4T-04

RNNLMとSVMを用いた日本語文の語順整序

高須 恵 †, 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統括本部 †

1 はじめに

日本語は語順が比較的自由であると言われているが，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生成されることがある．従って，機械翻訳や文生成においては，読みやすい語順の文を生成する技術が重要となる．

語順整序に関する研究は，推敲支援や文生成などへの応用を目的として，これまでにも幾つか行われている．特に近年，深層学習ベースの言語モデルを用いた研究が成果を挙げており，RNN ベースの言語モデル（RNNLM）を用いた研究 [1] や，CNN ベースの言語モデルを用いた研究 [2] がある．一方，内元ら [3] は，言語モデルを用いず，日本語の語順決定に関わる様々な要因を素性として最大エントロピー法により学習したモデルを用いて語順整序を行う手法を提案している．

そこで本稿では，RNNLM と，内元ら [3] のモデルを SVM により再構築したモデルとを統合することにより，日本語文を構成する文節の語順を読みやすく並べる手法を提案する．本手法は，文生成への応用を念頭に，文を構成する文節集合と，それら文節間の係り受け関係を入力とし，入力文節集合内の順序関係を同定する．新聞記事文を用いて語順整序実験を行った結果，本手法は，RNNLM 及び SVM モデルをそれぞれ単独で用いた場合よりも高い精度を達成した．

2 文生成における語順整序

本研究では，文を構成する文節集合と，それら文節間の係り受け関係は既知であるとして，それらを入力とし，その入力文節集合内の文節を読みやすく並べることを試みる．これらの入力が既知であるとの仮定は，文生成や機械翻訳への応用を念頭においたものであり，文を生成するあたって，その文で表したい内容は決まっている状況を想定したものである．

この仮定は，内元らの先行研究 [3] の問題設定においても見られる．内元らの先行研究では，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 $B_r = \{b_1, b_2, \cdots, b_n\}$ に対して，B_r から考えられる順列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．係り受け構造を表す木を考え，受け文節に係る複数の文節から成る集合ごとに内元らが定義した語順整序をボトムアップに繰り返すことにより，1 文全体の語順整序を行うことができる．そのため，本研究では，内元らの研究と同じ問題設定を採用することとする．

3 RNNLMとSVMを用いた語順整序

本手法では，文を構成する文節集合と，それら文節間の係り受け関係を入力とし，その入力文節集合内の文節を読みやすく並べた文を生成する．その際，内元ら [3] と同様に，任意の受け文節 b_r に係る文節の集合 B_r に対して最も読みやすい順列を求める．この計算を係り受け木の最も深い箇所からボトムアップに繰り返せば，1 文全体の語順整序を行えることになる．なお，形態素解析は事前に施されていることとする．

また，各 B_r に対する語順整序では，RNNLM と，内元ら [3] のモデルを SVM により再構築したモデルとを統合したモデルを用いて，B_r 内の読みやすい順列 $\mathbf{b}$ を決定する．以下では，内元ら [3] のモデルを SVM により再構築したモデルを用いた手法，RNNLM を用いた手法，これらを統合した本手法を順に説明する．いずれも，順列 $\mathbf{b}$ に対して，その読みやすさを表すスコア関数 $S(\mathbf{b})$ を定義し，B_r から考えられる全ての順列 $\mathbf{b}^k (1 \le k \le n!)$ の中で，$argmax_{\mathbf{b}\in\{\mathbf{b}^k|1\le k\le n!\}} S(\mathbf{b})$ を読みやすい語順の順列とする手法である．そのため，各手法のスコア関数を中心に説明する．

3.1 SVM を用いた手法

内元ら [3] は最大エントロピー法を用いて文節内外に含まれる情報から語順の傾向を学習したモデルを作成し，そのモデルを用いることで語順の推定を行っている．本研究では，内元らの手法と同じ素性を用いて語順の傾向を学習したモデルを作成する．ただし，最大エントロピー法ではなく，SVM を用いることとした．

