

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

荒木 駿介^{†, a)} 大野 誠寛^{†, b)} 松原 茂樹[‡]
 東京電機大学未来科学部[†] 名古屋大学情報連携推進本部[‡]

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、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無意識に作成される。また読点についても同様に、その有無や位置によっては文が読みにくくなる。このような読みにくい文は、語順を整え、適切に読点を挿入し直せば読みやすくなる。

語順整序や読点挿入に関する研究は、推敲支援や文生成などに応用でき、いくつか行われている（例えば、[1, 2, 3]）。その中でも宮地ら [3] は、Shift-Reduce アルゴリズムを拡張することにより、係り受け解析・語順整序・読点挿入を同時実行する手法を提案している。語順や読点と係り受けとの相互依存的な関係に着目した有望な手法であると考えられるものの、Shift などの操作を選択する機械学習モデルやその学習データに対する検討は十分ではなく、改善の余地がある。

そこで本稿では、宮地らの手法 [3] に BERT[4] を組み込むことにより、係り受け解析・語順整序・読点挿入を更に高精度に同時実行することを試みる。

2 先行研究

宮地ら [3] は、Shift-Reduce アルゴリズムにおいて、Shift と Reduce の他に、読点挿入のための Shift-Comma と Reduce-Comma、語順変更のための Swap という3つの操作を追加するとともに、キューやスタックの他に、語順変更のためのスタック（以下、語順変更スタック）を新たに用意することにより、係り受け解析・語順整序・読点挿入の同時実行を実現している。

宮地らの手法 [3] の動作例を図1に示す。入力文「五十歳を今年は迎える。」に対して、係り受け解析と語順整序、読点挿入を同時に施し、読みやすい文「今年は、五十歳を迎える。」に整形する過程が示されている。まず入力文の文節列^{*1}がキューに格納され、時刻1で Shift により、キューの先頭「五十歳を」がスタックにプッシュされる。次に時刻2で Swap により、スタックのトップ「五十歳を」とキューの先頭「今年は」がこの順で語順変更スタックにプッシュされる。時刻3で Shift-Comma により、語順変更スタックのトップ「今年は」が読点を付与されスタックにプッシュされる。なお、語順変更スタックの要素はキューの要素より優先的に操作対象となる。時刻4で「五十歳を」が Shift される。時刻5で Reduce により、スタックの「今年は、」と「五十歳を」がこの語順でキューの先頭「迎える。」

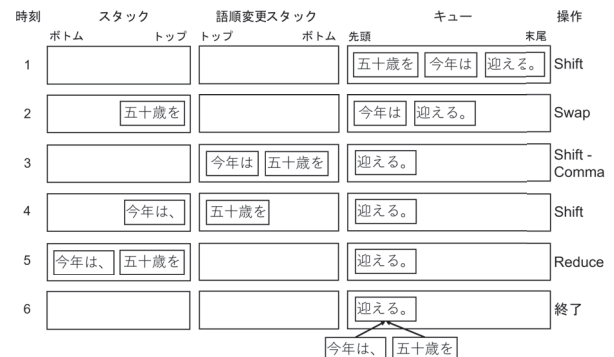


図1: 宮地らの手法の動作例

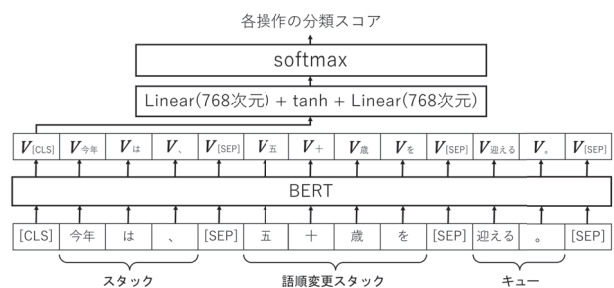


図2: 提案手法の操作選択モデル（図1の時刻4）

に係る^{*2}という係り受け木が構成され、終了する。

図1では、説明のため、各時刻で最適な操作が選択されているが、実際には機械的に選択する必要がある。宮地らの手法では、5つの各操作を仮に実行した各結果に対して、その読みやすさや係り受けの妥当性を推定し、その値が最も高くなる操作を選択している。その推定には、読みやすい文のコーパスで学習した勾配ブースティングマシン (GBM) を用いているが、操作選択の精度は十分ではない。

3 提案手法

提案手法では、図1で説明した宮地らのアルゴリズム [3] をそのまま採用し、各時刻での操作選択において、BERTを用いた新たな方法を導入する。具体的には、スタックと語順変更スタック、キューの各状態の3つ組から成る計算状況を分類対象、その計算状況に対する操作を分類クラスとする分類問題として捉え、各時刻においてBERTを用いたモデルにより操作を選択する。なお、本研究の問題設定は、宮地ら [3] と同様であり、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が入力されることを想定する。

3.1 BERTを用いた操作選択モデル

提案手法の操作選択モデル（図1の時刻4での計算例）を図2に示す。スタック・語順変更スタック・キュー

^{*2}入力文の語順と構文的制約 [5] から、1回の Reduce によって複数の文節の係り先が一度に決定されることもある。

Simultaneous Execution of Dependency Parsing, Word Reordering and Comma Insertion for Japanese Sentences

Shunsuke Araki^{†, a)}, Tomohiro Ohno^{†, b)}, Shigeki Matsubara[‡]

[†] School of Science and Technology for Future Life, Tokyo Denki University.

