択する手法が考えられる．しかし，全ての語順を生成する手法は，文節数の増加に伴い計算量が爆発的に増加するという問題がある．そのため，何らかの制約により語順の候補を絞るというアプローチに基づいた語順整序手法が提案されている．これらの手法は，係り受け情報を使用する手法と，使用しない手法に大別され，係り受け情報を使用する手法については，係り受け情報を既知とする手法と，係り受け情報を未知とする手法に分けられる．

係り受け情報を既知とする手法 [2, 4, 9] では，日本語の構文的制約に基づき，同一の文節に係る文節

1)　日本語の係り受けは係り受け関係が交差しない，後方にしか係らない，係り先は一つであるという 3 つの制約 [7]

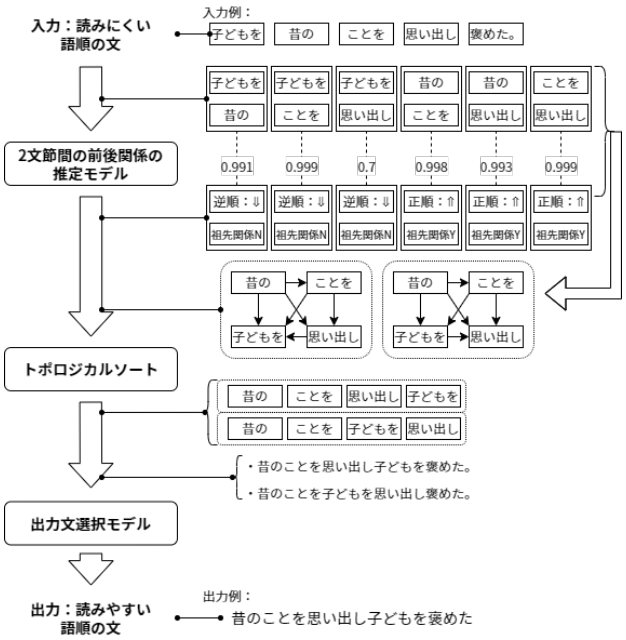


図 1　トポロジカルソートによる語順整序の概要

の中で語順の入れ替えを繰り返すことにより語順整序する．しかし，読みにくい日本語文に対しては，正しい係り受け解析結果を得にくく，語順整序の精度が低下する可能性があるという問題があった．そこで，係り受け解析・語順整序を同時に行う手法が提案された [10]．さらに，語順の変更によって読点位置も変化することに着目し，読点挿入も同時に行う手法 [3, 5, 6] も存在している．係り受け情報を使用する手法は，係り受け解析結果が正しいとすると，日本語の構文的制約を満たす語順整序結果が必ず得られるという利点があるが，係り受け解析が失敗すると，その影響を受けて，語順整序の精度も低下するという問題がある．

係り受け情報を使用せずに語順整序を行う手法 [8, 11] では，文節間の前後関係を推定し，その結果を用いて語順整序を行う．このうち，Kanouchi ら [11] は機械翻訳のための語順整序を対象としているが，孫ら [8] は，読みにくい日本語文の語順整序を対象としている．孫ら [8] の手法では，入力文の全文節対に対する前後関係の推定結果に基づいて，有向グラフを構築し，そのグラフに対してトポロジカルソートを適用することにより，語順整序する．しかし，係り受け情報を考慮していないため，構文的制約を満たさない結果が出力される可能性があること，また，2 文節間の前後関係の推定誤りに対する頑健性が低いことが課題として報告されている．

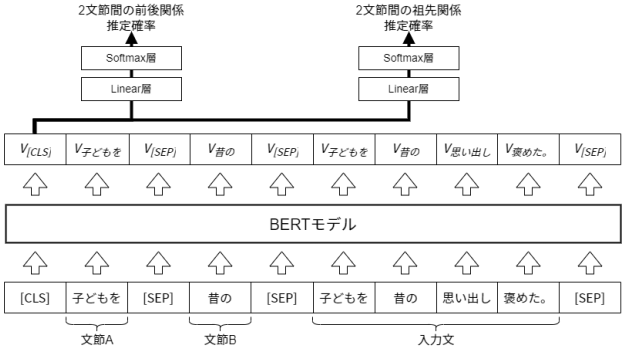


図 2　前後関係推定モデル

3　提案手法

本稿では，孫らのトポロジカルソートに基づく手法 [8] を拡張し，新たな語順整序手法を提案する．本研究の問題設定は，従来研究 [6, 8] と同一であり，意味は伝わるものの読みにくい語順を持った文の文節列を入力とし，その文節列を読みやすい語順に整える．その概要を図 1 に示す．具体的には大きく次の 3 点を拡張した．

