情報処理学会論文誌 Vol.59 No.10 1911–1921 (Oct. 2018)

Recurrent Neural Network Language Modelを用いたべた書きかな文の形態素解析

森山 柊平[1,a)] 大野 誠寛[1,b)] 増田 英孝[1] 絹川 博之[1]

受付日 2018年1月6日, 採録日 2018年6月8日

概要：外国人向け初級日本語教育では，日本語の読みを学ぶために，学習者は最初にかなのべた書きで作文を行う．このため，習い始めの学習者を対象とした学習支援システムはべた書きかな文を形態素解析する必要がある．しかし，従来の形態素解析器は，主に漢字かなまじり文により学習されており，べた書きかな文の解析にそのまま適用することはできない．一部，べた書きかな文により学習し直した解析器を用いて，かなで構成された絵本テキストの形態素解析を試みた研究が存在するが，漢字かなまじり文に対する解析と比べて，十分な解析精度は得られていない．そこで本稿では，誤りを含まないべた書きかな文を対象として，形態素周辺確率とRecurrent neural network language model（RNNLM）を用いた形態素解析手法を提案する．RNNLMの効果により単語系列の意味的自然さをとらえた解析を，また，形態素周辺確率の効果によりビームサーチにおける最適経路の取りこぼしの軽減を期待できる．評価実験では，新聞記事から生成したべた書きかな文に対する形態素解析を実施した．RNNLMによる悪影響や最適経路の取りこぼしの残存などによる失敗があるものの，単語分割と単語素性すべての一致を正解とする最も厳しい基準において，提案手法のF値は95.52を達成し，従来手法よりも有意（$p < 0.01$）に上回ることを確認した．

キーワード：リカレントニューラルネットワーク，RNN言語モデル，条件付き確率場，形態素解析，ひらがな

Morphological Analysis of Unsegmented Kana Strings Using Recurrent Neural Network Language Model

Shuhei Moriyama[1,a)] Tomohiro Ohno[1,b)] Hidetaka Masuda[1] Hiroshi Kinukawa[1]

Received: January 6, 2018, Accepted: June 8, 2018

Abstract: In elementary Japanese language education for foreigners, students only use kana characters for writing in order to learn how to pronounce Japanese words. Therefore, an elementary Japanese language learning system needs to analyze unsegmented kana strings as a method of preprocessing to find errors and to give advice for their correction. Conventional morphological analyzers are trained on native-speech sentences, which contain characters other than kana characters. Thus, analyzers cannot simply be applied to sentences composed of only kana characters. Although there has been research that performs morphological analysis of kana-string sentences on picture books using an analysis tool re-trained by kana-string sentences, its analytical accuracy is not high enough. We propose a morphological analysis method integrating a conventional method and recurrent neural network language model (RNNLM) for kana-string sentences not containing grammatical errors. Our method can perform morphological analysis catching semantic plausibility of a word sequence through the RNNLM. We conducted an experiment on morphological analysis of kana-string sentences. Although there were some errors caused by the harmful effects of the RNNLM, we confirmed that our method achieved an F-measure of 95.52 on the hardest evaluation criterion and significantly outperformed the conventional methods ($p < 0.01$).

Keywords: recurrent neural network, recurrent neural network language model, conditional random fields, morphological analysis, hiragana

[1] 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, Adachi, Tokyo 120–8551, Japan

[a)] 16fmi35@ms.dendai.ac.jp
[b)] ohno@mail.dendai.ac.jp

1. はじめに

外国人向けの初級日本語教育では，習い始めの学習者に日本語の読みを学ばせるため，かなのべた書きで作文を行わせる．そのため，習い始めの初学者を対象とした外国人向け日本語学習支援システムは，べた書きかな文が入力されるものとして，形態素解析を行う必要がある．

日本語形態素解析器として，JUMAN [1]，ChaSen [2]，MeCab [*1] [3]，KyTea [*2] [4] などが開発され公開されているが，これら従来の形態素解析器は，主に漢字かなまじり文からなるコーパスで形態素解析パラメータが推定されており，漢字かなまじり文の形態素解析を主眼としていることから，べた書きかな文の形態素解析にそのまま適用することはできない．べた書きかな文からなるコーパスで形態素解析パラメータを推定したとしても，一般に，漢字かなまじり文の場合と比べて，べた書きかな文は，考えられる単語候補の組合せが増大するなど，はるかに曖昧性が多いことが知られており [5]，解析精度が低下するという問題がある．一部，ほとんどがひらがなで構成されている絵本のテキストを解析対象とした研究があり，ひらがなで書かれた形態素のデータを用いて KyTea を学習し直すことにより，従来手法と比べて高い形態素解析精度を得ている [6]．しかし，従来手法による漢字かなまじり文に対する解析精度と比べると，いまだ十分とはいえない．加えて，初級日本語学習者の作文中には単語綴りや文法上の誤りの存在が想定され，そうした文の解析はさらに困難をともなうと考えられる．そのほか，かなを扱うツールとして仮名漢字変換ツール [5], [7], [8], [9], [10] があるが，ユーザによる同音異義語の適切な選択が逐次行われることを前提としている場合が多く，そのまま適用して高精度な形態素解析を実現することは難しい．

そこで本稿では，初級日本語学習者の中でも特に習い始めの初学者が作文を独習できるシステムの開発を目標に，べた書きかな文の形態素解析手法を提案する．提案手法では，Recurrent neural network language model（RNNLM）を従来の形態素解析手法に組み合せることにより，べた書きかな文に対する高精度な形態素解析を実現する．なお本稿は，その端緒として，単語綴りや文法上の誤りを含まないべた書きかな文を対象とする．

Recurrent neural network language model（RNNLM）[11] は Recurrent neural network（RNN）に基づく言語モデルであり，長文脈や単語の並びの意味的自然さをとらえることが可能とされている．この RNNLM を用いた形態素解析の実現をごく単純に考えると，形態素ラティス上の最適経路を探索する際に動的計画法を利用できないため，ラティス上の全経路について RNNLM を適用し確率を求めることになるが，計算量が非常に大きくなり現実的ではない．この計算量を削減する探索アルゴリズムにビームサーチ [12] がある．ビームサーチはビーム幅と呼ばれる探索幅の上限をあらかじめ定め，探索の全候補ノード群のうちビーム幅分の有力なノード群だけを探索するアルゴリズムである．ビームサーチの適用で RNNLM による形態素ラティスの探索が現実的になるが，ここで形態素ラティス上の最適経路が経路の前方で低い確率をとる場合を考えると，最適経路のノードをビーム内にとらえられないことが想定される．この想定に鑑みれば，RNNLM による形態素ラティスの探索にビームサーチを用いるには探索を補助する手段が必要となる．