[‡] Information and Communications, Nagoya University

a) 19f008@ms.dendai.ac.jp

b) ohno@mail.dendai.ac.jp

^{*1}入力文にある読点はすべて事前に削除される。

表 1: 語順整序の正解率

手法	2 文節単位	文単位
宮地ら	64.50%	9.60%
[3]	(20,486/31,760)	(96/1,000)
提案	86.11%	48.90%
手法	(27,350/31,760)	(489/1,000)

表 2: 読点挿入の再現率・適合率・F 値

手法	再現率	適合率	F 値
宮地ら	36.17%	58.62%	44.74
[3]	(17/47)	(17/29)	
提案	70.47%	83.28%	76.34
手法	(284/403)	(284/341)	

表 3: 係り受け解析の正解率

手法	係り受け単位	文単位
宮地ら	47.20%	11.40%
[3]	(3,581/7,587)	(114/1,000)
提案	84.55%	49.60%
手法	(6,415/7,587)	(496/1,000)

に格納された各文節列を結合し、先頭に [CLS] を、各々の後に [SEP] を付与したうえで、サブワード分割を施したものを各時刻での計算状況とみなし、BERT に入力する。BERT の出力のうち、[CLS] に対応する出力のみを取り出し、2 層の Linear 層と Softmax を介し分類スコアを得て、分類スコアが最も高い操作を選択する。

3.2 学習データの作成

前節で述べたモデルを学習するには、各時刻での計算状況と、それに対する適切な操作の組が大量に必要となる。このような学習データを構築するには、読みにくい文と、それを読みやすく整形した文に対して係り受け構造を付与したデータが大量に必要となるが、そのようなデータはない。そこで本研究では、係り受け構造が付与された読みやすいと想定される文を収録したコーパスを元に、疑似的な読みにくい文を機械的に生成し、それを元の読みやすい文に戻す過程を自動生成し、計算状況と適切な操作の組を大量に構築した。

4 評価実験

提案手法の有効性を確認するために、新聞記事文から疑似的に作成した読みにくい文を用いた実験を行った。

4.1 実験概要

テストデータには、京大テキストコーパス [6] を元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（宮地ら [3] と同一）を用いた。学習データには、京大テキストコーパス Ver.4.0[6] の 32,506 文^{*3}に対して、3.2 節の手順を適用して得られる計算状況と適切な操作の組 1,767,470 件を用いた。学習データとテストデータの間で、元となった新聞記事文に重複はない。

評価では、宮地らの手法 [3] と比較するため、彼らと同じ指標を用いた。語順整序では、2 文節単位正解率（文末文節以外の文節を 2 つずつ取り上げたとき、それらの前後関係が正解と一致している割合）と文単位正解率（正解の語順と完全一致している文の割合）を測定した。読点挿入では、語順が正解と完全一致している文のみを対象に、読点位置に関する再現率、適合率、F 値を測定した。係り受け解析では、係り受け正解率（文末文節以外の全文節のうち、正解と係り先が一致している文節の割合）と文単位正解率（正解の係り受け構造と完全一致している文の割合）を測定した。

モデルは PyTorch を用いて実装し、BERT の事前学習モデルには、東北大学が公開しているモデル^{*4}を用いた。学習はミニバッチ学習（バッチサイズ 16、学習率 $1e-5$ ）で行った。エポック数は 1 とした。

^{*3}構文的制約から 1 通りの語順しか考えられない文や、構文的制約を満たさない文は事前に排除した。

^{*4}cl-tohoku/bert-base-japanese-whole-word-masking

入力

一大観光都市として	脚光を、	中国からの	流民も	増え	浴びていた。
-----------	------	-------	-----	----	--------

宮地ら[3]の出力
(不正解)

一大観光都市として、	中国からの	流民も	脚光を	増え、	浴びていた。
------------	-------	-----	-----	-----	--------

提案手法の出力
(正解)

中国からの	流民も	増え、	一大観光都市として	脚光を	浴びていた。
-------	-----	-----	-----------	-----	--------

図 3: 提案手法の成功例

4.2 実験結果

実験結果を表 1 から表 3 に示す。提案手法はすべての評価指標において宮地らの手法 [3] を大幅に上回っており、提案手法の有効性を確認した。

図 3 に、宮地らの手法では不正解だったが、提案手法では正解した例を示す。宮地らの手法は、「一大観光都市として」と「脚光を」が「増え」に係ると誤って解析しており、それに伴って、これらの文節に関わる語順整序や読点挿入に失敗している。入力文が読みにくい理由は、「中国からの流民も増え」が埋め込み節になっており、「一大観光都市として」の係り先が遠く離れているためであり、その長距離の係り受け関係の解析が難しく失敗したと考えられる。一方、提案手法は、埋め込み節を含む文に対しても頑健に動作し、元文の語順・読点位置・係り受け構造を正しく復元できている。提案手法は、入力文全体の情報を BERT に入力しており、遠く離れた係り先となりうる文節の情報も考慮して、操作を選択できているものと考えられる。

5 おわりに

本稿では、Shift-Reduce 係り受け解析を拡張した係り受け解析・語順整序・読点挿入の同時実行手法において、操作の選択に BERT を用いる手法を提案し、その有効性を確認した。今後は、操作選択モデルの精緻化などにより、更なる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] 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, 2000.
- [2] 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.
- [3] 宮地ら, “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と読点挿入,” NLP2020 発表論文集, pp. 243-246, 2020.
- [4] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proc. ACL2018, pp. 4171-4186, 2019.
- [5] 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, 40(9), pp. 3397-3407, 1999.
- [6] 黒橋, 長尾, “京都大学テキストコーパス・プロジェクト,” NLP1997 発表論文集, pp. 115-118, 1997.