語彙制約付きニューラル単語分割器を用いた後処理としての単語分割の後段タスクへの最適化

平岡 達也　　岩倉 友哉
富士通株式会社
{hiraoka.tatsuya, iwakura.tomoya}@fujitsu.com

概要

本稿では，後段タスクのモデルを改変することなく，単語分割器をモデルに最適化することで，精度改善を行う手法を提案する．本手法は，精度を改善したいタスクのモデルが与えられた際に，モデルが扱える単語に出力を制限しつつ，モデルにおける損失関数を最小化する単語列を生成するような単語分割器を学習する．また，文脈情報をよりよく捉えるために，LSTM 上で語彙制約付きの単語分割手法を提案する．本手法を，日本語，中国語，英語の文書分類タスクで評価した結果，ユニグラム言語モデルによる単語分割の最適化手法，単語の限定を行わないニューラル単語分割器と比較し，提案手法が性能の向上に寄与することが示された．

1　はじめに

単語分割は，さまざまな自然言語処理に共通する前処理の一つである．文書分類や機械翻訳などの後段タスクでは，それぞれのタスクやドメイン、モデル構造に応じて適切な単語分割を用いることで性能の向上が得られることが知られている [1, 2, 3]．近年では後段タスクで学習済みのモデルに対して適切な単語分割を求める手法が提案されている [4, 5]．これらの手法は，すでに学習されている後段モデルのパラメータを固定し，より性能が向上するような単語分割を後処理として求めることができる．

従来の後処理としての単語分割の最適化手法には，2 つの課題が残されている．1 つ目の課題は，従来手法 [5] では文脈に応じた単語分割がされにくいという点である．これは従来手法が単語分割器としてユニグラム言語モデル採用しているためである．タスクの性能が向上するような単語分割を探索した時，ある文脈でタスクの性能向上に貢献する単語が，別の文脈でもタスク向上に繋がるとは限らない．この問題は，例えば BI タグ付けによる文字レベルでのニューラル単語分割モデル [6, 7] を用いれば，解決されるように思われる．しかしながら，文字レベルのタグ付けによる単語分割では，実際に後段モデルで利用できる語彙には含まれていない未知語が出力されてしまう可能性がある [8]．

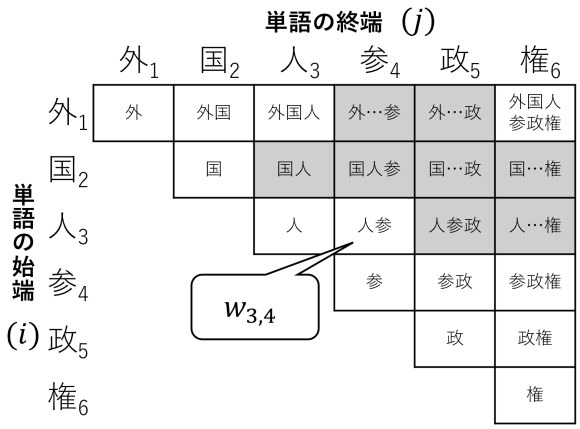

図 1　本手法で用いる語彙の制約（式 (4)）をテーブルとして表現した図．網掛けセルは未知語を表す．

2 つ目の課題は，単語分割の最適化のために後段モデルの構造を修正する必要があるという点である [4, 5]．広く利用されているモデルをそのまま使うことが出来ないため，後処理としての単語分割の最適化技術の利用の障壁となっている．

本研究では 1 つ目の課題への対処として，出力可能な語彙を制限したニューラル単語分割器による，後処理としての単語分割の最適化手法を提案する．これにより，従来のユニグラム言語モデルを用いた手法よりも高い表現力を持った単語分割器を用いて，後処理としての単語分割の最適化を行うことができる．さらに 2 つ目の課題への対処として，後段モデルの処理と単語分割の最適化処理の独立性を高めた学習方法を提案する．日本語，中国語，英語での文書分類の実験より，提案手法が後処理としての単語分割の最適化として妥当であることを示す．