具体的には，内元ら [3] と同様に，ある順列 $\mathbf{b}^k = b^k_1 b^k_2 \cdots b^k_n$ に対するスコア関数を式 (1) により定義する．なお，b^k_i は順列 $\mathbf{b}^k$ における i 番目の文節を意味する．

$$P_{svm}(\mathbf{b}^k) = \prod_{i=1}^{n-1} \prod_{j=1}^{n-i} p_{svm}(o^k_{i,i+j}) \quad (1)$$

ここで，$o^k_{i,i+j}$ は，文節 b^k_i が b^k_{i+j} よりも文頭側に現れる順序関係を表すものであり，$p_{svm}(o^k_{i,i+j})$ は，$o^k_{i,i+j}$ であることが読みやすさにおいて妥当である確率を SVM により推定した値を意味する．なお本研究では，SVM から確率を得るため，Platt's scaling[4] によって，境界面からの距離を分類確率に変換した値を用いた．

3.2 RNNLM を用いた手法

ある順列 $\mathbf{b}^k = b^k_1 b^k_2 \cdots b^k_n$ に対するスコア関数を，その形態素列 $w^k_{1,1} w^k_{1,2} \cdots w^k_{1,m_1} \cdots w^k_{n,1} \cdots w^k_{n,m_n}$ を言語モデルで生成する確率，すわなち，式 (2) により定義する．

$$P_{rnnlm}(\mathbf{b}^k) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} p_{rnnlm}(w^k_{i,j} | w^k_{1,1}, w^k_{1,2}, \cdots, w^k_{i,j-1}) \quad (2)$$

ここで，$w^k_{i,j}$ は文節 b^k_i の j 番目の形態素，$m_i - 1$ は文節 b^k_i を構成する形態素数とし，w^k_{i,m_i} は文節境界を意味する特殊記号を表すものとする．また，$p_{rnnlm}(w^k_{i,j} | w^i_{1,1}, w^k_{1,2}, \cdots, w^k_{i,j-1})$ は，LSTM[5] を用いた 2 層の隠れ層を持つニューラルネットワークにより推定する．

3.3 RNNLM と SVM を統合した手法

上記で説明した 2 つのモデルの混合モデルを考える．ここで，SVM を用いた手法では文節を単位として式 (1)

Japanese Word Ordering by Combining RNNLM and SVM
Megumi Takasu[†, a)], Tomohiro Ohno[†, b)], Shigeki Matsubara[‡]
† School of Science and Technology for Future Life, Tokyo Denki University.
‡ Information and Communications, Nagoya University
a) 16fi071@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

を計算しているのに対し，RNNLM を用いた手法では形態素を単位として式 (2) を計算しており，式 (1) と式 (2) では因子の数が異なる．そのため，2 つの確率を直接混合することはできない．そこで本研究では，式 (1) と式 (2) を，式 (3) と式 (4) によりそれぞれ正規化する．

$$\hat{P}_{svm}(\mathbf{b}^k) = \frac{P_{svm}(\mathbf{b}^k)}{\sum_{\mathbf{b}\in\{\mathbf{b}^{k'}|1\le k'\le n!\}} P_{svm}(\mathbf{b})} \quad (3)$$

$$\hat{P}_{rnnlm}(\mathbf{b}^k) = \frac{P_{rnnlm}(\mathbf{b}^k)}{\sum_{\mathbf{b}\in\{\mathbf{b}^{k'}|1\le k'\le n!\}} P_{rnnlm}(\mathbf{b})} \quad (4)$$

RNNLM と SVM を統合した手法では，式 (3) と式 (4) を混合比率 α により混合させた式 (5) を順列 $\mathbf{b}^k$ に対するスコア関数として用いる．

$$P(\mathbf{b}) \equiv \alpha\hat{P}_{rnnlm}(\mathbf{b}^k) + (1-\alpha)\hat{P}_{svm}(\mathbf{b}^k) \quad (5)$$

4 評価実験

本手法の有効性を検証するため，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，新聞記事文は読みやすい語順となっているとみなす．

4.1 実験概要

実験データには京大コーパス 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て使用した．また，1 月 9 日と 6 月 10 日から 6 月 30 日までの 2,368 文のうち，前半の 1,132 文を開発データ，後半の 1,236 文をテストデータとして使用した．形態素情報及び文節間の係り受け情報については，当該コーパスのものをそのまま利用した．