計算量を削減し，かつ最適経路の取りこぼしを低減する手法の 1 つとして，従来の形態素解析手法で得られる上位の最適経路群，いわゆる N-best 経路群のみに RNNLM を適用する手法がある [13], [14]．この手法の欠点は，RNNLM を考慮したときの最適経路が N-best 経路群の外に存在する場合に対応できないことである．RNNLM の適用範囲である N を大きくすることは上述のビームサーチと同様に計算量の問題があり，N のとれる大きさは限られてくる．このため，限られた N-best 経路群の範囲だけではなく，ラティス上を全体的かつ効率的に探索できる手法が必要となる．

ビームサーチを用いた RNNLM による形態素ラティスの探索を，従来のコスト最小法に基づく形態素解析手法を活用して補助することを考えたとき，ラティスのそれぞれの深さで利用できる情報として単語コスト・連接コストがある．一例として，単語コスト値・連接コスト値と RNNLM が与える確率を何らかの方法で組み合わせてビームサーチを行う方式を考える．この場合も，最適経路が経路の前方で高いコスト値をとりうることを想定すると，単語コスト・連接コストはラティス全体を考慮したものではないため，探索を補う手段としては不十分である．最適経路の取りこぼしを低減させるには，ラティス全体を考慮したコスト値のようなものが必要となる．

この解決策として，本研究では形態素周辺確率 [15] を利用する．形態素周辺確率は，従来のコスト最小法に基づく形態素解析を利用し，ラティス上で可能なすべての経路のコスト値を加味して算出される．このことから，上述のような最適経路の取りこぼしの低減が期待できる．すなわち，提案手法の特徴は，この形態素周辺確率を考慮することにより，従来のコスト最小法に基づく形態素解析手法の上で RNNLM の適用を実現している点にある．

関連研究として，Morita らの RNNLM を用いた日本語形態素解析 [16] がある．この研究では，本研究とは異なり，素性ベクトルに対応する重みベクトルを Soft confidence-weighted learning（SCW）により推定し，この 2 つのベクトルに基づく形態素解析モデルを新規に作成することによ

[*1] http://taku910.github.io/mecab
[*2] http://www.phontron.com/kytea/

り，RNNLM を用いた形態素解析を実現している．

以下，2 章では従来のコスト最小法に基づく形態素解析から得られる形態素周辺確率と RNNLM を組み合わせた形態素解析手法を提案する．3 章では，べた書きかな文のコーパスを作成する方法について述べる．このコーパスを用いて，べた書きかな文の形態素解析パラメータを推定する．4 章では，従来手法と提案手法をそれぞれ，べた書きかな文の形態素解析に適用した実験を通し，従来手法と比較して提案手法が高い解析精度を持つことを示す．5 章では，RNNLM の適用がべた書きかな文の形態素解析にもたらす影響について述べる．また，本研究の目的である外国人日本語学習者の作文データに対して提案手法による形態素解析を実施し，その実用上の課題について述べる．

2. 形態素周辺確率と RNNLM を用いたべた書きかな文の形態素解析手法

本研究は，RNNLM を用いた形態素解析手法の実現のために形態素周辺確率を利用する．本章では，形態素周辺確率と RNNLM が与える確率とを用いた形態素解析手法を提案する．

2.1 形態素周辺確率

形態素周辺確率は，コスト最小法に基づく形態素解析における，ラティス上で可能なすべての経路の累積のコスト値から算出される．コストの低い経路に含まれる形態素の形態素周辺確率は大きく，コストの高い経路に含まれる形態素の形態素周辺確率は小さくなる．なお，本研究では形態素周辺確率を Conditional random fields（CRFs）[17] に基づく形態素解析 [3] から算出する．

2.2 形態素周辺確率と RNNLM を用いた提案手法

提案手法では，形態素周辺確率と RNNLM が与える確率を線形補間によって組み合わせ，ラティス上の任意の経路 y のスコアを次式で計算する．

$$score(y) = \prod_{i=1}^{N} \{(1-\alpha) P_{mp}(y_i;\theta) + \alpha P_{rnnlm}(y_i|y_0^{i-1})\} \tag{1}$$

最適経路は $score(y)$ が最大となるような経路 $\hat{y}$ である．

$$\hat{y} = argmax\ score(y) \tag{2}$$

N は経路長を，α は線形補間の補間係数を，y_0 は文の開始記号を，y_i（$i \geq 1$）は経路の i 番目の形態素を，$P_{mp}(y_i;\theta)$ は y_i の形態素周辺確率を，$P_{rnnlm}(y_i|y_0^{i-1})$ は y_0 から y_{i-1} までの系列 y_0^{i-1} が入力されたときに RNNLM が与える y_i の生起確率をそれぞれ示す．形態素周辺確率 $P_{mp}(y;\theta)$ のパラメータ θ は，形態素周辺確率の確率分布の鋭さを与えるパラメータである．θ を大きくすると，コスト最小法に基づく形態素解析でコストが最小となる経路上の形態素の周辺確率を “1” に，他を “0” にするような効果が得られる．θ と α の値はグリッドサーチなどで発見的に決定する．なお，形態素周辺確率はラティス上で可能な語彙で正規化されているのに対し，RNNLM の与える生起確率は分類クラス数，すなわち言語モデルの語彙全体で正規化されているため，α の値は RNNLM が与える確率を大きくする値に偏る傾向にあることに注意が必要である．

ラティス上の最適経路を探索するために，すべての経路について RNNLM でスコアを算出するのは計算量の観点から現実的ではないため，ビームサーチを利用しラティス探索時の幅を制限する．ある時点で探索しているラティス上の頂点から接続しうる形態素の候補のうち，任意のビーム幅分だけ $score(y)$ が与えるスコアが高い候補を保持して同様の探索を繰り返す．

3. 京都大学テキストコーパスを用いたかなコーパスの生成

従来の機械学習手法と比較して，RNNLM の訓練には大量のテキストが必要となるが，日本語学習教材の電子テキストを大量に用意し，さらに形態素情報のアノテーションまで行うことは容易ではない．そのため，本研究では，比較的日本語学習教材の日本語文に近い新聞記事の日本語文に基づいていて[*3]，かつ，形態素情報が人手で付与された大量のテキストを得ることができる京都大学テキストコーパス Version 4.0[*4] [19] を利用し[*5]，このコーパスから疑似的に生成したべた書きかな文を用いることとした．

京都大学テキストコーパス中の単語系列の例を**図 1** に示す．単語素性は左から品詞（大分類），品詞（細分類），活用型，活用形，原形，読みである．図 1 に示すような単語素性の情報を使用して，見出し語のかな表記化をはじめとした変形を施し，かなコーパスを構築する．図 1 からの生成例を**図 2** に示す．単語素性は左から品詞（大分類），品詞（細分類），活用型，活用形，見出し語の漢字かな表記，読みである．京都大学テキストコーパスで存在していた原形の情報に関しては，べた書きかな文の形態素解析において同音異義語の識別を考慮した場合，見出し語の漢字かな表記のほうが識別への寄与が見込めると考え，置き換える

[*3] 実際，日本語学習教材では，日常的な日本語使用に堪える力をつける目的で，新聞・雑誌などの表記が採用されている（たとえば，文献 [18]）．

[*4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス

[*5] 現代日本語書き言葉均衡コーパス（BCCWJ）[20] のコアデータにも人手で修正された形態素情報（UniDic 形式）が付与されているが，新聞・雑誌などの出版物に限ると，そのデータ量は京都大学テキストコーパスよりも少ないこと，また，後述する 5.2 節の実験において，形態素解析器 JUMAN++（ver. 1.01）[*6] [16] と比較するために JUMAN 形式の形態素情報が付与されている必要があったことから，京都大学テキストコーパスを利用することとした．

[*6] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

見出し語	単語素性
方針	名詞, 普通名詞,*,*,*, ほうしん
を	名詞, 格助詞,*,*,*, を
固めた	動詞,*, 母音動詞, タ形, 固める, かためた
。	特殊, 句点,*,*,*,。

※単語素性中の*は該当する情報なしを示す.