2 後処理としての単語分割の最適化

本研究で対象とするタスクは，後処理としての単語分割の最適化である．本タスクでは文書分類などの後段タスクの学習データ D と単語分割器 T，単語分割済みのデータ D' で学習を行なった後段モデル θ を前提とする．また，$s \in D$ は学習データに含まれる入力テキスト，$s' \in D'$ は単語分割器 T で分割を行ったテキストで，$s' = T(s)$ である．様々な単語分割器を T として利用できるが、本研究ではユニグラム言語モデルによる単語分割器 [9] を使用する．

s の可能な単語分割は N 種類あり，$s'_1 ... s'_N$ と表す．ここで，N の最大数は s の文字数 $|s|$ について $N = 2^{|s|-1}$ である．実際には，後段モデルで使用できる語彙 V_θ が限られているため，N はある程度の大きさに収まる．後処理としての単語分割の最適化の目的は，後段タスクで後段モデルの性能が高くなる単語分割を出力する $\hat{T}$ を求めることである．

3 提案手法

3.1 学習の概要

本手法では，後段モデルの構造を変更せずに後処理としての単語分割の最適化を行えるように，後段モデルとは別に新しく単語分割器を学習する．具体的には以下に説明する通り，学習データでの性能が高くなるような単語分割の収集と，単語分割器の学習の 2 ステップに分けることで，後段モデルと単語分割器の独立性を高める．

学習データにおいて損失が小さい単語分割の収集 はじめに，$s \in D$ の可能な N 種類の単語分割 $s'_1 ... s'_N$ のうち，学習データでの損失が最も小さくなるような単語分割のみを収集した学習データ $\hat{s} \in \hat{D}$ を構築する．各テキストの N 種類の単語分割については，例えば N-best の単語分割 [10] を用いたり，サブワード正則などで利用される単語分割のサンプリング [11, 12, 13] によって収集する方法を用いたりすることができる．本研究では，$N = 100$ とし，N-best 単語分割を用いた．$\hat{D}$ の構築には，学習済みの後段モデル θ に N 種類の単語分割を入力し，それぞれの正解ラベルに対する損失を計算すれば良い．そのため $\hat{D}$ から学習することで、後段モデルの構造を変えずに後処理としての単語分割の最適化を行える．

$\hat{D}$ を再現する単語分割器 $\hat{T}$ の学習 次に，$\hat{D}$ の単語分割を再現するような新たな単語分割器 $\hat{T}$ を学習する．学習データにおいて後段モデルの損失が小さくなるような単語分割を学習した単語分割器 $\hat{T}$ は，検証データやテストデータにおいても正解ラベルとの損失が低い単語分割を出力すると期待される．

3.2 語彙制約付きニューラル単語分割器

$\hat{D}$ の単語分割を学習するための単語分割器 $\hat{T}$ には，様々な手法を用いることができる．この単語分割器の表現力が高いほど，損失が低くなるような単語分割の再現性能が高くなると考えられる．

表現力の高い単語分割器の例として，単語分割を BI タグで表現してニューラルネットワークを学習する手法 [6, 7] がある．しかし，この方法では後段モデル θ が使うことができる語彙，すなわちオリジナルの単語分割器 T で用いている語彙 V_θ には含まれない未知語を出力してしまう場合がある．このような未知語は，後段モデルにおいて適切に単語埋め込みへと変換できないため，後段タスクにおける性能低下につながると考えられる．

そこで本研究では出力可能な単語を考慮し [14]，利用できる語彙を制限したニューラル単語分割器を作成する．K 文字のテキスト $s = c_1 ... c_K$ について，i 文字目で始まり j 文字目で終わる単語 $w_{i,j}$ の出現確率 $p(w_{i,j}|s)$ を，次のように計算する．

$$\mathbf{h}_k = \mathrm{BiLSTM}(\mathbf{v}_{c_1} ... \mathbf{v}_{c_K})_k, \tag{1}$$

$$\mathbf{h}_k^{(\mathrm{begin})} = \mathrm{MLP}_{\mathrm{begin}}(\mathbf{h}_k), \tag{2}$$

$$\mathbf{h}_k^{(\mathrm{end})} = \mathrm{MLP}_{\mathrm{end}}(\mathbf{h}_k), \tag{3}$$

$$p(w_{i,j}|s) = \begin{cases} \sigma(\mathbf{h}_i^{(\mathrm{begin})\top} \mathbf{h}_j^{(\mathrm{end})}) & \text{if} \quad w_{i,j} \in V_\theta \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

