トポロジカルソートと BERT を用いた日本語文の語順整序

孫 鵬 [1]　大野 誠寛 [2]　松原 茂樹 [1]
[1] 名古屋大学大学院情報学研究科　[2] 東京電機大学未来科学部
sun.peng.z4@s.mail.nagoya-u.ac.jp　ohno@mail.dendai.ac.jp
matsubara.shigeki.z8@f.mail.nagoya-u.ac.jp

概要

日本語では，文法的には間違っていないものの読みにくい語順を持った文が作成される場合がある．本稿では，推敲支援のための要素技術として，読みにくい語順をもった日本語文を読みやすい語順に整える手法を提案する．本手法では，BERT を用いて 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，文節を並べ替える．

1　はじめに

日本語は語順が比較的自由であるとされているが，実際には語順に関して選好が存在している．そのため，文法的には間違っていないものの読みにくい語順を持った文が作成されることがある．

そのような文の推敲支援を目的に，語順整序に関する研究がいくつか行われている [1, 2, 3, 4]．いずれも，既知の係り受け情報，あるいは，同時的に解析し得られる部分的な係り受け情報を用いている．しかし，読みにくい語順の入力文は係り受け解析の精度は低下し，その結果，語順整序の精度も低下するという問題がある．

そこで本稿では，係り受け解析を陽に施すことなく，読みにくい語順をもった日本語文を読みやすい語順に整える手法を提案する．本手法では，BERT を用いて 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，文節を並べ替える．係り受け情報を利用できなければ語順の候補を絞ることはできず，膨大な候補の中から最適な語順を探索する必要があるが，本手法ではトポロジカルソートを用いることにより効率的に探索する．

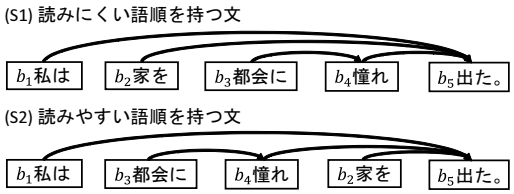

図 1　読みにくい文と読みやすい文の例

2　日本語の語順と係り受け

本研究では，日本語母語話者の推敲支援を目的とするため，文法的に間違っていないが読みにくい語順の文を入力として想定する．日本語母語話者であれば，文法的に誤った語順の日本語文を作ることは稀であるが，語順に関する選好を強く意識しなければ，読みにくい語順の文を無意識に作り出すことがあるためである．

日本語の語順に関する選好としては，格要素は基本的にガ格，二格，ヲ格の順に配置されること，また，その基本的な順序は長い従属関係など他の要因の影響を受けてしばしば変更されることなど [5, 6, 7] が知られている．図 1 の 2 文は，共に文法的に正しく，同じ意味を持っており，語順だけが異なる．ここで，四角と矢印はそれぞれ文節，係り受け関係を表す．S1 の語順は，S2 の語順に比べ，読みにくい．これは，文節の「家を」とその係り先文節「出た」の距離が長いためである [8]．このことは，読みやすく文節を並び替える際に，係り受け情報が有益であることを示唆している．

一方，文法的に正しい文は，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）[9] を満たす．そのため，語順整序の前に 1 文の係り受け構造を解析できれば，それに基づいて最適な語順の候補を絞ることができる．例えば図 1 の係り受け構造の場合，語順の候補は 6 通り (“$b_1b_2b_3b_4b_5$”, “$b_1b_3b_4b_2b_5$”, “$b_2b_1b_3b_4b_5$”, “$b_2b_3b_4b_1b_5$”, “$b_3b_4b_1b_2b_5$”, “$b_3b_4b_2b_1b_5$”) となる．

語順整序の従来手法 [1, 10, 11, 12, 13, 14, 15] では，係り受け解析が予め行われていることを前提とし，語順と係り受けの間の選好を用いて，日本語の構文的制約を満たす語順の候補の中から最も適切な語順を同定している．しかし，係り受け解析器は通常，S2 のような適切な語順を持つ文に係り受け情報を付与したコーパスを用いて学習しているため，一般に S1 に対する解析精度は S2 より低下する．

語順整序に係り受け解析を利用しないと，上記の構文的制約に基づいて語順の候補を絞ることはできず，入力文中の全文節から考えられる膨大な数の順列の中から最適な順列を選択する必要がある．入力文が n 個の文節を持つ場合，考えられる順列の数は $n!$ 個である（図 1 では, $5! = 120$）．そのため，$n!$ 個の候補の中から最適な順列を効率的に探索するアルゴリズムを採用する必要がある．加えて，文節の並び替えに有効な係り受け情報を利用できないため，係り受け情報を直接使用することなく適切な語順を予測できるモデルを採用する必要がある．