図 1 京都大学テキストコーパス中の単語系列の例
Fig. 1 Example of a word sequence in Kyoto Corpus.

見出し語	単語素性
ほうしん	名詞, 普通名詞,*,*, 方針, ほうしん
を	名詞, 格助詞,*,*, を, を
かためた	動詞,*, 母音動詞, タ形, 固めた, かためた
。	特殊, 句点,*,*,。 ,。

※単語素性中の*は該当する情報なしを示す.

図 2 図 1 から生成したかなコーパス
Fig. 2 Kana corpus generated from Fig. 1.

形で削除している.

4. べた書きかな文の形態素解析評価実験

べた書きかな文の形態素解析における提案手法の有効性を評価するため, 3 章で作成したべた書きかな文を用いて, 従来手法との比較実験を実施した.

4.1 実験データ

実験では, 3 章において京都大学テキストコーパスを変換し作成したかなコーパス（38,400 文 972,894 単語）を利用する. 異なり単語数は 43,213 語である. **表 1** に詳細を示す. 訓練データに 36,400 文 927,997 単語を, 開発データに 1,000 文 22,198 単語を, テストデータに 1,000 文 22,699 単語をそれぞれ使用した.

訓練データは CRFs に基づく形態素解析パラメータの推定, および, RNNLM の訓練に用いた. 開発データは RNNLM の訓練時のハイパーパラメータ調整, 提案手法のパラメータ α, θ の調整に用いた.

4.2 比較手法

提案手法（以下, [CRFs+RNNLM] と記す）との比較のため, 以下 4 つの比較手法を設けた.

- [KyTea]：点予測による形態素解析器 KyTea を用いる手法. ただし, 付属の学習機能を利用して, 本実験の訓練データにより学習し直したものを用いた. 使用する単語素性も提案手法と同一である. なお, KyTea は絵本テキストの形態素解析 [6] において用いられた解析器である.
- [MeCab]：CRFs に基づく形態素解析器 MeCab を用いる手法. ただし, 付属の学習機能を利用して, 本実験の訓練データにより学習し直したものを用いた. 使用する単語素性も提案手法と同一である.

表 1 実験データの詳細
Table 1 Details of experimental data.

内訳	データ総数	訓練データ	開発データ	テストデータ
文数	38,400	36,400	1,000	1,000
単語数	972,894	927,997	22,198	22,699
単語素性	品詞（大分類・細分類）, 活用型, 活用形, 見出し語の漢字かな表記, 読み			

- [CRFs]：提案手法 [CRFs+RNNLM] において, $\alpha = 0$ として（すなわち, 形態素周辺確率のみを用いて）スコアを算出する手法.
- [RNNLM]：提案手法 [CRFs+RNNLM] において, $\alpha = 1$ として（すなわち, RNNLM が与える確率のみを用いて）スコアを算出する手法.

4.3 評価指標

評価では適合率・再現率・F 値, および, 1 文あたりの平均解析時間を測定した[*7]. なお, 適合率・再現率・F 値の計算式は以下のとおりである.

$$Precision = \frac{\text{正解と判定された単語数}}{\text{出力された単語数}} \tag{3}$$

$$Recall = \frac{\text{正解と判定された単語数}}{\text{正解の単語数}} \tag{4}$$

$$F = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{5}$$

ここで, 正解データは京都大学テキストコーパスのアノテーションとし, 正解の判定基準として次の 5 つを設けた.

(1) level0：単語分割が正解.
(2) level1：level0 に加え品詞（大分類）が正解.
(3) level2：level1 に加え品詞（細分類）が正解.
(4) level3：level2 に加え活用型と活用形が正解.
(5) level4：level3 に加え見出し語の漢字かな表記と読みが正解.

なお, level4 の正解は同音異義語の識別の成功に相当する.

4.4 実験環境の設定

形態素解析器 KyTea には ver. 0.4.6 を, MeCab には ver. 0.996 を用いた. なお, 本稿で KyTea や MeCab を用いる際は共通してこのバージョンのものを使用している.

形態素周辺確率の算出には MeCab に実装されているものを利用した[*8]. 形態素周辺確率の算出に必要な形態素解析辞書のパラメータ推定にあたっては, 頻度によるフィルタをかけず, 訓練データに出現するすべての単語素性を用いた. また, 学習の強さを決定する, CRFs のハイパーパラメータ C はデフォルト値である 1.0 を指定した. MeCab

[*7] 適合率・再現率・F 値の計測には MeCab 付属の mecab-system-eval を, 解析時間の計測には bash の time コマンドにおける real time の値をそれぞれ利用した. なお, 提案手法は Python により実装しており, 実行には Python 3.4.2 を利用した.
[*8] MeCab は「ソフトわかち書き」という名称で実装している.

表 2 実験結果
Table 2 Experimental results.

	解析時間	F 値（適合率 [%]/再現率 [%]）				
	[ms/文]	level0	level1	level2	level3	level4
[KyTea]	4.30	95.62 (95.85/95.39)	94.98 (95.21/94.75)	94.38 (94.61/94.15)	94.34 (94.57/94.11)	93.28 (93.51/93.06)
[MeCab]	0.07	96.28 (96.83/95.74)	93.82 (94.35/93.29)	90.52 (91.04/90.01)	90.39 (90.91/89.89)	83.10 (83.58/82.63)
[CRFs]	31.58	96.35 (96.88/95.83)	93.81 (94.32/93.30)	90.49 (90.99/90.00)	90.39 (90.88/89.90)	82.06 (82.51/81.61)
[RNNLM]	7404.38	93.15 (90.55/95.91)	92.51 (89.93/95.25)	92.04 (89.47/94.76)	92.01 (89.44/94.73)	90.96 (88.42/93.65)
[CRFs+RNNLM]	7424.71	98.68 (98.50/98.87)	98.01 (97.82/98.19)	97.28 (97.09/97.46)	97.21 (97.02/97.39)	95.52 (95.34/95.70)