ここで $\mathrm{BiLSTM}(\cdot)_k$ は，k 番目の入力 $\mathbf{v}_{c_k}$ に対応する BiLSTM [15, 16] の出力を得る操作である．また $\mathbf{v}_{c_k}$ は，c_k に対応する文字埋め込み表現である．$\mathrm{MLP}_{\mathrm{begin}}$ と $\mathrm{MLP}_{\mathrm{end}}$ はそれぞれ異なる多層パーセプトロン，$\sigma(\cdot)$ はシグモイド関数である．

$p(w_{i,j}|s)$ は s ごとにまとめて計算することができ，その処理は図 1 に示すような上三角行列として表せる．図と式 4 に示すように，後段モデルの語彙 V_θ に含まれない単語は確率が 0 になるようにマスクされる．これにより，単語分割器が利用できる語彙にハードな制約を与え，未知語の出力を防ぐ．

単語分割器の学習時は，各学習サンプル $\hat{s}$ につい

言語	データセット	最適化なし	最適化あり				
			ユニグラム	BI タグ付け	Optok	提案手法	オラクル
日本語	Twitter	86.28	**86.30**	83.21	**86.29**	**86.36**	94.57
	WRIME	44.83	44.76	43.35	**45.00**	**45.41**	75.05
中国語	Weibo	93.00	92.99	92.83	**93.15**	**93.13**	97.76
	Genre	48.15	48.13	47.65	**48.18**	**48.24**	71.94
	Rate	47.92	**47.97**	**48.58**	**47.96**	**48.75**	79.85
英語	Twitter	77.50	77.39	77.22	**77.77**	**77.64**	90.57

表 1　文書分類での性能（F1 値，テストデータ）．太字は単語分割の最適化を行う手法のうち，オリジナルの後段モデル（最適化なし）の値を超えるもの，下線は比較手法間での最大値を示す．

て以下の損失を最小化するように最適化する．

$$\mathcal{L}_{\hat{s}} = -\sum_{w \in \hat{s}} \log p(w|s). \tag{5}$$

推論時は，$\sum_{w \in s'} \log(p(w|s))$ が最も大きくなるような系列 s' をビタビアルゴリズム [17] で求める．

4　実験

4.1　設定

データセット　実験では，後段タスクとして文書分類を用いた．単語分割の後段タスクへの影響が大きいと考えられる言語として，日本語と中国語のデータセットを利用した．**Twitter** [18] と **WRIME** [19] は，日本語の SNS への投稿から作成された感情分類データセットである．**Weibo** は中国語の SNS への投稿から作成された感情分類データセット[1)]，**Genre** と **Rate** は E コマースサイトに投稿された中国語のレビューから作成した商品のジャンル予測とレート予測のためのデータセット [20] である．さらに，スペース区切りで単語境界を明示する言語での実験のために，英語の Twitter への投稿から作成した感情分類データセット[2)] を用いた．

後段モデル　後段モデル θ は，各文書分類の学習データで事前に学習を行う．本研究では，BiLSTM で単語分散表現の系列をエンコードし，線形層でラベルのスコアを予測するモデルを採用した．単語分散表現の次元数は 64，BiLSTM のレイヤ数は 1，隠れ層の次元数は前向き・後ろ向きそれぞれ 256 とした．各データセットで 20 エポックの学習を行い，検証データでの性能が最大となるモデルを θ として利用する．後段モデルの学習のための単語分割器 T には，SentencePiece の Unigram モードを用いた．語彙 V_θ の大きさは 16,000 とし，後段モデルの学習には $\alpha = 0.2$ でサブワード正則化 [11] を用いた．

比較手法　本実験では，$\hat{D}$ を学習する単語分割器として 2 種類の比較対象を用いる．**ユニグラム**は，$\hat{D}$ に含まれる単語の頻度を数え上げることでユニグラム言語モデルを作成し，単語分割に利用する最も単純な手法である．**BI タグ付け**は，BiLSTM-CRF による系列ラベリング手法を用いて，単語分割を表す BI タグの系列を予測する手法 [7] である[3)]．さらに，後段モデル θ を学習に組み込んで単語分割を最適化する従来手法として **OpTok** [5] を用いた．

4.2　後段タスクでの性能

各データセットでの実験結果を表 1 にまとめた．「最適化なし」の列には，オリジナルの単語分割で学習した後段モデル θ の性能を示した．「最適化あり」の各列には，学習済みの後段モデルに対して後処理として単語分割を最適化した結果[4)] を示した．

