E-036

言語モデルと構文情報を用いた日本語文のボトムアップ語順整序
Bottom-up Japanese Word Reordering using Language Model and Syntactic Information

山添 壮登[†,a)]　大野 誠寛[†,b)]　松原 茂樹[‡]
Masato Yamazoe　Tomohiro Ohno　Shigeki Matsubara

1　はじめに

日本語の語順は比較的自由であるものの，選好が存在する．そのため文法的に誤りではないものの，読みにくい語順を持った文が生成される場合がある [5]．読みやすい語順を生成する技術は，機械翻訳や文生成にとって重要な要素技術となる．

語順整序に関する研究は，文章推敲支援や文生成への応用を目的にこれまでに幾つか行われている [1]〜[4]．中でも高須ら [3] は，RNNLM による言語モデルと，構文情報を考慮した SVM によるモデルを共に用いた手法を提案し，それぞれを単独で用いた場合と比べ，高精度な語順整序を実現している．しかし高須らは，同じ文節に係るもの同士でまとめた各文節集合の中で語順整序を行っているにすぎず，1 文全体の語順整序は行っていない．

そこで本稿では，高須らの手法 [3] に基づいて，1 文全体をボトムアップに語順整序する手法を提案する．提案手法では，RNNLM でスコアを計算する際に，1 文全体の文節の並びを考慮できる．また本稿では，語順整序において RNNLM を併用することの影響を考察する．

2　先行研究

内元ら [4] は，1 文の係り受け構造は既知であるとして，任意の受け文節 b_r に係る文節の集合 $B_r = \{b_1, b_2, \ldots, b_n\}$（$n \geq 2$ に限る）に対して，B_r から考えられる順列 $\mathbf{b}^k (1 \leq k \leq n!)$ の中で最も読みやすい順列を求める問題として語順整序を定義している．内元ら [4] は，構文情報を中心とした素性に基づき機械学習したモデルを用いて語順整序する手法を提案している．

高須ら [3] は，内元ら [4] の問題設定を引継ぎ，$\mathbf{b}^k$ に対して，内元らと同一の構文情報を含む素性に基づいた SVM によるモデルだけでなく，RNNLM による言語モデルによっても語順の読みやすさを示すスコアを計算し，それらを混合比率 α により混合させたスコア $S(\mathbf{b}) = \alpha S_{RNNLM}(\mathbf{b}) + (1-\alpha) S_{SVM}(\mathbf{b})$ が最も大きくなる順列 $argmax_{\mathbf{b} \in \{\mathbf{b}^k | 1 \leq k \leq n!\}} S(\mathbf{b})$ を出力している．

上述の両研究とも，B_r のみを語順整序の対象にしており，実際には 1 文全体の語順整序を行っていない．また，そのため，高須らの手法 [3] では，各文節を修飾する文節列を考慮せず，B_r 内の各文節を単に並べ替えた単語列に対して RNNLM を適用しスコアを計算している．

3　提案手法

提案手法では，1 文全体を構成する文節集合と，その係り受け構造を入力とし，入力文節集合内の文節を読みやすく並べたものを出力する．その際，1 文の文節集合とその係り受け構造を表す構文木を作成し，その木に対してボトムアップに処理を施し，1 文全体を語順整序する．具体的には以下の手順で語順整序を行う．

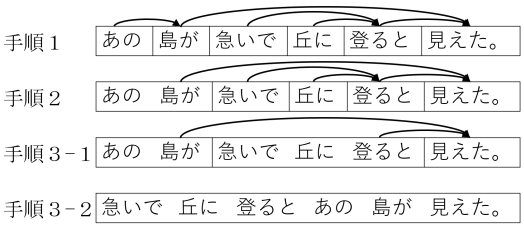

図 1　ボトムアップな語順整序の例

1. 入力の文節集合と，その係り受け構造を表す構文木を作る．具体的には，入力文の各文節を 1 つのノードとして配置し，係り受け関係を表すエッジを用いて．それらの間を結び，構文木を作る．なお，以下の手順 2 と手順 3 において，複数のノードが 1 つのノードにまとめ上げられる操作がある．したがって，ノードは文節列（長さ 1 も含む）を表し，エッジは子ノード（の最終文節）が親ノードに係る係り受け関係を表すものとする．

