E-033

トポロジカルソートに基づく日本語文の語順整序
Japanese Word Reordering based on Topological Sort

孫 鵬† 大野 誠寛‡ 松原 茂樹†
Peng Sun Tomohiro Ohno Shigeki Matsubara

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．例えば，以下の 2 つの例文

例文 1 私は家を都会に憧れ出た．

例文 2 私は都会に憧れ家を出た．

では，例文 1 はそのままでは読みにくいが，例文 2 のように文節を並べ替えることにより読みやすくなる[1]．

文を読みやすい語順に整えるという語順整序は，推敲支援や文生成などに応用でき，これまでにも幾つか研究されている[2]~[6]．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に基づいている．しかし，入力文が読みにくい語順である場合，係り受け解析の精度は低下する傾向にあり，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．

2. トポロジカルソート

トポロジカルソートは，有向非巡回グラフ内の全ノードを順序を付けて一次元に並べるアルゴリズムである[7]．あるノード u から v へのエッジ $u \rightarrow v$ を持つとき，u は v よりも前にくるように順序付けられる．トポロジカルソートの自然言語処理タスクへの応用として，Prabhumoy ら[8]の文整序を行う研究がある．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[9]，複数文書要約[10,11]や料理手順生成[12]などに応用される．Prabhumoy らの研究[8]では，文書内のあらゆる 2 文間の前後関係を推定し，文をノード，その前後関係をエッジとする有向グラフに対して，トポロジカルソートを適用し，文書中のすべて文を一次元に並べる手法が提案されている．

本研究で取り組む語順整序タスクは，文内の文節を並べる問題であるため，文整序タスクと同様に，トポロジカルソートを活用することができると考えられる．

3. トポロジカルソートに基づく語順整序

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が入力されることを想定し，その文内の文節を読みやすく並べ替える．本手法の概要を図 1 に示す．

まず，読みにくい語順の文の文節列が入力され（図 1 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 1 では，｛私は，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．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．

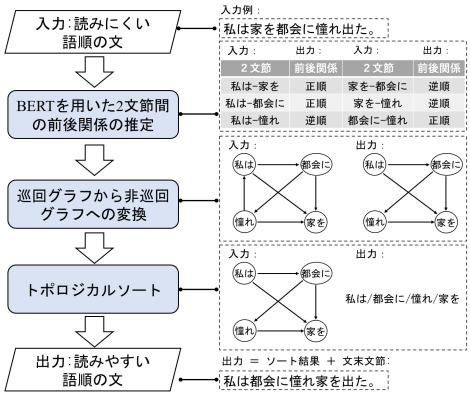

図 1 本手法の概要

次に，その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．変換手法は，Prabhumoye らの手法[8]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というものである．

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を適用し，各ノード（各文節）を順に並べる．なお，本研究では，深さ優先探索に基づいたトポロジカルソート[7]を用いる．

3.1 2 文節間の前後関係を推定する BERT モデル

2 文節間の前後関係の推定する BERT モデルの概要を図 2 に示す．入力文の文節列を $B = b_1 \cdots b_n$ とするとき，文末文節 b_n を除いた文節集合 $\{b_1, \cdots, b_{n-1}\}$ 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_n［SEP］"として，サブワード分割を施したものとする．

ここで文末文節は，例え読みにくくても文法的に誤っていない文であるならば，その文の主節の述語であり，係り受け構造を表した木の根にあたる．2 文節間の前後関係は，それらの係り先文節との関係で決まる可能性が示唆されるため，文末文節の情報は，2 文節間の前後関係の推定に寄与すると考え，BERT への入力に文末文節を加えた．

BERT の出力は，文節 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）ほうが読みやすい確率と，文節

†名古屋大学 Nagoya University
‡東京電機大学 Tokyo Denki University

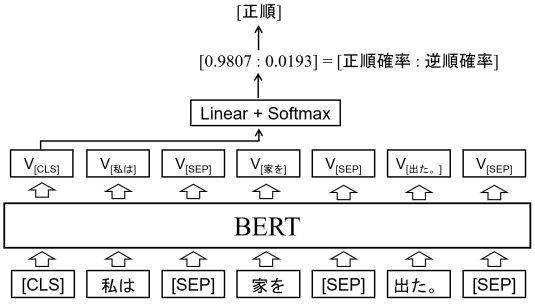

図 2　2 文節間の前後関係の推定例

b_i が文節 b_j よりも文末側にある（入力文の語順とは逆順である）ほう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.　評価実験

