E-020

深層学習を用いたかな書き文の語分割の評価と改良
Word Segmentation of Non-Segmented Kana String Using Deep-Learning

森山 柊平　絹川 博之
Shuhei Moriyama　Hiroshi Kinukawa

1. はじめに

近年，外国人を対象とした日本語学習においてコンピュータが広く利用されるようになってきた．しかし，外国人日本語学習者が作成した文章を添削するようなシステムは見受けられず，日本語教師により人手で添削されているのが現状である．

そこで我々は初級日本語学習者が独学で文章作成を学べることを目標として日本語学習支援システムを開発している．学習支援のため彼らの書く文の誤りを検出・訂正しようとする際に，問題となるのが文中に現れるかなの多さである．初級日本語学習，特にその初期学習に際しては，文の一部のみならず文全体をもひらがなのみで記述することは珍しくない．文中にかなが多く存在する文字列の形態素解析は，漢字かなまじり文字列の形態素解析と比較すると精度が格段に落ち込む．

本稿では，文字列の表層が全てかなからなる所謂かな書き文の語分割に深層学習を適用して処理する方式について述べる．提案手法は，リカレントニューラルネットワーク(RNN)を利用し，従来の形態素解析手法によって算出されるラティスの最尤経路群からさらに最尤経路を推定する．

2. RNN によるラティスの最尤経路の推定

提案手法の概略図を図 1 に示す．

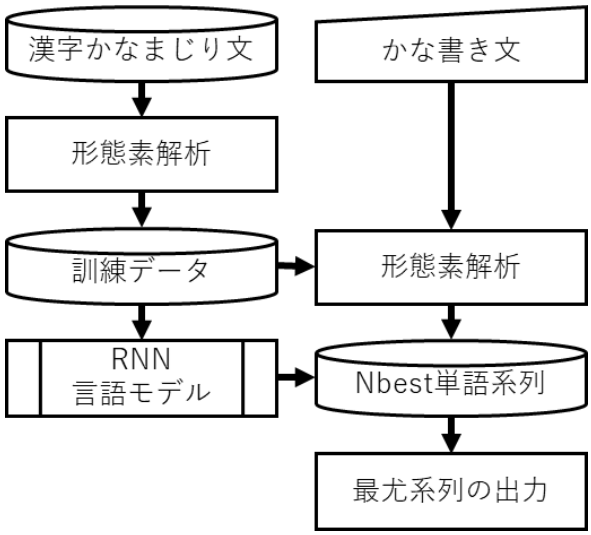

図 1　提案手法の概略図

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

本稿では，特に明記しない限り形態素解析器は MeCab[1]を指し，形態素解析辞書は ipadic[2]を指すものとする．

提案手法は，かな書き文を形態素解析し，任意の 1st-best から N-best の経路中から RNN を利用して最尤経路を推定することで，かな書き文の単語分割を実現する．入力系列を$\mathbf{x}$，系列中でとりうる要素の候補をXとしたとき，次式から最尤経路$\mathbf{y}$を求める．

$$\mathbf{y} = \text{argmax} \left(\prod_{i=1}^{n} softmax(\mathbf{x}_i) \right)^{\frac{1}{n}} \quad \mathbf{x}_i \in X$$

nは系列長を示し，$\mathbf{x}_i$は$\mathbf{x}$のi番目の要素を示す．関数$\text{softmax}(\mathbf{x}_i)$は RNN により$\mathbf{x}_i$の生起確率を求める関数である．

漢字かなまじり文の形態素解析は高い精度で可能であり，系列の尤もらしい分割の境界を得ることができる．この解析により得られる系列の分割の境界を訓練データとして RNN の学習に利用する．

2.1 形態素解析辞書のかな書き文への対応

かな書き文の形態素解析に際して問題となるのが汎用のかな形態素解析器がないことと，形態素解析辞書の生起確率・連接確率が漢字かなまじり文で調整されていることである．

かな書き文の形態素解析のため，漢字かなまじりの形態素解析辞書の読み仮名を見出しとするひらがな辞書を作成することで対応した．

生起確率・連接確率に関しては，RNN を訓練するのに用いる訓練データと同様のかな単語系列を上記のひらがな辞書に適用し，MeCab の再学習機能を利用して生起確率・連接確率を元モデルから更新することで対応する．このようにしてかな書き文用の形態素解析辞書を構築する．