における辞書の構築に必要な素性テンプレートなどの設定ファイルの記述は，京都大学テキストコーパスに基づいて MeCab 向けにビルドされた JUMAN 辞書と基本的に同一であり，1-2gram までの単語素性を利用している．ただし，辞書の動作を指定するオプション eval-size については，品詞（大分類），品詞（細分類），活用型，活用形だけでなく，見出し語の漢字かな表記，読みも含んだ単語素性すべてを考慮するように設定を変更している．上記で構築した形態素解析辞書は，[MeCab] の形態素解析，ならびに [CRFs+RNNLM]，[CRFs] の形態素周辺確率の算出に共通して利用した．

本研究で用いた RNNLM は，word2vec モデル [21] に基づく単語素性の Word embeddings を入力とする，順方向 LSTM（Long Short-Term Memory）[22] ネットワークを基に構築している．隠れ層のニューロン数をはじめとする構成およびその正則化は Zaremba らの large LSTM [23] と同一である．実装には TensorFlow [*9] [24] を利用した．上記で構築した RNNLM を，[CRFs+RNNLM] と [RNNLM] に共通して利用した．

[CRFs+RNNLM], [CRFs], [RNNLM] の適用の際のビーム幅は, [CRFs+RNNLM] を用いた予備実験において, level4 の F 値が最大となった 13 と定めた．なお，ビーム幅ごとの解析精度については 5.1 節で後述する．[CRFs+RNNLM] 適用の際の θ と α の値，および [CRFs] 適用の際の θ の値は，[CRFs+RNNLM] を用いた開発データ上でのグリッドサーチから $\theta = 0.001$，$\alpha = 0.999$ と定めた．

4.5 使用プラットフォーム

実験は次の 2 つのプラットフォーム上で行った．

(1) Debian 9 64-bits（CPU: Intel Core i7-6700 3.40 GHz, RAM: 16.0 GB, GPU: MSI GeForce GTX1080 GAMING X 8G 2-Way SLI）：MeCab の形態素解析辞書のパラメータ推定，および RNNLM の訓練に使用した．RNNLM の訓練には約 310 分要した．

(2) Debian 8 64-bits（CPU: Intel Core i5-6500 3.20 GHz, RAM:8.0 GB）：[KyTea], [MeCab], [CRFs], [RNNLM], [CRFs+RNNLM] の評価実験に使用した．

[*9] https://www.tensorflow.org/

4.6 評価実験結果

表 2 に実験結果を示す．解析時間は，提案手法 [CRFs+RNNLM] が最大となり，[RNNLM] もほぼ同等に時間がかかっていることが分かる．[CRFs] と比べ，提案手法や [RNNLM] の解析時間は大幅に増加していることから，LSTM に基づく RNNLM の利用に時間がかかっているものと考えられる．提案手法の解析時間の改善は今後の課題である．

一方, 解析精度では, 全 level の適合率・再現率・F 値において，提案手法 [CRFs+RNNLM] が，[KyTea]，[MeCab]，[CRFs], [RNNLM] の各々を有意（$p < 0.01$）に上回った[*10]．

[KyTea] は，level0 において [RNNLM] を除く他の手法と比べ低い F 値となっているが，他の level では提案手法 [CRFs+RNNLM] に次ぐ F 値を達成していた．次に [MeCab] と [CRFs] は，どちらも CRFs を用いた手法であり，ほぼ同程度の F 値を示した．一方，[RNNLM] は，[MeCab] や [CRFs] と比べると，level0 と level1 では低い F 値となったが，level2 から level4 では高い F 値を達成していることが分かる．

[CRFs] と [RNNLM] を統合した提案手法 [CRFs+RNNLM] は，形態素周辺確率と RNNLM の相乗効果が期待どおり得られたと考えられ，全体的に解析精度が飛躍的に向上しており，比較手法の中で 2 番目に解析精度が高いと考えられる [KyTea] と比べても，全 level において，F 値が 2 から 3 ポイント向上していた．以上の結果より，べた書きかな文に対する形態素解析精度の向上において，提案手法が有効であることを確認した．

5. 考察

本章では，実験結果に基づき，べた書きかな文の形態素解析における RNNLM の適用の影響を考察する．また，実際の初級日本語学習者の作文を提案手法により形態素解析し，実用上の課題について述べる．

5.1 形態素解析性能へのビーム幅の影響

1 章や 2 章で述べたとおり，提案手法では，RNNLM を利用することにともない，ビームサーチによる探索を行っ

[*10] テストデータを 10 分割した t 検定を実施し確認した．

表 3 ビーム幅ごとの解析時間および level4 における F 値（適合率/再現率）

Table 3 Analysis time and f-measure (precision/recall) on level4 by beam size.

ビーム幅	解析時間 [ms/文]	F 値（適合率 [%]/再現率 [%]）
1	654.29	82.62 (78.99/86.61)
2	1172.31	91.64 (90.37/92.95)
3	1736.43	94.01 (93.32/94.71)
4	2295.41	94.42 (93.92/94.93)
5	2865.67	94.94 (94.58/95.31)
6	3425.18	95.01 (94.68/95.34)
7	4004.85	95.27 (94.99/95.54)
8	4569.82	95.34 (95.10/95.58)
9	5126.92	95.44 (95.22/95.66)
10	5654.80	95.43 (95.20/95.66)
11	6280.68	95.47 (95.28/95.66)
12	7083.38	95.47 (95.28/95.67)
13	7424.71	95.52 (95.34/95.70)
14	7971.78	95.49 (95.32/95.67)

ている．4 章の評価実験では，4.4 節で説明したとおり，予備実験の結果に基づいて，ビーム幅を 13 と定めたが，ビーム幅による形態素解析性能への影響を明らかにすることは有益な知見を与えると考えられる．そこで本節では，予備実験におけるビーム幅ごとの解析時間および解析精度を示し考察する．なお，予備実験は，4 章と同一の実験設定を用いて，提案手法のビーム幅を 1 から順に 1 ずつ増加させて実施した．

表 3 にビーム幅ごとの提案手法の解析時間と level4 における解析精度を示す．解析時間は，想定どおり，ビーム幅に比例して増加していることが分かる．一方，F 値をみると，ビーム幅が大きくなるほど増加傾向にあるが，ビーム幅が 5 のときあたりから，徐々に収束している様子が分かる．予備実験はビーム幅 14 で打ち切ったが，これは，ビーム幅が 10 と 14 のときの 2 回，F 値が落ち込んだことから，解析精度は収束状態に入ったと判断したためである．ビーム幅を単純にこれ以上大きくしたとしても，解析時間がいたずらに増加するだけであり，大幅な解析精度の向上は見込めないと考えられる．