本手法の実現可能性を評価するため，新聞記事て語順整序実験を実施した．新聞記事中の文から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．

4.1 実験概要

評価用データには，宮地ら[5]と同じ手順で作成された読みにくい語順の文データ 1,000 文を用いた．学習には，京大テキストコーパス Ver.4.0 のうち，評価用データの作成に用いた文を除く 35,404 文を用いた．

評価では，2 文節単位一致率（2 つずつ文節を取り上げ，その順序関係が元の文と一致しているものの割合）[2]と文単位一致率（元の文の語順と完全に一致している文の割合）[4]を測定した．

モデルは Pytorch を用いて実装した．学習アルゴリズムは AdamW を採用した．パラメータの更新はミニバッチ学習(学習率 1e-6，バッチサイズ 16)により行い，エポック数は 3 とした．BERT は東北大学が公開している事前学習済み BERT モデル [1] を用いてファインチューニングを行った．

比較のために，以下の 2 つのベースラインを設けた．

- **ランダム**:文節単位で語順をランダムに変更する．
- **語順整序前**：評価用データの語順(語順整序前の語順）をそのまま出力する．

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．文単位一致率では，本手法が最も高い一致率となった．本手法による語順整序の成功例を図 3 に示す．図 3 では，本手法の出力語順が正解の語順と 1 文全体で一致しており，このような例が 175 文存在していた．

一方，2 文節単位一致率では，本手法は語順整序前よりわずかに下回った．ここで，BERT による 2 文節間の前後関係の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．その結果のうち，2 文節単位一致率を表 2 に示す．非巡回グラフとなった文に対して，本手法は語順整序前より 10%以上，上回っていることが分かる．なお，本手法の文単位一致率は，非巡回グラフの文は 31.00% (155/500)，巡回グラフの文は 4.00% (20/500)であった．

以上より，本手法の実現可能性を確認した．

1)　https://github.com/cl-tohoku/bert-japanese

表 1　実験結果

	2 文節単位一致率	文単位一致率
本手法	74.80% (23,758/31,760)	17.50% (175/1,000)
ランダム	50.59% (16,070/31,760)	4.60% (46/1,000)
語順整序前	75.48% (23,973/31,760)	0.00% (0/1,000)

表 2　語順整序前後の 2 文節単位一致率

	非巡回グラフ	巡回グラフ
本手法	78.43% (5,625/7,172)	73.75% (18,133/24,588)
語順整序前	68.11% (4,885/7,172)	77.63% (19,088/24,588)

【入力文】
外圧をこれらの業界は比較的受けにくくまた政治的発言力が強いという特徴がある。
【本手法の語順整序結果（正解）】
これらの業界は比較的外圧を受けにくくまた政治的発言力が強いという特徴がある。

図 3 語順整序結果の成功例

5.　おわりに

本論文では，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，その結果の有向グラフに対してトポロジカルソートを適用する．読みにくい語順の文データを用いた評価実験の結果，本手法の実現可能性を確認した．今後は，2 文節間の前後関係の推定モデルや巡回グラフから非巡回グラフへの変換アルゴリズムを見直すことにより，精度向上を図りたい．

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] 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] 内元ら, “コーパスからの語順の学習,” 自然言語処理, 7(4), pp.163–180, 2008.
[3] 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, 45(5), pp.1451–1459, 2004.
[4] 大野ら, “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,” 信学論, J99-D(2), pp.201–213, 2016.
[5] 宮地ら, “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.
[6] 宮地ら, “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.
[7] R. E. Tarjan, “Edge-disjoint spanning trees and depth-first search,” Acta Informatica, 6(2), pp.171–185, 1976.
[8] S. Prabhumoye et al., “Topological sort for sentence ordering,” Proc. ACL2020, pp.2783–2792, 2020.
[9] R. Barzilay, L. Mirella, “Modeling local coherence: An entity-based approach,” Computational Linguistics, 34(1), pp.1–34,2008.
[10]R. Barzilay, N. Elhadad, “Inferring strategies for sentence ordering in multidocument news summarization,” J. Artif. Intell. Res., 17(1), pp.35-55, 2002.
[11]R. Nallapati et al., “Summarunner: A recurrent neural network based sequence model for extractive summarization of documents,” Proc. AAAI2017, pp.3075–3081, 2017.
[12]K. Chandu et al., “Storyboarding of recipes: grounded contextual generation,” Proc. ACL2019, pp.6040–6046, 2019.