3. 実験

京都テキストコーパス[3]から漢字かなまじり文を抽出して形態素解析をしたのち，異なり語彙数の調整を施した約 38 万かな単語列を訓練データに用いて実験を行った．訓練データはかな書き文の形態素解析に利用される辞書のモデルの更新と RNN の学習に使用している．

実験では 2 種類の単語系列を文頭からの順方向・文末からの逆方向の入力で学習した以下の 4 種類の RNN を用意した．

・品詞列を学習した RNN（順方向・逆方向）
・読み仮名列を学習した RNN（順方向・逆方向）

RNN の各種ハイパーパラメータの調整は Recurrent Neural Network Regularization[4]を参考にした．実装には Tensorflow[5]を利用した．

漢字かなまじり文の 1st-best 語分割を正解とした相対的な語分割の成功数・失敗数を表 1，2 にそれぞれ示す.

3.1 実験 1

訓練データ量： 約 38 万単語列
訓練データ： 品詞
異なり語彙数： 241 品詞
テストデータ量： 200 文

表 1　実験 1 の語分割結果

	順方向		逆方向	
N-best	成功数	失敗数	成功数	失敗数
2	184	16	182	18
4	173	27	173	27
8	165	35	163	37
16	161	39	159	41
32	143	57	139	61
64	141	59	135	65

3.2 実験 2

訓練データ量： 約 38 万単語列
訓練データ： 単語読み仮名
異なり語彙数： 7462 語
テストデータ量： 200 文

表 2　実験 2 の語分割結果

	順方向		逆方向	
N-best	成功数	失敗数	成功数	失敗数
2	191	9	192	8
4	191	9	194	6
8	191	9	194	6
16	191	9	193	7
32	191	9	190	10
64	188	12	188	12

4. 考察

漢字かなまじり文の形態素解析と比較して差異が生じた語分割を検討すると，大別して以下の傾向がみられた.

（1）数詞や複合語の語分割の差異
（2）文脈によっては正しい語分割
（3）文脈によらず誤りの語分割
（4）漢字かなまじり文の語分割と同様の経路が推定範囲の N-best 中に存在せず失敗

（1）は語によっては複合語が 1 つの形態素として登録されていることに起因すると考えられる．誤り検出・訂正といった後段の処理で許容できるのであれば問題にならないが，許容できない場合は接尾辞等，あるいは複合語の形態素を辞書から除く必要があるとみられる.

（2）は具体例には以下のような文である.
「これらは教育の力で解決可能であろう。」
「コレラは教育の力で解決可能であろう。」
上記のように入力文が短く文脈に乏しい上にあいまいである場合，改善のためにはあらかじめ入力されうる文のトピックに応じてドメイン適応を行っておく必要があると考えられる.

（3）のような誤りは系列の意味的自然さをより強く捉えられるように，双方向 RNN 等の利用で低減が見込めると考えられる．また，実験 2 に関しては単なる読み仮名に Word Embeddings を適用したことで同音異義語の特徴がひとつの読み仮名に集約された影響も考えられる．これは RNN に学習させる入力の情報（品詞+読み仮名，同音異義語の区別等）を改良することで低減が図れると考えられる.

（4）については形態素解析辞書の再学習を行ったことで，再学習をしない場合と比べ格段に改善されることを確認しているが，再学習によりモデルを更新しているとはいえ，元々は漢字かなまじり文のコーパスで調整されたモデルに基づいていることが一因と考えられる．改善の方策としては，今回再学習させた以上のかな書き文でモデルをさらに更新するか，一からかな書き文で調整されたモデルを構築し利用することで改善が見込めるものと考えられる.

5. おわりに

本稿では，従来の形態素解析と RNN を併用したかな書き文の語分割について述べた．RNN を用いた最尤経路の推定には従来の形態素解析と比して大きな計算量が伴うため，RNN で推定する経路を制限することで従来手法以上の精度を保ちつつ，計算量の低減が図れる感触を得た.

今後の性能改善へ向けた展望としては，双方向 RNN の利用，RNN に学習させる入力の改善のほか，訓練データ量の拡大等が挙げられる.

謝辞

MeCab，ipadic，京都大学テキストコーパス，および Tensorflow の開発に携わった方々に感謝いたします.

参考文献

[1] https://taku910.github.io/mecab/
[2] https://ja.osdn.net/projects/ipadic/
[3] 京都大学テキストコーパス Version 4.0
http://nlp.ist.i.kyoto-u.ac.jp/
[4] W.Zaremba, I.Sutskever, and O.Vinyals.
Recurrent Neural Network Regularization.
In arXiv:1409.2329, 2015.
[5] https://www.tensorflow.org/