2. 葉ノードのみを子に 1 つもつ親ノードと，その子ノードとをまとめ上げ 1 つのノードにする．その際，係り受けの後方修飾性を考慮し，子と親とをこの順に連接した文節列を新たなノードとする．この手順は，葉ノードのみを子に 1 つもつ親ノードがなくなるまで繰り返す．

3. 葉ノードのみを子に複数もつ親ノードと，その子ノード集合とをまとめ上げ 1 つのノードにする．その際，子ノード集合内の適切な語順を高須らの手法 [3] を準用し求め，その語順で連接した文節列の後ろに親ノードを繋げた文節列を新たなノードとする．

4. 手順 2 と手順 3 を，構文木が根ノード 1 つになるまで繰り返す．

上記手順の具体例を図 1 に示す．まず手順 1 では，図 1 最上部の構文木を作る．手順 2 では，『あの』と『島が』が『あの島が』にまとめ上げられる．手順 3 では，まず，『登ると』に係る『急いで』と『丘に』の語順整序が行われ，これら 3 つのノードが『急いで丘に登ると』にまとめ上げられる．続いて，『見えた。』に係る『あの島が』と『急いで丘に登ると』の語順整序が行われ，これら 3 つのノードが『急いで丘に登るとあの島が見えた。』にまとめ上げられ，終了する．

ここで，高須らの手法 [3] と提案手法との違いを図 1 により概説する．高須らの手法では，各文節を修飾する文節を考慮しないため，例えば，『見えた。』に係る『島が』と『登ると』の兄弟ノードのみを対象として，『島が登ると見えた。』や『登ると島が見えた。』に対して RNNLM を適用している．一方，提案手法では，『あの島が急いで丘に登ると見えた。』や『急いで丘に登るとあの島が見えた。』に対して RNNLM を適用することになる．

4　評価実験

提案手法の有効性を示すために，新聞記事文を用いた語順整序実験を実施した．なお，本研究では新聞記事文

† 東京電機大学大学院未来科学研究科, Graduate School of Science and Technology for Future Life, Tokyo Denki University.
‡ 名古屋大学情報連携推進本部, Information and Communications, Nagoya University.
a) 21fmi21@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

表 1　実験結果

	二文節単位	文単位
[SVM]	85.49% (4,661/5,452)	57.47% (669/1,164)
[RNNLM]	81.42% (4,439/5,452)	51.54% (600/1,164)
[RNNLM+SVM]	85.75% (4,675/5,452)	59.19% (689/1,164)
[RNNLM⁻]（[3]）	81.69% (4,454/5,452)	50.17% (584/1,164)
[RNNLM⁻+SVM]（[3]）	85.84% (4,680/5,452)	58.59% (682/1,164)

は読みやすい語順であるとみなす.

4.1　実験概要

実験データは高須ら [3] と同一であり，京大コーパス Ver.4.0 のうち，1 月 1 日から 8 日までと 1 月 10 日から 6 月 9 日までの 25,388 文を学習データとし，1 月 9 日と 6 月 10 日から 6 月 30 日までの 2,368 文から，構文情報のみから 1 文全体の語順が確定する文を取り除いたもののうち，1,050 文を開発データ，1,164 文をテストデータとした．評価指標は二文節単位一致率（2 つずつ係り文節を取り上げ，その順序関係が元の文と一致しているものの割合）[4]，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）を採用した．SVM には LIBSVM V3.24[*1] を用い，オプションの設定は，type を NuSVC，確率を得るため-b を 1 とする以外，すべてデフォルトのままとした．また，RNNLM の学習は Chainer V6.4[*2] を介して行った．学習アルゴリズムには SGD を採用し，パラメータの更新はミニバッチ学習（学習率 0.01，バッチサイズ 40）により行った．エポック数は 35 とし，embedding 層及び隠れ層（LSTM2 層）の次元数はいずれも 400 とした．入力の one-hot ベクトルの次元数は 19,316 とした．これは，学習データ中の異なり語数に未知語タグ及び文節境界タグを加えた数である．出力層も同じ次元数とした．