評価では，内元ら [3] と同じく，二文節単位一致率と完全一致率を測定した．これらは，1 つの受け文節に係る複数の係り文節から成る集合ごとに，その語順整序結果が元の文とどの程度一致するかを評価する指標である．このうち，完全一致率は，テストデータ 1,236 文から得られる係り文節集合 2,360 個のうち，各集合を構成するすべての係り文節の語順が元の文と完全に一致しているものの割合である．一方，二文節単位一致率は，係り文節集合ごとに，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である．例えば,「先方の/迅速な/対応で/問題が/全て/解決した」という文からは,「解決した」に係る「対応で」「問題が」「全て」から成る集合と,「対応で」に係る「先方の」「迅速な」から成る集合の 2 つの集合が評価対象となる．語順整序結果が「問題が/先方の/迅速な/対応で/全て/解決した」であると，二文節単位一致率は 75.00% (3/4)，完全一致率は 50.00% (1/2) となる．

比較手法として，RNNLM 及び SVM をそれぞれ単独で用いた以下の手法を用意した．

・[RNNLM]：式 (5) において $\alpha = 1$ とする手法
・[SVM]：式 (5) において $\alpha = 0$ とする手法

なお，本手法 [RNNLM+SVM] における式 (5) の α については，$0 \le \alpha \le 1$ の下で 0.01 毎に値を変え，開発データにおいて完全一致率が最高となる α を選択した．なお，本実験においては α が 0.15 のとき，二文節単位一致率，完全一致率共に最も高くなった．

SVM の学習は LIBSVM V3.24[*1] を用いて行った．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，

表 1: 実験結果

手法	二文節単位	完全一致
[SVM]	85.49 % (4,661/5,452)	73.94 % (1,745/2,360)
[RNNLM]	81.69 % (4,454/5,452)	68.90 % (1,626/2,360)
本手法 [RNNLM+SVM]	85.84 % (4,680/5,452)	75.04 % (1,771/2,360)

本手法[RNNLM+SVM]の出力(正解)

最高裁判例は	なく、	下級審で	判断が	分かれた	ままである

[RNNLM] の出力

最高裁判例は	なく、	判断が	下級審で	分かれた	ままである

[SVM] の出力

下級審で	判断が	分かれた	最高裁判例は	なく、	ままである

図 1: 本手法の成功例

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 を採用した．パラメータの更新はミニバッチ学習（学習率 1.0，バッチサイズ 20）により行った．エポック数は 39 とした．embedding 層及び隠れ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 17,077 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ものである．出力層も同じ次元数とした．

4.2 実験結果

表 1 に実験結果を示す．いずれの一致率においても，本手法は，[SVM] や [RNNLM] を上回った．図 1 に本手法のみが成功した例を示す．この例では，[RNNLM] と [SVM] がそれぞれが異なる箇所で失敗しているが，本手法では，それらを補い合って成功している．以上より，語順整序における本手法の有効性を確認した．

なお，内元らの研究 [3] では完全一致率が 75.41%であり，これと比べ，本稿の実験結果はやや低いものとなった．これは，用いた京大コーパスの Version が異なる等，完全に同一の実験となっていないためと考えられ，単純に比較することはできない．

5 おわりに

本論文では，RNNLM と SVM を用いて文節集合を読みやすい語順に並び替える手法を提案した．語順整序実験の結果，RNNLM と SVM を併用することの有効性を確認した．今後は，RNNLM を単独で用いた際の精度があまり高くなかった原因を調査し，その精度向上を図ることにより，本手法の改善を試みたい．

謝辞 本研究は，一部，科学研究費補助金基盤研究（C）No. 16K00300 及び No. 19K12127 により実施した．

参考文献

[1] A. Schmaltz et al., “Word Ordering Without Syntax,” Proc. EMNLP2016, pp. 2319 – 2324, 2016.
[2] 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,” NLP2019 発表論文集, pp.1053 – 1056, 2019.
[3] 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.
[4] J. C. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods,” ADVANCES IN LARGE MARGIN CLASSIFIERS, pp. 61 – 74, 1999.
[5] M. Sundermeyer et al., “LSTM Neural Networks for Language Modeling,” Proc. INTERSPEECH 2012, pp. 194 – 197, 2012.

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