1. 2 文節間の前後関係の推定時において，マルチタスク学習により，その 2 文節が係り受け木における祖先関係にあるか否かも同時推定するモデルを導入する．
2. トポロジカルソートの適用前において，文節間の前後関係の推定確率に基づいて複数の有向グラフを作成し，それらの各有向グラフにトポロジカルソートを適用することにより，複数の出力文候補を生成可能にする．
3. 生成された出力文候補の中から語順整序として最も適切な文を選択するモデルを組み込む．

以下では各拡張について詳述する．

3.1　2 文節間の前後関係の推定モデル

2 文節間の前後関係を推定するモデルを図 2 に示す．入力文中の文末を除いた文節集合から作成可能なあらゆる文節対の各々に対して，一方の文節（文節 A）が，もう一方の文節（文節 B）の前に来る確率を BERT により推定する．日本語では，係り受け木においての祖先となる文節は必ず後方に位置するという構文的制約が存在していることから，祖先関係と語順には強い関連が存在している．この特性を利用することで，語順整序の精度向上が期待できる．そこで，文節 A が文節 B の祖先であるか否かを推定するサブタスクを追加し，本来の前後関係推定と同

時に学習させるマルチタスク学習の枠組みを利用する．BERT の入力には，文節 A と文節 B，更に入力の 1 文全体をそれぞれサブワード分割し，特殊トークンで区切って連結したものを使用する．1 文全体を入力することで，それぞれのタスクの推定に必要な文脈情報を考慮できるようにする．各タスクの出力層は，Linear 層と Softmax 層を用い，2 値分類を行う．前後関係推定タスクの損失関数 $L_{\text{前後関係}}$ と祖先関係推定タスクの損失関数 $L_{\text{祖先関係}}$ はクロスエントロピー誤差を用い，重みを λ として，式 (1) で表される総合損失関数 L を最小化するように学習する．

$$L = \lambda L_{\text{前後関係}} + (1-\lambda) L_{\text{祖先関係}} \tag{1}$$

学習データには，先行研究 [8] と同様に，構文的制約を満たすようにランダムに語順を入れ替えた文を使用する．それに加えて，祖先関係のラベルも付与することで，マルチタスク学習に対応した学習データを構築する．

3.2 トポロジカルソート

3.1 節のモデルを用いて推定した前後関係をエッジ，各文節をノードとする有向グラフを作成する．次に，前後関係の推定確率が閾値以下のエッジ（最小のものから最大 k 個）をあらゆる組み合わせで逆向きに変更した有向グラフを新たに複数作成する．作成できる有向グラフの数は，逆向きにするエッジの数が k のときに 2^k 通りとなるため，計算量を抑えるために k を設定する．作成された各有向グラフが巡回グラフであった場合，トポロジカルソートを適用可能な有向非巡回グラフに変換する．Prabhumoye ら [12] の手法を参考にし，トポロジカルソートを行う中で，閉路が見つかるたびに閉路を構成する最後のエッジを削除することを繰り返し，非巡回グラフに変換する．最後に，各有向非巡回グラフに対して，トポロジカルソートを適用し，重複を取り除いたものを語順整序の出力候補とする．

3.3 出力文選択モデル

出力文選択モデルでは，3.2 節で生成した複数の出力文候補から語順整序として最も適切な文を選択する．本研究では，本タスクを多肢選択問題として捉え，Talmor ら [13] の手法を参考に基づいて出力文選択モデルを設計した．具体的には，入力文と各出力文候補の 2 文を特殊トークンで連結したものを BERT への入力とする．推論時には，各候補に対してモデルが出力する語順整序としての適切さを表す確率を比較し，最も値が高い候補を最終的な出力文とする．ここで，語順整序としての適切さとは，入力文と同義かつ読みやすいことであるとして，それらをモデルに学習させる．

学習データには，京大コーパス [14] の元文を正解として，不正解の選択肢を複数用意することで，多肢選択問題の形式に対応した学習データを構築する．不正解には，同義だが読みにくい文と，同義ではない文の 2 種類を用意する．同義だが読みにくい文は，3.1 節の学習データと同様に，構文的制約を満たすようにランダムに語順を入れ替えることによって作成する．同義ではない文は，元の文節を構文的制約を満たさないように語順をランダムに入れ替えた文を生成し，日本語として不自然な語順を除くために，MLM スコア [15] が元の文より十分に低い文を取り除くことで作成する．正解文を正例とするのに対し，入力文と同義であるが読みにくい文を負例にすることにより，1 文全体の日本語としての読みやすさを，また，元の文とは同義でない文を負例とすることにより入出力文間の同義性を，それぞれ学習することを期待する．