実験結果より，OpTok と提案手法はすべてのデータセットで，単語分割の最適化による性能の向上が得られた．また，提案手法は複数のデータセットで OpTok の性能を上回ることが確認された．一方で，最も単純なユニグラム言語モデルによる手法や，広く使われるニューラル単語分割器である BI タグ付による手法では，オリジナルのモデルよりも性能が下がる場合が多いことが分かった．

「オラクル」の列には，テストデータの 100-best 分割を学習済みの後段モデルに入力し，損失が最も小さくなるような単語分割だけを選択したときの性能を示した．すなわちこの値は，単語分割のみを修正して到達可能な性能の上限である．実際にはテスト

1）https://github.com/wansho/senti-weibo
2）https://www.kaggle.com/c/twitter-sentiment-analysis2

3）https://github.com/jidasheng/bi-lstm-crf の実装を用いた．文字分散表現の次元数は 128，BiLSTM の隠れ層の次元数は 256 とし，200 エポックの学習を行った．
4）BI タグ付け，OpTok，提案手法については 3 回試行の平均．検証データでの性能が最大となる単語分割器で評価．

言語	データセット	学習データの正解率 (%) ユニグラム	BI タグ付け	提案手法	検証データでの未知語割合 (%) ユニグラム	BI タグ付け	提案手法
日本語	Twitter	93.8	97.2	98.3	0.0	4.2	0.0
	WRIME	93.2	100.0	99.9	0.1	11.5	0.1
中国語	Weibo	93.1	97.0	93.9	0.0	1.0	0.0
	Genre	89.4	93.6	90.9	0.0	2.5	0.0
	Rate	88.8	93.2	92.1	0.0	2.3	0.0
英語	Twitter	96.7	97.0	97.5	0.0	5.6	0.0

表 2　各手法による学習データの単語分割の再現性能（正解率）と，検証データでの未知語出現割合.

データの正解ラベルは未知であるため、この値への到達は困難だが，後処理としての単語分割による性能向上の余地は大いに残されていると言える.

4.3　単語分割の再現性能と未知語割合

本手法では，学習データでの損失が低くなるような単語分割 $\hat{D}$ を再現するように，ニューラル単語分割器の学習を行った．表現力の低い単語分割器では，$\hat{D}$ を再現することも難しいと考えられる．そこで各単語分割手法が，それぞれどの程度単語分割の学習データを再現できているかを調べた.

表 2 の「学習データの正解率」として，ユニグラム言語モデル，BI タグ付け，提案手法のそれぞれについて，学習データでの単語分割の正解率を示した[5)]．ユニグラム言語モデルは単語の頻度数え上げで作成したモデル，BI タグ付けと提案手法はそれぞれ 200 エポックの学習を行ったモデルで評価した．また正解率は，単語分割を BI ラベルに変換した上で，ラベルの一致率として計算した.

学習データの単語分割の正解率より，ユニグラム言語モデルよりもニューラルネットワークを用いた BI タグ付けや，提案手法の再現性能が高いことが分かる．提案手法に比べて BI タグ付けの手法は，文字レベルで単語境界を予測できる点で表現力が高く，学習データの正解率も高くなっている．提案手法の表現力は，語彙をハードに制約する点で BI タグ付けの手法に劣り，正解率もやや低くなっている．しかし，単語分割最適化の従来手法 OpTok が採用しているユニグラム言語モデルよりも，提案手法は学習データの単語分割を再現できている．このため 4.2 節の実験においては，単語分割の最適化にニューラル単語分割器を用いる提案手法が，OpTok よりも高い性能の向上幅を得られていると考えられる.

表 2 の「検証データでの未知語割合」では，ユニグラム言語モデル，BI タグ付け，提案手法のそれぞれが検証データで後段モデルの語彙 V_θ に含まれない単語を出力した割合を示している.

ユニグラムと提案手法は，その性質として未知文字以外の未知語を出力することが出来ないため，検証データでの未知語割合はほぼゼロになっている．一方で BI タグ付けの手法は，出力できる単語についての制約がないため，最大で 11.5%もの未知語を出力している．これらの未知語が，4.2 節の実験における後段モデルの性能低下につながったと考えられる．実際に表 1 にまとめた BI タグ付けによる手法の性能は，多くのデータセットで「最適化なし」のモデルよりも低い値になっている.