3 トポロジカルソート

トポロジカルソート [16, 17] は，有向非巡回グラフ内の全ノードを順序付けて一次元に並べるアルゴリズムである．概説すると，全てのノードは，それぞれの出力エッジの先にある隣接ノードよりも先行するように，線形に並べられる．例えば，あるノード v からノード u へのエッジ $v \rightarrow u$ を持つとき，v は u よりも前に来るように順序付けられる．

トポロジカルソートは，何らかのソートを行う自然言語処理タスクに応用されている．例えば，Prabhumoye らは，トポロジカルソートを用いた文整序手法を提案している [18, 19]．文整序とは，テキスト中の一貫性を最大化するように，その内部の文を並び替えるタスクであり，複数文書要約 [20]，調理手順生成 [21] などに応用される．彼らの研究では，テキスト内の各 2 文間の相対順序を予測し，その予測結果を各 2 文間の前後関係の制約とみなしている．この制約条件の集合を表現するために，有向非巡回グラフを作成し，それにトポロジカルソートを適用することで最適な文の順序を求める．文内の語順整序は，構成要素を並べ替えるという点で文整序と似ているため，トポロジカルソートを同様に利用できると考えられる．

トポロジカルソートの時間計算量は $O(|V| + |E|)$ である．ここで，V と E はそれぞれ，有向非巡回グラフの全ノードと全エッジの集合を意味する．トポロジカルソートを用いて日本語文を語順整序する場合，有向非巡回グラフは，文中の文節を表すノードと，文節間の前後関係を表すエッジを持つことになる．入力文が n 個の文節を持つ場合，$|V|$ は n，$|E|$ は ${}_nC_2 = \binom{n}{2} = n * (n-1)/2$ となるため，トポロジカルソートの時間計算量は $O(n^2)$ となる．したがって，トポロジカルソートは，2 節で述べた計算コストが高くなるという問題を解決することが期待できる．

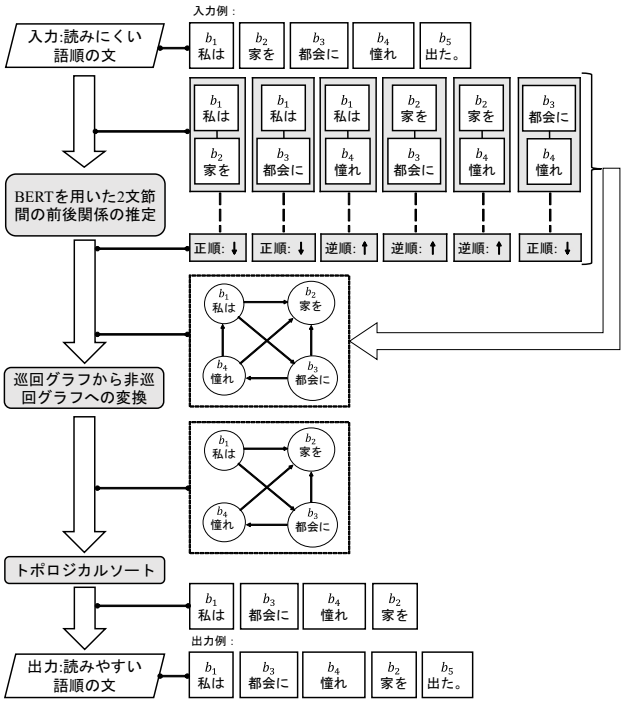

図 2 本手法の概要

4 提案手法

本手法では，文法的に誤っていないものの読みにくい語順を持った文の入力を想定し，文内の文節を読みやすく並べ替える．概要を図 2 に示す．

まず，読みにくい語順の文の文節列が入力され（図 2 では「私は/家を/都会に/憧れ/出た.」），このうち，文末文節を除いた文節集合（図 2 では，{私は, 家を, 都会に, 憧れ}）から，あらゆる 2 文節の組み合せを取り出し，その 2 文節間の前後関係を BERT を用いて推定する[1)]．次に，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し，それが有向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．

1) なお，日本語の場合，文法的に誤っていない文が入力されたならば，その入力文の文末文節は，語順整序後の文でも必ず文末となるため，並べ替える対象から外している．

変換手法は，Prabhumoye らの手法 [18] のソースコードを参照したものとなっており，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（探索済みのノードに再び戻ってくるエッジ）を削除するということを閉路が存在しなくなるまで繰り返す．最後に，上記で作成された有向非巡回グラフに対してトポロジカルソート[2]を適用し，各ノード（各文節）を順に並べる．

