学会記事

“Character-to-Word Attention for Word Segmentation”
の研究過程

東山 翔平†

1 はじめに

本稿では我々の単語分割の研究を紹介する。本研究は、国際会議版論文(Higashiyama et al. 2019)がNAACL-HLT 2019に採択され、更に表題の論文(Higashiyama et al. 2020b)が論文誌「自然言語処理」に採択され、言語処理会論文賞を受賞した。本稿は「研究の過程を含めた論文紹介記事」という趣旨で、実際の研究の進行過程に注目して研究を紹介する。特にこれから自然言語処分手例で論文投稿を目指す方にとって、本記事に参考になる点があれば幸いである1.

2 単語分割とその課題

単語分割は、（この文のように）分かち書きされていない文を単語の列に分割するタスクである。日本語の自然言語処理では、単語分割、品詞付与、原型推定を合わせて形態素解析とも呼ぶ。言語解析処理をパイプライン的に接続して応用タスクに繋ぐ伝統的な自然言語処理において、単語分割は最初のステップとしてほとんど必須とされてきた。形態素解析器としては、MeCab2など高精度かつ高速なツールが2000年代には存在している。しかし、単語分割の課題は現在も存在し、主なものとして、目標ドメインへの適用時、未知語や文脈依存の曖昧性などの要因によって解析精度が低下することが挙げられる。

学習コーパスとドメインが異なる専門分野のテキストや、新語や崩れた表記が頻繁に現れるユーザ生成テキスト(SNSの投稿など)を解析する場合、未知語の問題が顕著となる。Higashiyama et al. (2020b)では、新聞記事等で学習したモデルを、特許やレシピなど別ドメインのテキストに適用した際、分割精度(F値)が2〜4ポイント程度低下することを示している。

他の観点として、単語分割と、サブワード分割などの統計的分割法とを応用に応じてどう使い分ける（併用する）のが最適か明らかでないという点がある。近年は、ニューラルネットワーク（以下，NN）によるend-to-endの手法が発展し、単語という人間に有意義な単位を介さなく

† 国立研究開発法人情報通信研究機構 (National Institute of Information and Communications Technology)
1 本手法のソースコードは次で公開している。https://github.com/shigashiyama/seikanlp
2 https://taku910.github.io/mecab/
とても、応用タスクで高い精度が達成される事例もある3。一方、依存構造、述語項構造、照応などの統語・意味的な言語解析処理を行う場合には、処理単位（をまとめ上げた単位）が人間に有意味な単位と一致していなければ意味をなさず、これらのタスクでは依然として単語が処理単位であることが要求される。その他にも、単語分割の粒度や、同一文字列に対する分割の一貫性が応用タスクへどう影響するかといった点も問題になり得るが、ここでは観点を挙げるのみとして、更なる研究によって調査され明らかになっていくことを期待したい。

3 本研究の概要

本研究では、単語分割において、単語境界の曖昧性解消に単語の情報が有効という仮説の下、文字単位の系列ラベルの枠組みに注意機構に基づいて単語の情報を組み込んだNNベースの方法を提案した。図1に示す通り、提案手法のモデル構造は、文字のベクトル表現を得るための文字埋め込み層と双方向LSTM (Long Short-Term Memory) 層、単語のベクトル表現を得るための単語埋め込み層と注意機構、2種のベクトルからラベルを予測するCRF (Conditional Random Fields) 層からなる。

例として、“彼は日本人”という入力文では、文字“本”を含む部分文字列として“日本”、“本”、“本人”などの単語の候補が存在する。提案手法では、これらの文字および単語に対応するベクトル表現を取得し、文字から単語への注意（重み）を計算することで、文脈に応じた

図1 提案手法のモデル構造

3 たとえば、ユニグラム言語モデルによる分割を用いた機械翻訳 (Kudo and Richardson 2018) など。
単語候補の重要度を考慮して適切な分割ラベルの予測に役立てる。
日本語データセットを用いた実験により、本研究では以下のことを示した。
- 単語ベクトルの統合法として、注視機構を用いた方法で分割精度が向上し、特に重み付け連絡（WCON）で最良となることを示した。
- 自動分割テキストから学習した単語埋め込みベクトルを用いすることで未知語の問題を緩和し、学習時と異なるドメインに対しても従来手法と比べ頑健な精度を示した。
- 学習済みモデルの注意の重みを確率的に制御し、正しい単語へ注意する割合が高まるにつれて分割精度が単調に上昇し、単語への適切な重みの学習が正確な分割に貢献することが示唆された。
加えて、モデルパラメータ数、単語語彙に含める単語の最大長と最小出現頻度、ラベルなしテキストの使用の有無／使用法の違い／ドメインの違いと、それらの分割精度への影響をそれぞれ評価し、多面的な分析を行った。