比較のために提案手法 [RNNLM+SVM] において SVM と RNNLM をそれぞれ単独で用いた手法 [SVM] と [RNNLM] を用意した．また，高須ら [3] の手法を [RNNLM⁻+SVM]，その RNNLM 単独手法を [RNNLM⁻] とし，それらの実験結果を参考までに示す．なお，高須ら [3] は，1 文全体の語順整序を行っておらず，文単位一致率は算出していないが，高須らの実験データに基づき 1 文全体の語順を再現した．[RNNLM+SVM] の混合比率 α については，$0 \leq \alpha \leq 1$ の下で 0.01 毎に値を変え，開発データにおいて文単位一致率が最大となる $\alpha = 0.23$ とした．

4.2　実験結果

表 1 に実験結果を示す．提案手法 [RNNLM+SVM] は，二文節単位と文単位の両一致率において，[SVM] と [RNNLM] を上回っており，語順整序において構文情報と言語モデルを共に用いることの有効性を確認した．

一方，高須ら [3] の [RNNLM⁻+SVM] と比較すると，提案手法 [RNNLM+SVM] は，文単位で上回ったが，二文節単位では下回った．[RNNLM] と [RNNLM⁻] との比較でも同様の傾向となった．RNNLM の学習データは 2 万文強であるため，RNNLM 単独での精度はまだ十分ではなく，言語モデルの効果が十分に発揮されていないと考えられる．

4.3　言語モデルによる影響の分析

まず RNNLM の好影響を分析する．文単位において，提案手法 [RNNLM+SVM] が正解した 689 文のうち，[SVM] では不正解となった文は 53 文存在した．その典型例を図 2 に示す．この例では，慣用句『視野に入れる』を提案手法のみがひとまとまりにして正しく並べている．慣用句は，複数の語が固く結びつき，ひとまとまりとなっており，他の語がその内部に割り込むことは少ない．そのことを言語モデルは学習できていると考えられる．慣用句やそれに類する表現に対しては，言語モデルを考慮した提案手法だけが成功した例が複数見られた．

提案手法の出力(正解)
立法府も/時代の/流れを/視野に/入れた/施策に/心すべきだ 。
[SVM]の出力(不正解)
立法府も/視野に/時代の/流れを/入れた/施策に/心すべきだ 。

図 2　提案手法の成功例

提案手法の出力(不正解)
米国は/日本に/通知するだけで/「新しい/路線は/開設できる」と/いう/立場/を/とっている 。
[SVM]の出力(正解)
米国は/「新しい/路線は/日本に/通知するだけで/開設できる」と/いう/立場/を/とっている 。

図 3　提案手法の失敗例

次に RNNLM の悪影響を分析する．文単位において，[SVM] が正解した 600 文のうち，提案手法 [RNNLM+SVM] では不正解となった文は 33 文存在した．その典型例を図 3 に示す．提案手法は，元の文において引用の括弧‘「’と‘」’の内部に現れる『日本に通知するだけで』を誤って‘「’の前に出している．新聞記事文では，‘「’は文頭に現れにくく，その前に何らかの文節列が現れやすい傾向があり，その傾向を RNNLM は学習していると考えられる．しかし，これは 1 文単位で見た場合の傾向であり，1 文に含まれる部分木の範囲で見ると，‘「’が先頭に来ることは少なくない．提案手法では，ボトムアップに 1 文全体の語順整序を行う途中で，部分木に相当する単語列に対しても，1 文単位の傾向を学習した RNNLM を適用することになり，その悪影響が出たものと考えられる．

5　おわりに

本稿では言語モデルと構文情報を用いて，ボトムアップに日本語文を語順整序する手法を提案した．語順整序実験の結果，言語モデルと構文情報を併用することの有効性を確認した．今後は，RNNLM 以外の言語モデルを用いるなどして精度の向上を図りたい．

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]栗林ら, “言語モデルを用いた日本語の語順評価と基本語順の分析,” 言語処理学会第 25 回年次大会発表論文集, pp.1053 – 1056, 2019.
[2]A. Schmaltz et al., “Word Ordering Without Syntax,” Proc. EMNLP2016, pp. 2319 – 2324, 2016.
[3]高須ら, “RNNLM と SVM を用いた日本語文の語順整序,” 情報処理学会第 82 回全国大会講演論文集, pp.453–454, 2020.
[4]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.
[5]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.

*1　https://www.csie.ntu.edu.tw/~cjlin/libsvm/
*2　https://chainer.org/