4.1 2 文節間の前後関係の推定

2 文節間の前後関係の推定する BERT モデルの概要を図 3 に示す．入力文の文節列を $B = b_1 b_2 \cdots b_n$ とするとき，文末文節 b_n を除いた文節集合 $\{b_1, b_2, \cdots, b_{n-1}\}$ から取り出した 2 文節を $b_i, b_j\ (1 \leq i < j \leq n-1)$ とする．このとき，BERT への入力は，“[CLS] b_i [SEP] b_j [SEP] $b_1 b_2 \cdots b_n$ [SEP]” として，サブワード分割を施したものとする．

ここで入力文の語順は，文法的には誤っていないため，部分的には適切な語順，すなわち，多くの文節の相対的な位置は，語順整序後も入力時と変わらず，その情報を入力文は保持していると考えられる．また入力文には，語順整序に有益な文節間の係り受け情報が暗黙的に含まれていると考えられる．以上より，本手法では，BERT モデルへの入力に入力文全体 $b_1 b_2 \cdots b_n$ を加えた．

BERT の出力は，b_i が b_j よりも文頭側にある（入力文の語順の正順である）方が読みやすい確率と，文末側にある（入力文の語順とは逆順である）方が読みやすい確率の 2 値である．高い確率が与えられた前後関係を推定結果とする．

4.2 学習データの作成

BERT モデルの学習データとして，どのようなデータを使用するかも重要となる．単純には，新聞記事文などの読みやすい文に含まれる文節間の前後関係を使用することが考えらえる．このアプローチでは，1 文中から 2 文節 b_i と $b_j (i < j)$ を取り出したときに，その文の語順と同一の前後関係 “$b_i \rightarrow b_j$” に「正順」，その逆の前後関係 “$b_i \leftarrow b_j$” に「逆順」のラベルを付与して学習データを作る．その結果，正例「正順」と負例「逆順」は同数となる．しかし，人間が実際に作文した読みにくい文に，このような傾向があるとは考えにくい．

一方，人間が実際に作った読みにくい文の傾向を学習するためには，それらの文から 2 文節の組合せを取り出し，各々に人手でラベルを付与する必要があり，そのコストは非常に高い．

そこで，新聞記事文は読みやすい語順で書かれているという前提で，その各文から読みにくい語順の擬似文を機械的に作成することにした．具体的には，新聞記事文から読点を取り除いた上で，2 節で述べた構文的制約を保持させつつ，ランダムに語順を変更して作成した．このようにして作成した擬似文は，構文的制約を満たすため，文法的な正しさが維持され，新聞記事文の語順とは異なるため，読みにくい語順となっていることが想定できる．

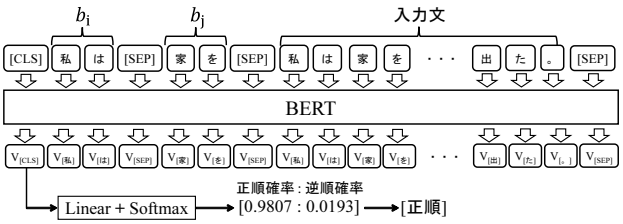

図 3 2 文節間の前後関係の推定例

5 評価実験

本手法の有効性を確認するため，語順整序実験を実施した．新聞記事中の文から人手で擬似的に作成した読みにくい語順の文に対して本手法を適用し，元の文の語順をどの程度再現できるかを測定した．

5.1 実験概要

テストデータ及び開発データには，宮地ら [3] と同じ手順で，京大テキストコーパス [22] をもとに人手を介して疑似的に作成された読みにくい語順の文 1,000 文をそれぞれ用いた．学習データには，4.2 節で作成した 34,199 文を用いた[3]．

評価では，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）[5] 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）[2] を測定した．

比較のため，以下の 4 つの手法を用意した．

[random]: 文末以外の語順をランダムに変更する．

[no reordering]: 入力文をそのまま出力する．

[Dep+TS]: 係り受け解析結果をもとに，構文的制約だけでは定まらない語順に対してのみトポロジカル

2） 深さ優先探索に基づくトポロジカルソート [17] を用いた．

3） 34,199 文のうち，新聞記事文と異なる順序を持つ文は 27,263 文，偶然に同じ順序を持つ文は 6,936 文であった．また，学習データの元となった新聞記事文は，テストデータや開発データの元となったものとはすべて異なる．