5.2 漢字かなまじり文の形態素解析精度との比較

漢字かなまじり文と比べて，べた書きかな文の形態素解析は，一般に難しいタスクであり，その解析精度が低下する傾向にあることは 1 章で述べたとおりである．RNNLM を用いた提案手法を適用することにより，その低下をどれだけ抑えることができているかを確認するため，漢字かなまじり文に対する形態素解析実験を実施し，べた書きかな文に対する実験結果との比較を F 値により行った．

漢字かなまじり文に対する実験の設定を述べる．実験データには，かなコーパスに変換する前のオリジナルの京都大学テキストコーパスを，表 1 と同様に，訓練データ，開発データ，テストデータに分割したものを用いた．また，比較手法として，[KyTea] や [MeCab] のほか，1 章で関連研究としてあげた Morita ら [16] の RNNLM を用いた形態素解析器 JUMAN++（ver. 1.01）を用いる手法（以下，[JUMAN++]）を設けた．JUMAN++のバージョン 1.X は，形態素定義を変更したデータで学習し直すことは非常に難しく，提案手法との公平な比較ができないため，べた書きかな文での比較は断念したが，JUMAN++は，提案手法と違う形ではあるものの RNNLM を用いているため，参考までに，漢字かなまじり文に対しては解析実験を実施した．なお，[KyTea] と [MeCab] は，オリジナルの京都大学テキストコーパスの訓練データにより学習し直したものを利用し，[JUMAN++] は JUMAN++（ver. 1.01）をデフォルトのまま利用した．また，提案手法 [CRFs+RNNLM] のパラメータについては，開発データを用いて新たに決定し，ビーム幅 9，$\theta = 0.001$，$\alpha = 0.990$ とした．その他の実験設定は，4 章と同一とした．

表 4 漢字かなまじり文とべた書きかな文の形態素解析結果（F 値）の比較

Table 4 Comparison between morphological analysis results (F-measure) for kanji and kana mixture sentences and those for kana sentences.

		level0	level1	level2	level3	level4
漢字かなまじり文	[JUMAN++]	99.20	98.64	98.02	97.96	94.49
	[KyTea]	99.44	99.10	98.89	98.86	98.78
	[MeCab]	98.63	97.48	95.74	95.66	92.92
	[CRFs+RNNLM]	99.30	98.74	98.05	98.01	97.13
べた書きかな文	[KyTea]	95.62	94.98	94.38	94.34	93.28
	[MeCab]	96.28	93.82	90.42	90.39	83.10
	[CRFs+RNNLM]	98.68	98.01	97.28	97.21	95.52

表 4 に漢字かなまじり文とべた書きかな文の形態素解析精度を示す．下 3 段のべた書きかな文の形態素解析精度は，比較の簡単化のために，表 2 の値を再掲載したものである．

漢字かなまじり文から，べた書きかな文に解析対象を変えたときの解析精度の変化に着目すると，提案手法は，[KyTea] や [MeCab] と比べ，その落ち込みが最も小さいことが分かる．また，提案手法によるべた書きかな文の形態素解析は，著名な日本語形態素解析器 MeCab による漢字かなまじり文の解析と同程度か若干上回る解析精度を達成していた．

一方，漢字かなまじり文に対する解析精度は，[KyTea] が提案手法 [CRFs+RNNLM] をわずかに上回り，最も高い F 値となった．参考のために解析を実施した [JUMAN++] は，配布されている標準モデルを利用しており，提案手法や [KyTea] とは訓練データが異なっているなど，公平な比較ができているとはいえないが，提案手法より若干下回っ

正解	米国/の/これ/まで/の/主張/は/ 先進/国/重視/で/、/性急な/要求/が/目立ち/、/ 発展/途上/国/から/の/批判/が/強かった/。
[KyTea]	米国/の/これ/まで/の/主張/は/ せんしんこく/重視/で、/性急な/要求/が/目立ち/、/ 発展/と/上告/から/の/批判/が/強かった/。
[CRFs+RNNLM]	米国/の/これ/まで/の/主張/は/ 先進/国/重視/で/、/性急な/要求/が/目立ち/、/ 発展/途上/国/から/の/批判/が/強かった/。

※べた書きかな文を便宜上漢字かな表記で記載している．

図 3 単語分割の改善例

Fig. 3 Example of improving word segmentation.

正解	警察/庁/科学/警察/研究/所/で/ 鑑定/した/ところ/、/ 有機/リン/系/化合/物/が/検出/さ/れた/。
[KyTea]	警察/庁/科/額/警察/研究/所/で/ 官邸/した/ところ/、/ 有機/リン/系/化合/物/が/検出/さ/れた/。
[CRFs+RNNLM]	警察/庁/科学/警察/研究/所/で/ 鑑定/した/ところ/、/ 有機/リン/系/化合/物/が/検出/さ/れた/。

※べた書きかな文を便宜上漢字かな表記で記載している．

図 4 同音異義語の識別の改善例

Fig. 4 Examples of improving identification of homonym.

ていた．漢字かなまじり文に対する提案手法の有効性に関する検証は今後の課題としたい．

5.3 RNNLM の適用による形態素解析の改善

[KyTea] によって正しく解析できなかったべた書きかな文が，RNNLM を適用した提案手法 [CRFs+RNNLM] で解析できた例を紹介する．べた書きかな文の解析を便宜上漢字かな表記で記載していることに注意されたい．

(1) 単語分割における解析精度の向上

Morita ら [16] も報告しているように，RNNLM の適用によって単語分割の改善がみられた．**図 3** に例を示す．「せんしん/こく（先進/国）」が正解であるのに対し，[KyTea] は「せんしんこく（せんしんこく）」と解析誤りを犯したほか，「はってん/とじょう/こく（発展/途上/国）」に対し「はってん/と/じょうこく（発展/と/上告）」と誤りを犯した．その半面，[CRFs+RNNLM] は正解と同一の自然な解析に成功している．

(2) 同音異義語の解析精度の向上

べた書きかな文を解析対象としたときの特有の改善として，同音異義語の解析精度の向上がみられた．**図 4** に例を示す．図 4 のべた書きかな文は「けいさつちょうかがくけいさつけんきゅうじょでかんていしたところ、ゆうきりんけいかごうぶつがけんしゅつされた。」である．「かんていした」の部分は「かんてい/した（鑑定/した）」という解析が正解であるのに対し，[KyTea] は「かんてい/した（官邸/した）」と解析した一方で，[CRFs+RNNLM] は正しく「かんてい/した（鑑定/した）」と解析した．

正解	その/結果/、/7/割/の/主婦/が/映画/好き/で/、/ 特に/20/代/、/30/代/と/高く/なって/おり/、/ 若い/ほど/映画/好き/……/と/いう/傾向/が/表れた/。
[KyTea]	その/結果/、/七/割/の/主婦/が/映画/好きで/、/ 特に/20/代/、/30/代/と/高く/なって/おり/、/ 若い/ほど/映画/好き/……/と/いう/傾向/が/現れた/。
[CRFs+RNNLM]	その/結果/、/七/割/の/主婦/が/英/が/好きで/、/ と/国/20/代/、/30/代/と/高く/なって/おり/、/ 若い/ほど/映画/すき/……/と/いう/傾向/が/現れた/。

※べた書きかな文を便宜上漢字かな表記で記載している．

※「……」は原文のままであり，省略を意味するわけではない．

図 5 提案手法のみ解析を誤った例

Fig. 5 Example of errors made only by the proposed method.