4 評価実験

実験により，提案手法の有効性を評価する．

4.1 実験概要

テストデータと開発データには，京大テキストコーパス [14] を元に人手を介して擬似的に作成された読みにくい語順の各 1,000 文（荒木ら [6]，孫ら [8] と同一）を使用した．学習データは，テストデータと開発データには含まれない文 31,798 文に対して，3.1，3.3 節で述べた方法をそれぞれ適用し，前後関係推定モデルの学習データ 1,711,265 件，出力文選択モデルの学習データ 300,850 件を構築し，使用した．

評価指標には，先行研究 [6, 8] と同様に，2 文節単位一致率（文末文節を除く文節を 2 文節ずつ取り上げ，その順序関係が元の文と一致する割合）と文単位一致率（元の文の語順と完全に一致する文の割合）を使用する．

比較として，以下を用意した．

荒木ら [6]：Shift-Reduce アルゴリズムを拡張した手法 [6].

孫ら [8]：トポロジカルソートを用いた手法 [8].

表 1　語順整序 実験結果

手法	アルゴリズム	2 文節単位	文単位
荒木ら [6]	Shift-Reduce	90.17% (28,638/31,760)	61.60% (616/1,000)
孫ら [8]	Topological	88.49% (28,105/31,760)	40.60% (406/1,000)
孫ら [8] 再現	Topological [$\lambda=1, k=0$]	95.56% (30,351/31,760)	78.50% (785/1,000)
孫ら [8] 再現 係り受け考慮	Topological [$\lambda=0.5, k=0$]	95.99% (30,485/31,760)	80.60% (806/1,000)
提案手法	Topological [$\lambda=0.5, k=10$]	96.45% (30,634/31,760)	83.30% (833/1,000)

孫ら [8] 再現：孫ら [8] の手法を提案手法と同一の環境で再現した手法．提案手法の前後関係推定モデルにおいて，$\lambda = 1$ とすることで，係り受けを考慮せず前後関係のみを推定させ，また，トポロジカルソートの適用前に有向グラフを作成する際に $k = 0$ とすることで，トポロジカルソート適用後の出力文数を 1 とし，出力文選択モデルを適用せずに出力文を確定させる．すなわち，提案手法において，2 文節間の前後関係の推定モデルの精緻化や，2 文節間の前後関係の誤りに対する頑健性向上を行っていない手法とみなすことができ，孫ら [8] の手法と同様の手法と解釈できる．

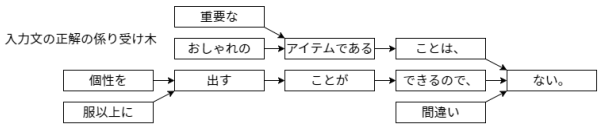
孫ら [8] 再現+係り受け考慮：提案手法において $\lambda = 0.5$ とすることで，係り受けを考慮した前後関係推定モデルの精緻化は行っているが，$k = 0$ とすることで，2 文節間の前後関係の誤りに対する頑健性向上は行っていない手法．

2 文節間の前後関係の推定モデルの構築にあたり，BERT[2]，RoBERTa[3]，ModernBERT[4] を比較検討した．その結果，開発データで最も高い精度を示した ModernBERT を採用した．学習はミニバッチ学習を用い，バッチサイズ 16，学習率 5e-6，エポック数 6，損失関数の λ は 0.5 で行った[5]．

トポロジカルソートによる語順整序において，逆向きにするエッジの最大数 k は 10 とした．また，前後関係推定モデルの出力確率の閾値については，開発データにおける softmax スコアの分布と，計算量を抑えつつ，多様な出力文を生成できるバランスを考慮し，閾値を 0.99 に設定した．

出力文選択モデルについても事前検証に基づいて

2）　tohoku-nlp/bert-base-japanese-whole-word-masking
3）　nlp-waseda/roberta-large-japanese-seq512-with-auto-jumanpp
4）　sbintuitions/modernbert-ja-310m
5）　これらハイパーパラメータは開発データで最も高い精度を示した組み合わせを採用した．なお，出力文選択モデルも同様である．



入力文	重要なおしゃれのアイテムであることは、個性を服以上に出すことができるので、間違いない。
孫ら[8]再現	重要なアイテムである服以上に個性を出すことができるのでおしゃれのことは間違いない。
提案手法	服以上に個性を出すことができるのでおしゃれの重要なアイテムであることは間違いない。