これらの結果より，後処理としての単語分割の最適化においては，提案手法のように語彙の制約を設けたニューラル単語分割器を用いることが妥当であると結論付けることができる。

5　おわりに

本稿では，後処理としての単語分割の最適化の手法にニューラル単語分割器を用いる方法を提案した．本手法では後段タスクの学習データのうち，学習済みモデルの損失が小さくなるような単語分割を収集し，単語分割器の学習に用いる．また，後段モデルが利用できる語彙は限られているため，出力できる単語を制限するようなニューラル単語分割器を作成した．文書分類タスクでの実験結果より，提案手法は従来手法による後処理としての単語分割の最適化よりも性能が高くなるような単語分割を獲得できることが分かった．また，一般的に用いられる BI タグ付けによる単語分割器とは異なり，未知語の発生を防ぐことができることが示された．今後は機械翻訳などのタスクでの実験や，後段モデルと単語分割の同時最適化にも本手法を応用していく.

5)　OpTok は単語分割に関する学習データを用いないため，本評価の対象外である.

謝辞

本研究は，JST，ACT-X，JPMJAX21AM の支援を受けたものです.

参考文献

[1] Tatsuya Hiraoka, Hiroyuki Shindo, and Yuji Matsumoto. Stochastic tokenization with a language model for neural text classification. In **Proceedings of the 57th Annual Meeting of ACL**, pp. 1620–1629, 2019.

[2] Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. In **Findings of ACL: EMNLP 2020**, pp. 4617–4624, 2020.

[3] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. BioMegatron: Larger biomedical domain language model. In **Proceedings of the 2020 Conference on EMNLP**, pp. 4700–4706, Online, November 2020. Association for Computational Linguistics.

[4] Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. Optimizing word segmentation for downstream task. In **Findings of ACL: EMNLP 2020**, pp. 1341–1351, Online, November 2020. Association for Computational Linguistics.

[5] Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. Joint optimization of tokenization and downstream model. In **Findings of ACL: ACL-IJCNLP 2021**, pp. 244–255, 2021.

[6] Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuan-Jing Huang. Long short-term memory neural networks for chinese word segmentation. In **Proceedings of the 2015 conference on EMNLP**, pp. 1197–1206, 2015.

[7] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. **IEEE Computational intelligence magazine**, Vol. 13, No. 3, pp. 55–75, 2018.

[8] Hiroyuki Deguchi, Masao Utiyama, Akihiro Tamura, Takashi Ninomiya, and Eiichiro Sumita. Bilingual subword segmentation for neural machine translation. In **Proceedings of the 28th COLING**, pp. 4287–4297, 2020.

[9] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on EMNLP: System Demonstrations**, pp. 66–71, 2018.

[10] Masaaki Nagata. A stochastic japanese morphological analyzer using a forward-dp backward-a* n-best search algorithm. In **Proceedings of the 15th conference on COLING**, pp. 201–207. Association for Computational Linguistics, 1994.

[11] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **Proceedings of the 56th Annual Meeting of ACL**, pp. 66–75, 2018.

[12] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In **Proceedings of the 58th Annual Meeting of ACL**, pp. 1882–1892, Online, July 2020. Association for Computational Linguistics.

[13] Tatsuya Hiraoka. Maxmatch-dropout: Subword regularization for wordpiece. In **Proceedings of the 29th COLING**, pp. 4864–4872, 2022.

[14] Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, and Isaac Okada. Incorporating word attention into character-based word segmentation. In **Proceedings of the 2019 Conference of NAACL: HLT**, pp. 2699–2709, 2019.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[16] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In **Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.**, Vol. 4, pp. 2047–2052. IEEE, 2005.

[17] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. **IEEE transactions on Information Theory**, Vol. 13, No. 2, pp. 260–269, 1967.

[18] Yu Suzuki. Filtering method for twitter streaming data using human-in-the-loop machine learning. **Journal of Information Processing**, Vol. 27, pp. 404–410, 2019.

[19] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In **Proceedings of the 2021 Conference of NAACL: HLT**, pp. 2095–2104, Online, June 2021. Association for Computational Linguistics.

[20] Yongfeng Zhang, Min Zhang, Yi Zhang, Guokun Lai, Yiqun Liu, Honghui Zhang, and Shaoping Ma. Daily-aware personalized recommendation based on feature-level time series analysis. In **Proceedings of the 24th international conference on WWW**, pp. 1373–1383, 2015.