図 3，図 4 から分かるように，べた書きかな文の解析における RNNLM の適用は，漢字かなまじり文の場合と同様に単語の並びの意味的自然さの考慮が可能であることを示している．

5.4 RNNLM の適用による形態素解析の誤り

RNNLM を適用したべた書きかな文の形態素解析の誤り事例として，[KyTea] は正解したが [CRFs+RNNLM] は解析を誤った例，[KyTea] と [CRFs+RNNLM] のどちらも解析を誤った例を紹介する．べた書きかな文の解析を便宜上漢字かな表記で記載していることに注意されたい．

5.4.1 KyTea は正解したが提案手法は解析を誤った例

[KyTea] は正解したが，[CRFs+RNNLM] は不正解となった例を**図 5** に示す．

図 5 に示す文のべた書きかな文は「そのけっか、ななわりのしゅふがえいがすきで、とくににぜろ[*11]だい、さんぜろだいとたかくなっており、わかいほどえいがすき……というけいこうがあらわれた。」である．「ななわりのしゅふがえいがすきで」の部分の正解が「な/わり/の/しゅふ/が/えいが/すき/で（7/割/の/主婦/が/映画/好き/で）」であるのに対し，[KyTea] は「な/わり/の/しゅふ/が/えいが/すきで（7/割/の/主婦/が/映画/好きで）」と最後の「すき/で（好き/で）」の箇所以外は正解したが，[CRFs+RNNLM] は「な/わり/の/しゅふ/が/えい/が/すきで（7/割/の/主婦/が/英/が/好きで）」と「えい/が（英/が）」の箇所も誤って解析した．[CRFs+RNNLM] の解析結果は，漢字かな表記であれば意味的に自然には並んでおり，「7 割の主婦が英国が好きで」という文意に解釈することができる．「英/が/好きで」と推定されたのは，単語の漢字かな表記を学習している RNNLM の作用が一因と考えられる．また，

[*11] 「にじゅう」となっていないのは，京都大学テキストコーパスの単語素性から機械的に生成したためである．同文後方の「さんぜろ」も同様．詳細は 5.5 節を参照されたい．

この誤った解析の影響を受け，[CRFs+RNNLM] では，その直後においても「とくに（特に）」を「と/くに（と/国）」と誤解析している．

「英/が」を通る経路が [CRFs+RNNLM] により最終的に選択されたのは，探索アルゴリズムであるビームサーチに起因する．ビームサーチの途中経過を分析したところ，一時的に「映画」も「英/が」もビーム内にとらえることに成功していたが，「映画」を通る経路が途中からビーム外に脱落することが分かった．「映画好きで」後方の「特に」周辺までは，「映画」を通る経路，「英/が」を通る経路のいずれもビーム内にとらえられていたが，以降はスコアが僅差で上回る「英/が」を通る経路のみが保持され，「映画」を通る経路はビーム外に完全に脱落した．

解決策としては，スコアの拮抗する複数の経路を保持できるような探索アルゴリズムを適用する方策が考えられる．本稿では，探索アルゴリズムとして，探索前にあらかじめ定めたビーム幅を探索中一貫して利用する基本的なビームサーチを適用したが，ビームサーチには，動的にビーム幅を変化させるもの [25]，探索に Backtracking を取り入れたもの [26] などの変種が存在する．べた書きかな文に曖昧性が多いという問題がある以上，N-best 解として複数の候補を保持できることが望ましく，有力な候補となりうる経路のビーム外への脱落は避けたい．有力な候補をより多く残すための具体的な探索アルゴリズムの検討は今後の課題である．

5.4.2 KyTea と提案手法のどちらも解析を誤った例

[KyTea] も [CRFs+RNNLM] も解析を誤った例を**図 6** に示す．

図 6 に示す文のべた書きかな文は「ななろく[*12]ねんのだいさんかいまではにほんますとがしゅりゅうだったが、だいよんかいからいちほんますとがちゅうしんになった。」である．「にほんますと」の部分の正解が「に/ほん/ますと（2/本/マスト）」であるのに対し，[KyTea] も [CRFs+RNNLM] も「にほん/ますと（日本/マスト）」と誤解析した．

[CRFs+RNNLM] も誤解析した一因として，適用している RNNLM が順方向 LSTM に基づいており，過去の情報しか利用できていないことが考えられる．直観的には，文の後方に「1 本マスト」という単語列が存在することが分かっていたならば，「2 本マスト」と推定することが可能だったはずである．

過去の情報に加え未来の情報も考慮する RNN として，RNN を双方向に拡張した Bidirectional RNN [27] がある．順方向 LSTM に基づく RNNLM の適用の代わりに，双方向 LSTM に基づく RNNLM の適用がアイデアとして浮かぶが，双方向 LSTM に基づく RNNLM は，単方向 LSTM に基づく RNNLM から，解析精度が向上しないことが Arisoy らによって報告されている [28]．反面，Arisoy らは LSTM を用いない双方向 RNN に基づく RNNLM に関しては，単方向のそれを大きく上回ったことを報告している．双方向 RNN は未来の情報を利用できるものの，LSTM を利用していないため，長文脈の考慮が必要な場合における解析精度の低下が懸念される．

解決策としては，順方向 LSTM・逆方向 LSTM に基づく 2 つの独立した RNNLM を用いることが考えられるが，単純計算で計算量は 2 倍になる．計算量の問題のほか，具体的な実現手法も含め，図 6 のような例の改善は今後の課題である．

[*12]「ななじゅうろく」となっていない理由は脚注*11 で述べたとおりである．同文後方の「いちほん」も同様．

正解	76/年/の/第/3/回/まで/は/2/本/マスト/が/主流/だった/が/、/第/4/回/から/1/本/マスト/が/中心/に/なった/。
[KyTea]	76/年/の/第/3/回/まで/は/日本/マスト/が/主流/だった/が/、/第/4/回/から/1/本/マスト/が/中心/に/なった/。
[CRFs+RNNLM]	76/年/の/第/3/回/まで/は/日本/マスト/が/主流/だった/が/、/第/4/回/から/1/本/マスト/が/中心/に/なった/。

※べた書きかな文を便宜上漢字かな表記で記載している．

図 6 [KyTea] も提案手法も解析を誤った例

Fig. 6 Example of errors made by both [KyTea] and the proposed method.

