7Q-01

機械学習によるかな書き文の語分割

森山 柊平　　絹川 博之
東京電機大学大学院 未来科学研究科

1. はじめに

近年，外国人日本語学習者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殆ど見られず，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．初級日本語学習者が扱う日本語の難易度は使用語彙数で表すと 1500 語ほどである．

学習支援のために彼らが書いた文章内の誤りを検出・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．初級日本語学習において，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．なお，カタカナ語や数字が含まれる場合はその限りでない．文中のかなの多さは十分な精度の形態素解析から困難となる．

本稿では，そういった殆どの単語がかなであるかな書き文の語分割を機械学習により行う方式について述べる．提案手法は，再帰型ニューラルネットワーク(RNN)を利用し漢字かな混じり文の形態素解析結果から単語分割境界を学習することにより，かな書き文の形態素解析結果からラティスの最適経路を推定する．

2. 外国人日本語学習者によるかな書き文

コンピュータを利用した外国人の日本語学習において問題となるのが，IME に搭載されている補完機能である．多少漢字の読みを間違えても，IME が補完・修正してしまうため，正しい読みが身に付かないといった弊害がある．それを避けるために，初期学習においては次のようなかな書き文が使われることが間々ある．

・これからひるごはんをたべます
・タクシーをよびましょうか
・5 じにあいましょう

こうした意図的に学習を目的として書かれるほか，漢字の語彙の不足からやむを得ず通常ひらがなで書かれることが少ない単語をひらがなで書き文を構成する場合がある．

3. 形態素解析辞書のひらがなへの対応

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

提案手法ではラティスの最適経路を推定するため，かな書き文の形態素解析に際して漢字かな混じり文と同程度のラティスを構築できる必要がある．通常の形態素解析辞書では，多くの場合，語彙の問題から困難である．

形態素解析辞書内の登録単語は，漢字かな混じりのものが多く，ひらがなのみからなる単語は機能語を始めとする一部の語に限られている．動詞や名詞の単語をひらがなで表現した単語はほとんど登録されていない．

以下はかな書き文の形態素解析例である．

表 1　かな書き文の形態素解析例

かな書き文	しんかんせんにのる
解析結果	しん/かんせん/に/のる

語分割の失敗は，文に未知語を含む場合と含まない場合の 2 通りに大別されると考えられる．通常の辞書によりかな書き文を解析するときに生じる語分割の失敗の多くは前者に起因する．これを低減するために，漢字かな混じりの形態素の生起確率といった各種パラメータを流用し，ひらがなの形態素を生成してひらがな辞書を作成する．ひらがな辞書によるかな書き文の解析は，通常の辞書による漢字かな混じり文の解析精度には及ばないとしても，通常の辞書を用いるよりは比較的もっともらしく解析することができるものと考える．

Word Segmentation of Non-Segmented Kana Sentence Using Machine-Learning
Shuhei Moriyama, Hiroshi Kinukawa
Graduate School of Science and Technology for Future Life, Tokyo Denki University

4. RNNによる最適経路の推定

2.で作成したひらがな辞書による形態素解析から得られるラティスの最適経路群から RNN によりひとつの最適経路を推定する．学習および推定には品詞列を用いる．

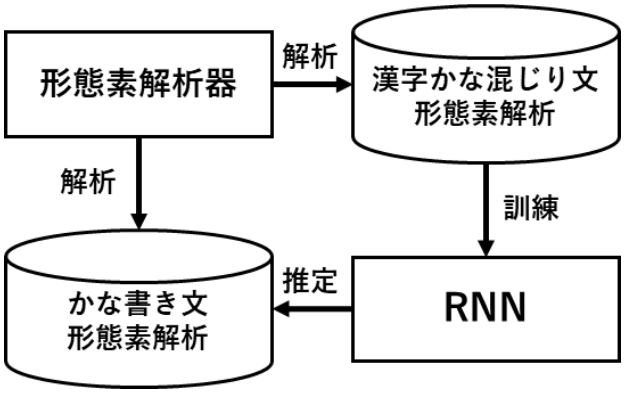

図1　提案手法の概要

RNN の訓練に関しては，漢字かな混じり文を形態素解析し，その 1Best 品詞列を訓練データとして RNN で学習して品詞に基づいた RNN 言語モデルを構築する．

構築した RNN 言語モデルによる推定では，かな書き文を形態素解析して得られる 1Best から NBest までの品詞列群からその中での最適経路を推定する．入力品詞列を$\mathbf{x}$，品詞列中でとりうる品詞の候補をXとしたとき，次式より最適経路$\mathbf{y}$を求める．

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

nは品詞列長を示し，$\mathbf{x}_i$は$\mathbf{x}$のi番目の品詞を示す．関数$\mathrm{softmax}(\mathbf{x}_i)$は RNN 言語モデルにより$\mathbf{x}_i$の生起確率を求める関数である．

5. 評価

今回，RNN 言語モデルの構築には日本語版 Wikipedia データベース・ダンプ[3]を利用した．ダンプ中から無作為に記事を選出させ各記事の本文を抽出し，それを形態素解析し品詞列としたものを訓練データとした．データ量は 2342 文ほどである．

評価用のテストデータに関しては，訓練データの作成において選出された記事と類似するドメインの記事から本文を抽出し，テストデータとした．

各種ハイパーパラメータの調整については，Recurrent Neural Network Regularization[4]を参考にした．実装には TensorFlow[5]を用いた．

表2　単語分割の精度

	MeCab (1Best)	MeCab (1~16Best) +RNN
分割成功箇所	49	57
分割失敗箇所	51	43

※MeCab は2.のひらがな辞書を利用

表 2 の単語分割の成功・失敗の分類に関しては，品詞は考慮せず単語がもっともらしく分割されているかどうかにのみ基づいている．

[MeCab]による単語分割が失敗した文と，[MeCab + RNN]による単語分割が失敗した文はおおよそ共通しており，文中には未知語が含まれていた．今回は ipadic をそのまま辞書に用いており，共通して分割に失敗している文はドメイン適応を行っておらず未知語が多数存在することに起因する単語分割の失敗と考えられる．また，[MeCab]により単語分割が成功しているにもかかわらず，[MeCab + RNN]で単語分割が失敗している文も確認された．これは訓練データに関してもドメイン適応を考慮していないため，漢字かな混じり文の形態素解析の段階で混入した未知語の影響がひとつの原因と考えられる．

6. おわりに

本稿では形態素解析と RNN 言語モデルによるラティスの最適経路の推定について述べた．Wikipedia データベース・ダンプという大規模な言語資源を利用しながら，計算資源の不足から訓練データ量・RNN の規模ともに抑えての実験であったが，単純なひらがなへの適応を施した辞書を用いての形態素解析より高い精度での単語分割が行えることが分かった．RNN の規模および学習量の拡大，辞書類のドメイン適応により，さらなる精度向上が見込めるものと思われる．

今回は品詞列を用いて学習・推定を行い，漢字かな混じり文とかな書き文の対応をとったが，単語対単語で対応をとった場合などに関しても今後実験を行っていく予定である．

参考文献

[1] https://taku910.github.io/mecab/
[2] https://ja.osdn.net/projects/ipadic/
[3] https://dumps.wikimedia.org/jawiki/
[4] W.Zaremba, I.Sutskever, and O.Vinyals. Recurrent Neural Network Regularization. In arXiv:1409.2329, 2015.
[5] https://www.tensorflow.org/