表 1　実験結果

	2 文節単位一致率	文単位一致率
本手法	88.49% (28,105/31,760)	40.60% (406/1,000)
[random]	50.59% (16,070/31,760)	4.60% (46/1,000)
[no reordering]	75.48% (23,973/31,760)	0.00% (0/1,000)
[Dep+TS]	86.42% (27,448/31,760)	36.10% (361/1,000)
[NLM]	75.65% (24,025/31,760)	11.50% (115/1,000)

入力：　一つには どう 総理が 率いる 社会党が なるか。
本手法（正解）：一つには 総理が 率いる 社会党が どう なるか。
[Dep+TS]：　一つには 総理が どう 率いる 社会党が なるか。

図 4　本手法が正解し，[Dep+TS] が不正解だった例

ソートを用いた語順整序を行う．まず入力文に係り受け解析を施し，その結果をもとに，直接的に係り受け関係にある 2 文節間の前後関係を構文的制約から決定する．直接的に係り受け関係にない 2 文節間の前後関係は，本手法の BERT モデルにより推定する．以上によって定めた全 2 文節間の前後関係から構築した有向グラフに対して，本手法と同じ手順でトポロジカルソートを適用し，語順整序を行う．なお，係り受け解析器には CaboCha[23] を用いた．

[NLM]: 係り受け解析を施すことなく，言語モデルを用いて語順整序を行う．具体的には，まず，入力文に対して，文末を除く 2 文節のみを入れ替えたあらゆる語順を考え，それらを候補文とする[4]．次に，これらの候補文と入力文の各文に対して言語モデルを適用しスコアを算出し，そのスコアが最大となる文を読みやすい文として出力する．言語モデルのスコアは各トークンの予測確率の対数の和とする．なお，言語モデルには日本語 GPT2[5] を用いた．

5.2　実験結果

語順整序結果を表 1 に示す．2 文節単位と文単位の両指標において本手法が最も高い一致率を達成しており，その有効性を確認した．図 4 に，本手法が正解し，次点の [Dep+TS] では不正解だった例を示す．本手法の出力語順は，正解の語順と 1 文全体で一致しているのに対し，[Dep+TS] では，「どう」が「率いる」に係ると誤って解析し，その誤りを引きずり「どう」の語順の同定に失敗している．

4）　候補文の数は，入力文が n 文節からなる場合，${}_{n-1}C_2$ となる．係り受け解析を施さないため，単純にあらゆる順列を考えると，2 節で述べた通り，$n!$ の候補文ができ，計算量が膨大となる．これを避けるため，比較手法では，入力文も部分的には適切な語順を持っていると想定し（4.2 節参照），単純に 1 組の 2 文節だけを入れ替えたものを候補文する．

5）　https://github.com/tanreinama/gpt2-japanese

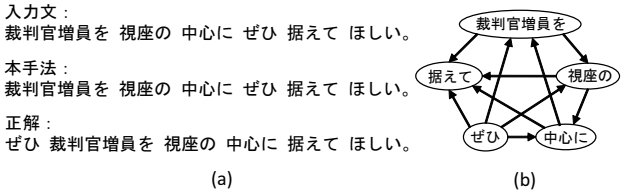

図 5　本手法の不正解例

表 2　巡回/非巡回に分けた再評価（各一致率）

		本手法	[no reordering]
2 文節単位	巡回	85.34% (13,264/15,542)	75.21% (11,689/15,542)
	非巡回	91.51% (14,841/16,218)	75.74% (12,284/16,218)
文単位	巡回	18.42% (56/304)	0.00% (0/304)
	非巡回	50.29% (350/696)	0.00% (0/696)

5.3　エラー分析

図 5(a) に本手法の不正解例を示す．この例では、文節「裁判官増員を」と「中心に」の前後関係の推定に失敗し，図 5(b) の有向巡回グラフが当初作られていた．このように前後関係の推定結果から当初は巡回グラフとなったものに不正解が多く見られた．

そこで，本手法の BERT による推定結果から作成した有向グラフが，巡回グラフとなった文（次の手順で非巡回グラフに変換された文）と，もともと非巡回グラフとなった文とに分けて再評価した．