4 研究の過程

本研究は、着想から論文採択まで一筋縄では進まなかった。一連の流れがわかるよう、三つの段階に分けつつ、時系列に沿って記述したい。

モデル開発試行 2017年5月、著者は現研究室に着任し、研究テーマを模索していた。NNを用いた高精度な形態素解析器の要望があったこともあり、単語分割にテーマを定めた。この時点でのNN（のみ）を用いた単語分割の研究は、日本語を対象としたものは查読のない国内会議文献が少数あるのみで5。中国語では先行研究が多数存在している状況であった。
それから1〜2か月の間に、第1のアイデアを考えた。文字単位の系列ラベルリングを採用し、推論時に、辞書を基に各入力文の単語ラティス構造を構築して、推論パスを制限する方法である。非NNの日本語形態素解析器で採用されているラティスに基づく方法に着想を得て、その利点を取り入れる意図であった。つまり、妥当性の低いパスを排除できる利点（さらに単語情報も利用できる利点）があった。この手法を実装したところ、多少の精度向上は確認できたが、ベースライン RNN (Recurrent Neural Network) モデルの10倍以上処理時間が必要であるという致命的な欠点があり、実装入となった。文ごとにラティス構造が異なるためミニバッチ化が難しく、オンライン処理を行ったためである。
次に第2のアイデアとして、文字単位の系列ラベルリングにおいて明示的に単語の情報を使用

---

4 本研究では BIES 形式の分割ラベルを採用した。“B”, “I”, “E” はそれぞれ該当文字が単語の先頭、内部、末尾であることを意味し、“S” は該当文字単独で１単語を構成することを意味する。
5 それ以降に公開された日本語単語分割の論文としては、(Kitagawa and Komachi 2018; Tolmachev et al. 2020) などがある。
る方法を考えた。品詞付与などの先行研究では、単語と文字両方を用いることで精度が向上している。単語分割では単語が確定していないとはいえ、文字に加えて妥当性の高い単語の情報を用いるのは自然な考えに思えた。著者の記録によると、同年10月には、注意機構を利用して、文字ごとに可能な単語への重みを計算する方法を考えていた。しかし、別件の開発作業に着手する事情があり、この研究は一旦保留となった。なお、同年11月、中国語単語分割で、文字単位で単語情報を用いる論文(Wang and Xu 2017)が発表されヒヤリとしたが、幸い前述のアイデアにおける注意機構を用いる点は彼らに濟んだ。

国際会議への投稿　2018年1月、研究を再開した。前述の第2のアイデアを試そうと、ベーシックラインモデルに単語ベクトルを導入した。単語ベクトルを平均する方法(AVG)で精度向上を確認したものの、そこから注意機構を加えても顕著な差は得られなかった。しかし、正解の単語にのみ注意させる“チート”の設定では大幅な精度向上が確認できたため、方向性は間違いしていないと判断し、試行錯誤を重ねた。結局、注意機構を、先行研究(Wang and Xu 2017)で使われている連結と組み合わせた方法(WCON)により、中国語単語分割ベンチマークデータの一つでstate-of-the-art(SOTA)を超える精度を達成した。この結果を基に国際会議論文を書くことにした。

5月、EMNLP 2018 に論文を投稿した。著者が自然言語処理の関連会議に投稿するのは初めての機会であったが、関連会議での採択経験が豊富な当研究室の熟練研究者の方々に助言を頂いて、論文のストーリーとしては説得力のあるものに改善できたのではないかと思う。ただし、査読者の3名によるスコアは5段階評価で2.3.4で、主な否定的なコメントとして、提案手法(WCON)と比較手法(CONや他の先行研究)との精度の差が小さいという点があった。Rebuttalにて反論を捻り出すも、結果は不採択だった。

8月、前述の査読者のコメントを踏まえて論文を修正し、AAAI 2019 に投稿した。査読者のスコアは10段階で4.6.6で、否定的なコメントには、類似する関連研究が引用されていないという指摘があった。著者にとって当該研究との違いは明確で、反論は容易であった。しかし、時差を勘違いしてrebuttalの登録期限を過ぎるという非常に初歩的なミスを犯したことも影響してか、結果は再び不採択となった。

12月、論文を再度修正し、NAACL-HLT 2019へ投稿した。同年10月に発表された論文で、中国語単語分割のSOTA精度が更新され、手間取っている間に状況は苦しくなっていた。この手法の論文を関連会議に投稿する機會はこれが最後という思いで臨んだが、2019年2月、採択の報せが届いた。査読者のスコアは6段階で5.4.4であった。

著者の限られた経験から帰納すると、論文の質が採択・不採択の境界付近に位置する場合、

---

6 文字単位の単語分割において、該当文字を含む文字1-gramから4-gramに相当する単語のベクトルを連結(concatenation)して一つの単語ベクトルとして用いる方法。