見出し語	単語素性
避難	名詞, サ変名詞,*,*,*, ひなん
所	名詞, 普通名詞,*,*,*, ところ
で	助詞, 格助詞,*,*,*, で
毛布	名詞, 普通名詞,*,*,*, もうふ
に	助詞, 格助詞,*,*,*, に
くるまる	動詞,*, 子音動詞ラ行, 基本形,*, くるまる

図 7 京都大学テキストコーパス中の未修正の単語の読み

Fig. 7 Example of the unmodified pronunciation of a word in Kyoto Corpus.

5.5 かなコーパスに起因する問題

本稿では，京都大学テキストコーパス中の単語に付与されている単語素性の情報を利用してかなコーパスを生成し，それをべた書きかな文の形態素解析のパラメータ推定用コーパスとした．京都大学テキストコーパスの利用に起因する問題として，読みが人手で確認されておらず，誤りが含まれている点があげられる [19] [*13]．例を**図 7** に示す．

図 7 の単語素性から読みを利用してべた書きかな文を作成すると，「ひなんところでもうふにくるまる」という文ができる．「避難所」の正しい読みは「ひなんじょ」であり，

[*13] なお，現代日本語書き言葉均衡コーパス（BCCWJ）[20] や日本語話し言葉コーパス（CSJ）[29] では読み情報が人手で付与されているが，BCCWJ は脚注*5 で述べた理由から，また，CSJ は話し言葉を収録したものであり新聞・雑誌などの出版物を含んでいないため，本研究では使用していない．

「ひなんところ」ではない．接尾語のほかに，数詞にも同様の問題がある．たとえば「1994」であれば，「せんきゅうひゃくきゅうじゅうよん」と読むのが自然であるが，コーパスから作成される読みは「いちきゅうきゅうよん」であり，現実の読みとは乖離がある．習い始めの初級日本語学習者がべた書きかな文で作文したデータの整備は今後の課題である．

5.6 初級日本語学習者の作文データに対する形態素解析

4 章の評価実験では京都大学テキストコーパスのきれいな日本語文を用いたが，実際に解析しなければならないのは学習者による作文である．日本語に習熟していない人間による作文である以上，文中に単語綴りや文法上の誤りが多分に含まれることは想像に難くない．そこで本節では，初級日本語学習者の作文データに対して，提案手法がどの程度の形態素解析精度を達成できるのかを示す．

実験のテストデータおよび開発データには，日本語学習者作文コーパス[*14] [30] のうち，日本語レベルが初級，かつ，学習期間が 5 年未満の学習者の作文データ 531 文（テストデータ：331 文，開発データ：200 文）を用いた．このデータには，MeCab + UniDic により解析後，人手で修正された形態素情報が付与されており [30]，このアノテーションを正解データとして評価した．なお，この作文データは，初級とはいえ，習い始めではない一定の学習期間を経た日本語学習者によるものであるため，漢字かなまじり文となっている．

漢字かなまじり文に対応するため，訓練データには，オリジナルの京都大学テキストコーパスと，それをべた書きかな文に変換したかなコーパスとを表 1 と同様に分割し，それらの両者をともに用いた．すなわち，訓練データの文数は，表 1 の訓練データの 2 倍の 72,800 文である．なお，テストデータの品詞体系に合わせるため，オリジナルの京都大学テキストコーパスの形態素情報を MeCab + 現代書き言葉 UniDic（ver. 2.2.0）[*15] [31] を用いて変換し，その後，その読み情報を使って，かなコーパスを作成した．

提案手法のビーム幅は 4 章の評価実験と同様に 13 とし，その他のパラメータは，開発データを用いて新たに決定し，$\theta = 0.001$，$\alpha = 0.800$ とした．評価指標は，4.3 節と基本的に同様であるが，level2 においては，UniDic 形式における最も細かい品詞情報まですべて一致した場合のみ正解とした．その他の実験設定は 4 章の評価実験と同一である．

初級日本語学習者の作文データに対する提案手法の F 値（適合率 [%]/再現率 [%]）は，level0 から level4 まで順に列挙すると，96.48（96.10/96.86），92.67（92.31/93.04），91.76（91.40/92.12），85.13（84.80/85.47），82.54（82.22/82.87）となった．新聞記事を変換し生成したべた書きかな文に対する実験結果（表 2）と比べると，初級日本語学習者の作文データに対しては，解析精度が大幅に低下していることが分かる．RNNLM は過去に入力された系列に影響を受け続けるため，誤りが含まれない学習データを用いて訓練した RNNLM は，解析時の入力に誤りが一部でも含まれると，その後の単語の生起確率を不正確に推定し続けることになり，解析精度が著しく低下するものと考えられる．解決に向けて，文法上の誤りについては誤用コーパスを活用した学習を行うほか，単語綴りの誤りは一種の未知語とも考えられるため，未知語処理の先行研究なども活用し取り組みたい．誤りが含まれる文への対応は今後の課題である．

[*14] http://sakubun.jpn.org/
[*15] http://unidic.ninjal.ac.jp/

6. おわりに

本稿では，従来のコスト最小法に基づく形態素解析から得られる形態素周辺確率と RNNLM とを組み合わせた形態素解析手法を提案した．京都大学テキストコーパス Version 4.0 から生成したべた書きかな文を利用した評価実験では，単語分割と単語素性すべての一致を正解として測定した F 値において，[KyTea] が 93.28，[MeCab] が 83.10 であったのに対し，提案手法は 95.52 を達成しており，2 ポイント以上の解析精度向上を確認した．実験結果に基づいて，RNNLM の適用がべた書きかな文の形態素解析にもたらす影響を検証したところ，単語分割における解析精度の向上や，同音異義語の解析精度の向上を確認した．

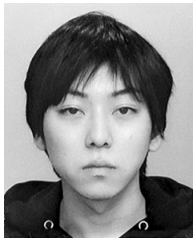
今後は，単純なビームサーチに代わる探索アルゴリズムの導入や，順方向 LSTM と逆方向 LSTM の両者を用いた手法の検討などを行い，べた書きかな文に対する形態素解析精度の向上を図る予定である．また，提案手法の解析時間の改善についても検討したい．さらに，本稿では単語綴りや文法上の誤りを含まない日本語文を扱ったが，誤りが含まれるべた書きかな文の形態素解析に向けて，誤用コーパスの活用方法などを検討し，誤りに頑健な形態素解析手法の開発にも取り組みたい．

謝辞 本研究の提案手法の実装に際しては MeCab と TensorFlow を利用した．評価に際しては，京都大学テキストコーパス，日本語学習者作文コーパス，UniDic，KyTea，JUMAN++を利用した．これらの開発に携わった方々に感謝する．