結果を表 2 に示す．[no reordering] の数値は，本手法により当初作成された有向グラフに基づいて各文を上述の 2 クラスに分類し，当該手法を適用した結果である．両クラスの間で一致率に大きな差はない．一方，本手法は，非巡回グラフとなった文に対する両一致率が巡回グラフとなった文に対するものよりも大きく上回っている．巡回グラフとなった文は，2 文節間の前後関係の推定結果に一部誤りが含まれることにより巡回グラフとなるため，たとえ，非巡回グラフに変換したとしても，その誤りのエッジが含まれたままであれば，適切な語順を生成できないと考えられる．2 文節間の前後関係の推定精度の向上，あるいは，巡回グラフの中の誤ったエッジの適切な除去に基づく非巡回グラフへの変換が求められるが，これらは今後の課題である．

6　おわりに

本稿では，係り受け解析を陽に施すことなく，読みにくい文を語順整序する手法を提案した．本手法では，BERT を用いて 2 文節間の前後関係を推定し，その結果に対してトポロジカルソートを適用する．評価実験の結果，本手法の有効性を確認した．

謝辞

本研究は，一部，科学研究費補助金基盤研究 (C) No. 19K12127 により実施した．

参考文献

[1] 横林博, 菅沼明, 谷口倫一郎ほか. 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. 情報処理学会論文誌, Vol. 45, No. 5, pp. 1451–1459, 2004.

[2] 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 D, Vol. 99, No. 2, pp. 201–213, 2016.

[3] 宮地航太, 大野誠寛, 松原茂樹. 読みにくい語順の文への読点の自動挿入. 言語処理学会第 25 回年次大会発表論文集, pp. 1308–1311, 2020.

[4] 宮地航太, 大野誠寛, 松原茂樹. 文末からのトップダウン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 27 回年次大会発表論文集, pp. 1840–1845, 2021.

[5] 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, 2008.

[6] Tatsuki Kuribayashi, Takumi Ito, Jun Suzuki, and Kentaro Inui. Language models as an alternative evaluator of word order hypotheses: A case study in Japanese. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 488–504, 2020.

[7] Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, and Kentaro Inui. Topicalization in language models: A case study on Japanese. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 851–862, 2022.

[8] Kazushi Yoshida, Tomohiro Ohno, Yoshihide Kato, and Shigeki. Matsubara. Japanese word reordering integrated with dependency parsing. In **Proceedings of the 25th International Conference on Computational Linguistics**, pp. 1186–1196, 2014.

[9] 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.

[10] Anja Belz and Eric Kow. Discrete vs. continuous rating scales for language evaluation in nlp. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 230–235, 2011.

[11] Katja Filippova and Michael Strube. Generating constituent order in german clauses. In **Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics**, pp. 320–327, 2007.

[12] Karin Harbusch, Gerard Kempen, Camiel Van Breugel, and Ulrich Koch. A generation-oriented workbench for performance grammar: Capturing linear order variability in german and dutch. In **Proceedings of the 4th International Natural Language Generation Conference**, pp. 9–11, 2006.

[13] Geert-Jan M. Kruijff, Ivana Kruijff-Korbayová, John Bateman, and Elke Teich. Linear order as higher-level decision: Information structure in strategic and tactical generation. In **Proceedings of the 8th European Workshop on Natural Language Generation**, pp. 74–83, 2001.

[14] Eric Ringger, Michael Gamon, Robert C. Moore, David M. Rojas, Martine Smets, and Simon Corston-Oliver. Linguistically informed statistical models of constituent structure for ordering in sentence realization. In **Proceedings of the 20th International Conference on Computational Linguistics**, pp. 673–679, 2004.

[15] James Shaw and Vasileios Hatzivassiloglou. Ordering among premodifiers. In **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics**, pp. 135–143, 1999.

[16] Arthur B. Kahn. Topological sorting of large networks. **Communications of the ACM**, Vol. 5, No. 11, pp. 558–562, 1962.

[17] Robert Endre Tarjan. Edge-disjoint spanning trees and depth-first search. **Acta Informatica**, Vol. 6, No. 2, p. 171–185, 1976.

[18] 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.

[19] Vishal Keswani and Harsh Jhamtani. Formulating neural sentence ordering as the asymmetric traveling salesman problem. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 128–139, 2021.

[20] Regina Barzilay and Noemie Elhadad. Inferring strategies for sentence ordering in multidocument news summarization. **Journal of Artificial Intelligence Research**, Vol. 17, No. 1, pp. 34–35, 2002.

[21] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In **Proceedings of the 31st AAAI Conference on Artificial Intelligence**, pp. 3075–3081, 2017.

[22] Sadao Kuroashi and Makoto Nagao. Building a Japanese parsed corpus while improving the parsing system. In **Proceedings of the 1st International Conference on Language Resources and Evaluation**, pp. 719–724, 1998.

[23] 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. Vol. 43, No. 6, pp. 1834–1842, 2002.