図 3　マルチタスク学習の成功例

ModernBERT を採用し，ハイパーパラメータはバッチサイズ 128，学習率 5e-6，エポック数 3 で行った．

4.2　実験結果

実験結果を表 1 に示す．提案手法は，2 文節単位一致率で 96.45%，文単位一致率で 83.30%を達成し，従来手法（荒木ら [6]，孫ら [8]，孫ら [8] 再現）を大幅に上回る結果となった．孫ら [8] 再現では失敗したものの，提案手法では正解した例を図 3 に示す．孫ら [8] 再現では，「おしゃれの」と「アイテムである」の前後関係を誤って推定したため失敗しているが，提案手法では，これらの祖先関係を捉えることによって，前後関係の推定が成功し，正解の文を出力できている．

孫ら [8] 再現+係り受け考慮は孫ら [8] 再現と比較して，2 文節単位一致率が 0.43 ポイント，文単位一致率が 2.10 ポイント向上している．係り受け情報を考慮した前後関係推定モデルの精緻化の有効性が確認できる．

また，提案手法は，孫ら [8] 再現+係り受け考慮と比較して，2 文節単位一致率で 0.46 ポイント，文単位一致率で 2.70 ポイント上回った．有向グラフを複数作成し，出力文選択モデルを導入することにより，2 文節間の前後関係に基づいたトポロジカルソートでは捉えられきれない文全体としての同義性や読みやすさを考慮できるようになったためと考えられる．これにより，提案手法において，2 文節間の前後関係の誤りに対する頑健性を向上させたことの有効性を確認できる．

以上より，提案手法の有効性を確認した，

5　おわりに

本稿では，係り受け情報を考慮したトポロジカルソートによる語順整序手法を提案した．評価実験の結果，提案手法は，従来手法を大幅に上回る精度を達成しており，その有効性を確認した．今後は，主観評価を実施する予定である．

謝辞

本研究は JSPS 科研費 JP19K12127, JP24K15076 の助成を受けたものです.

参考文献

[1] 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.

[2] 内元清貴, 村田真樹, 馬青, 関根聡, 井佐原均. 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.

[3] 大野誠寛, 吉田和史, 加藤芳秀, 松原茂樹. 係り受け解析との同時実行に基づく日本語文の語順整序. 電子情報通信学会論文誌, Vol. J99-D, No. 2, pp. 201–213, 2016.

[4] Masato Yamazoe, Tomohiro Ohno, and Shigeki Matsubara. Bottom-up Japanese Word Ordering Using BERT. In **Proceedings of the 15th International Conference on Agents and Artificial Intelligence**, Vol. 3, pp. 673–681, 2023.

[5] 宮地航太, 大野誠寛, 松原茂樹. 係り受け解析との同時実行に基づく日本語文の語順整序と読点挿入. 言語処理学会第 26 回年次大会発表論文集, pp. 243–246, 2020.

[6] 荒木駿介, 大野誠寛, 松原茂樹. 処理途中での非文生成の回避を考慮した日本語文に対する係り受け解析・語順整序・読点挿入の同時実行. 言語処理学会第 30 回年次大会発表論文集, pp. 3182–3186, 2024.

[7] 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.

[8] Peng Sun, Tomohiro Ohno, and Shigeki Matsubara. Japanese Word Reordering Based on Topological Sort. In **In Proceedings of the 15th International Conference on Agents and Artificial Intelligence**, Vol. 3, pp. 768–775, 2023.

[9] Shota Nemoto, Eiji Kamioka, Manami Kanamaru, and Phan Xuan Tan. Correcting Ambiguous Japanese Sentences Based on Japanese Dependency Analysis. In **Proceedings of the 2022 AppliedInformatics International Conference**, pp. 13–18, 2022.

[10] 吉田和史, 大野誠寛, 加藤芳秀, 松原茂樹. 係り受け解析を伴った日本語文の語順整序. 言語処理学会第 20 回年次大会発表論文集, pp. 701–704, 2014.

[11] Shin Kanouchi, Katsuhito Sudoh, and Mamoru Komachi. Neural Reordering Model Considering Phrase Translation and Word Alignment for Phrase-based Translation. In **Proceedings of the 3rd Workshop on Asian Translation**, pp. 94–103, 2016.

[12] Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. Topological Sort for Sentence Ordering. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2783–2792, 2020.

[13] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4149–4158, 2019.

[14] 黒橋禎夫, 長尾眞. 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.

[15] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked Language Model Scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, 2020.