7 本稿執筆時点で、関連国際会議に4回投稿し、うち2回採択された。
採取されるものとされないものとの差は微妙な差であり、異なる要素（他の投稿論文の数と質、査読者の割り当て）も絡む。論文の質の要因を、提案手法（またはデータセット、評価尺度、その他アイデア）自体の良さと論文の文章の質とに分解すると、前者については、時間的な制約もある中で手法の中核部の改善のために着想・実装・検証の過程をいくらでも繰り返してはいられず、適当な時点で折り合いをつけずのを得ないのではないかだろう。一方、後者は、論理展開の妥当性や英文の質など、執筆者の力量に依存し、訓練・経験で高められる余地が大きい。

論文誌への投稿　2019年8月頃、次の研究テーマとして単語分割のドメイン適応の問題に取り組んでいた。そこで、先の提案手法について、学習時と異なるドメインでの性能評価を行ったため、新しい実験結果を含めて国際会議版論文を加筆修正し、11月に論文誌「自然言語処理」に投稿した。2020年1月、査読結果が通知され、2名の査読者の判定はそれぞれ条件付採録と照会後判定であった。論文誌では追加実験などが認められる一方で、緻密な評価が要求される。査読者の指摘は、査読要因に対するモデル性能の追加評価を求めるのが中心で、一つ一つ着実に対応して論文を修正・再投稿し、4月に無事採択されるに至った。

本研究が成仏したことにより、2021年1月、忘れられた頃に論文賞受賞との報せがあった。著者にとっては思いがけない歓喜であったが、査読時の指摘の恩恵も多分に受けて内容が充実し、多面的な観点での評価・分析を行った点を評価いただいたのかと解釈している。

5 おわりに

本研究の詳細に興味を持たれた方は（Higashiyama et al. 2020b）を、またその後に行った単語分割のドメイン適応の研究に興味を持たれた方は（Higashiyama et al. 2020a）を参照いただきたい。

著者は、次の研究として、ユーザ生成テキストにおける単語分割と崩れた表記の正規化の問題に取り組んでいる。同テキストは、多様性に富んだ大量のユーザの意見が入るできる有用な情報源として関心が高まってから10年程度が経つが、現在も十分な解析・認識が実現されているとは言い難い状況である。固有名や新語の多さ、崩れた表記などの特徴が解析を困難にしている反面、人間の言語表現の自由さ、柔軟さを窺い知れる点で非常に興味深い言語媒体と感じている。著者は、2020年代現在の自然言語処理・機械学習技術の可能性に期待しつつ、今後の研究でこの領域の発展に多少なりとも貢献したいと考えている。

注8 2011年の言語処理学会年次大会では、従来の整ったテキストと異なるウェブ上の言語表現を対象とした処理を指して「不自然言語処理」と題するテーマセッションが設定され、翌年にも情報処理Vol.53 No.3や自然言語処理Vol.19 No.5で同様の趣旨の特集が組まれた。
謝 辞

本研究に関して、論文の査読、論文賞の選考をしていただいた編集委員会、査読者、関係者の皆様に感謝申し上げます。また、本研究に助力・コメントいただいた皆様に感謝申し上げます。

参考文献

Higashiyama, S., Utiyama, M., Matsumoto, Y., Watanabe, T., and Sumita, E. (2020a). “Auxiliary Lexicon Word Prediction for Cross-Domain Word Segmentation.” Journal of Natural Language Processing, 27 (3), pp. 573–598.

Higashiyama, S., Utiyama, M., Sumita, E., Ideuchi, M., Oida, Y., Sakamoto, Y., and Okada, I. (2019). “Incorporating Word Attention into Character-Based Word Segmentation.” In Proceedings of NAACL-HLT, pp. 2699–2709.

Higashiyama, S., Utiyama, M., Sumita, E., Ideuchi, M., Oida, Y., Sakamoto, Y., and Okada, I. (2020b). “Character-to-Word Attention for Word Segmentation.” Journal of Natural Language Processing, 27 (3), pp. 499–530.

Kitagawa, Y. and Komachi, M. (2018). “Long Short-term Memory for Japanese Word Segmentation.” In Proceedings of PACLIC, pp. 279–288.

Kudo, T. and Richardson, J. (2018). “SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing.” In Proceedings of EMNLP, pp. 66–71.

Tolmachev, A., Kawahara, D., and Kurohashi, S. (2020). “Design and Structure of The Juman++ Morphological Analyzer Toolkit.” Journal of Natural Language Processing, 27 (1), pp. 89–132.

Wang, C. and Xu, B. (2017). “Convolutional Neural Network with Word Embeddings for Chinese Word Segmentation.” In Proceedings of IJCNLP, pp. 163–172.

略歴

東山 翔平：2012年、神戸大学工学部情報工学科卒業。2014年、同大学院システム情報学研究科博士課程前期課程修了。2014年に日本電気株式会社に入社し、2017年から国立研究開発法人情報通信研究機構に出向、2019年より同機構に有期研究技術員として在籍。同年より奈良先端科学技術大学院大学先端科学技術研究科博士後期課程に在学。自然言語処理の研究に従事。言語処理学会会員。