参考文献

[1] 黒橋・河原研究室：日本語形態素解析システム JUMAN version 7.0 (2012)，入手先 ⟨http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN⟩．
[2] 松本裕治，高岡一馬，浅原正幸：形態素解析システム『茶筌』version 2.4.3 使用説明書 (2008)，入手先 ⟨http://chasen-legacy.osdn.jp/⟩．
[3] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP-2004*),

pp.230–237 (2004).
[4] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (*ACL-HLT2011*), pp.529–533 (2011).
[5] 長尾 真（編）：自然言語処理，岩波講座 ソフトウェア科学 15，pp.122–132, 岩波書店 (1996).
[6] 藤田早苗，平 博順，小林哲生，田中貴秋：絵本のテキストを対象とした形態素解析，自然言語処理，Vol.21, No.3, pp.515–539 (2014).
[7] 牧野 寛，木澤 誠：べた書き文の分かち書きと仮名漢字変換 —二文節最長一致法による分かち書き，情報処理学会論文誌，Vol.20, No.4, pp.337–345 (1979).
[8] 吉村賢治，日高 達，吉田 将：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol.24, No.1, pp.40–46 (1983).
[9] 荒木哲郎，池原 悟，土橋潤也，笹島伸一：2 重マルコフモデルを用いたべた書きかな文の仮文節境界の推定方法，情報処理学会論文誌，Vol.38, No.6, pp.1116–1125 (1997).
[10] 小林龍生：漢字・日本語処理技術の発展 仮名漢字変換技術，*IPSJ Magazine*, Vol.43, No.10, pp.1099–1103 (2002).
[11] Mikolov, T., Karafiat, M., Burget, L., Cernocky, J. and Khudanpur, S.: Recurrent Neural Network Based Language Model, *Proc. 11th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2010*), pp.1045–1048 (2010).
[12] Lowerre, B.: The Harpy Speech Recognition System, Ph.D. Thesis, Carnegie Mellon University (1976).
[13] 森山柊平，絹川博之：機械学習によるかな書き文の語分割，第 79 回情報処理学会全国大会講演論文集，Vol.2, pp.593–594 (2017).
[14] 森山柊平，絹川博之：深層学習を用いたかな書き文の語分割の評価と改良，第 16 回情報科学技術フォーラム講演論文集，Vol.2, pp.193–194 (2017).
[15] 工藤 拓：形態素周辺確率を用いた分かち書きの一般化とその応用，言語処理学会第 11 回年次大会発表論文集，pp.592–595 (2005).
[16] Morita, H., Kawahara, D. and Kurohashi, S.: Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model, *Proc. 2015 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2015*), pp.2292–2297 (2015).
[17] Lafferty, J., McCallum, A. and Pereira, C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. 18th International Conference on Machine Learning* (*ICML 2001*), pp.282–289 (2001).
[18] 佐々木仁子，松本紀子：「日本語能力試験」対策 日本語総まとめ N3 文法，アスク出版 (2010).
[19] 黒橋禎夫，居蔵由衣子，坂口昌子：形態素・構文タグ付きコーパス作成の作業基準 version 1.8 (2000)，入手先 ⟨http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス⟩.
[20] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y.: Balanced Corpus of Contemporary Written Japanese, *Language Resources and Evaluation*, Vol.48, No.2, pp.345–371 (2014).
[21] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *CoRR*, Vol.abs/1310.4546, pp.1–9 (2013) (online), available from ⟨http://arxiv.org/abs/1310.4546⟩.
[22] Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, *Neural Computation*, Vol.9, No.8, pp.1735–1780 (1997).
[23] Zaremba, W., Sutskever, I. and Vinyals, O.: Recurrent Neural Network Regularization, *CoRR*, Vol.abs/1409.2329, pp.1–8 (2014) (online), available from ⟨http://arxiv.org/abs/1409.2329⟩.
[24] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I.J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D.G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P.A., Vanhoucke, V., Vasudevan, V., Viégas, F.B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, *CoRR*, Vol.abs/1603.04467, pp.1–19 (2016) (online), available from ⟨http://arxiv.org/abs/1603.04467⟩.
[25] Norvig, P.: *Paradigms of Artificial Intelligence Programming: Case Studies in Common LISP*, Morgan Kaufmann (1991).
[26] Furcy, D. and Koenig, S.: Limited Discrepancy Beam Search, *Proc. 19th International Joint Conference on Artificial Intelligence* (*IJCAI 2005*), pp.125–131 (2005).
[27] Schuster, M. and Paliwal, K.: Bidirectional Recurrent Neural Networks, *IEEE Trans. Signal Processing*, Vol.45, No.11, pp.2673–2681 (1997).
[28] Arisoy, E., Sethy, A., Ramabhadran, B. and Chen, S.: Bidirectional Recurrent Neural Network Language Models for Automatic Speech Recognition, *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2015*), pp.5421–5425 (2015).
[29] 国立国語研究所：日本語話し言葉コーパスの構築法，国立国語研究所報告 124，国立国語研究所 (2006).
[30] 林 炫情，李 在鎬，宮岡弥生，柴崎秀子，趙 坰熙：言語処理技術を利用した日本語学習者作文コーパスの開発，日本文化学報，Vol.56, pp.129–142 (2013).
[31] 岡 照晃：CRF 素性テンプレートの見直しによるモデルサイズを軽量化した解析用 UniDic：unidic-cwj-2.2.0 と unidic-csj-2.2.0，言語資源活用ワークショップ発表論文集，Vol.2, pp.144–153 (2017).



森山 柊平 （正会員）

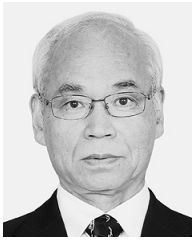
2016 年東京電機大学工学部第二部情報通信工学科卒業．2018 年同大学大学院未来科学研究科情報メディア学専攻修士課程修了．外国人向け日本語学習支援システムの研究に従事．2017 年情報処理学会第 79 回全国大会学生奨励賞受賞．同年第 16 回情報科学技術フォーラム FIT 奨励賞受賞．

大野 誠寛 （正会員）

2003 年名古屋大学工学部電気電子・情報工学科卒業．2007 年同大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．同年同大学院国際開発研究科助教．2011 年同大学情報基盤センター助教．2017 年より東京電機大学未来科学部情報メディア学科准教授．この間，日本学術振興会特別研究員．自然言語処理，音声言語処理の研究に従事．電子情報通信学会，言語処理学会各会員．

増田 英孝 （正会員）

1995 年東京電機大学大学院工学研究科電気工学専攻博士後期課程修了．博士（工学）．同年同大学工学部電気工学科助手．同大学工学部情報メディア学科講師，助教授を経て，現在，同大学未来科学部情報メディア学科教授．Web マイニング，ソーシャルメディアの活用等の研究に従事．ACM，言語処理学会各会員．



絹川 博之 （正会員）

1970 年東京大学理学部数学科卒業．1987 年理学博士（東京大学）．1970 年株式会社日立製作所に入社．以来，かな漢字変換，日本語文書処理，情報検索，自然言語インタフェースの研究に従事．1999 年東京電機大学教授．工学部情報通信工学科を経て，未来科学部情報メディア学科所属．以後自然言語処理の研究に従事．1987 年情報処理学会論文賞受賞．2018 年東京電機大学名誉教授．電子情報通信学会会員．本会終身会員．