品詞解析の学習者英語への分野適応

永田　亮†,††　水本　智也†††　菊池　悠太††††　川崎　義史†††††　船越孝太郎††††††
† 甲南大学　†† 国立研究開発法人科学技術振興機構さきがけ　††† 理化学研究所 AIP センター
†††† 株式会社 Preferred Networks　††††† 東京大学　†††††† 京都大学
E-mail: †nagata-nlp2018@ hyogo-u.ac.jp.

1. はじめに

品詞解析は，言語処理や言語の分析に広く使われているが，学習者英語への分野適応に関する研究は少ない．よく用いられるのは，品詞ラベル付き学習者英文を訓練データに単純に加える方法 [1], [10] である．しかしながら，利用可能な品詞付き学習者英文の量は非常に限られているため，適用範囲が狭い．この制限のため，学習者英語を扱う研究では，代わりに，母語話者の英文用に開発された品詞解析器を使うことが多い．例えば，文法誤り訂正 [4] とその自動評価 [2]，エッセイの自動採点 [3] など，数多くの例がある．

しかしながら，母語話者向けの品詞解析器では，学習者の英文を十分な精度で解析できない可能性がある．その理由として，学習者特有の言語現象を挙げることができる．例えば，

*Becose/NNP I/CD like/IN reading/NN ,/,
I/PRP want/VBP many/JJ Books/NNPS ./.

という英文を考えてみよう(注 1)．この例から，綴りに関する誤りと文法誤りが解析ミスを引き起こすことがわかる．更に，誤りがある単語だけではなく，周辺単語の解析ミスにつながることもわかる．別の要因として，母語の影響により，品詞の分布や品詞列の分布が異なる [9] ことも挙げることができる．これも品詞解析の性能低下につながる可能性がある．言い換えれば，母語ごとに品詞解析の分野適応が必要であることを示唆する．

このような背景を考慮し，本稿では，品詞解析の学習者英語への分野適応について検討を行う．具体的には，母語話者向けの品詞解析で高い性能が報告されている深層学習に基づいた手法を用いると，品詞付き学習者コーパスなしでも，効率的かつ効果的に学習者英語への分野適応が可能なことを示す．更に，学習者の母語へ品詞解析を適応させる方法についても検討する．これらの分野適応が品詞解析の性能へ及ぼす効果を実験により明らかにする．また，解析結果の分析により，分野適応が効果的である学習者特有の言語現象とそうでないものを明らかにする．

(注 1)：Stanford CoreNLP 3.8.0（https://stanfordnlp.github.io/CoreNLP/）を用いて解析した結果である（解析ミスを太字で表示）．

2. 品詞解析モデル

母語話者英語向けの品詞解析に関する従来研究 [6], [7] を参考にし，本研究では，図 1 に示すようなモデルを検討した．入力となるのは，文中の各単語と書き手の母語の情報である．これらの情報が，図中下部に示される Embed layer により，ベクトルに変換される．続いて，そのベクトルは Bi-directional LSTM（BLSTM）に渡される．最後に，Softmax layer において，各単語が取りうる品詞（の確率）が推定される．

Embed layer は，単語／文字列／母語モジュールの三つのモジュール（ネットワーク層）から成る．それぞれ，単語自身，単語内の文字列，母語の情報をベクトルに変換する．

単語モジュールは，単語の表層形に基づいて単語の情報をベクトル（分散表現）に変換する層である．ただし，ネットワークのパラメータを減らすため，全て小文字に変換した単語を入力とする．また，頻度が閾値未満の単語は未知語を表す特殊な記号に変換したのち分散表現を得ることとする．単語の分散表現により，用法や意味が似た単語は，似た値をもつベクトルに変換される．分散表現のこの特性は，品詞解析における綴り誤りからの影響を低減させるのに効果的であると期待できる．

ここで重要なのは，単語モジュールは，品詞付き学習者コーパスを用いずに事前学習が可能であるということである．品詞情報なしの学習者コーパスを母語話者コーパスに加えて事前学習を行うことで，綴り誤りも含めて分散表現の訓練が行われる．幸いなことに，品詞情報なしの学習者コーパスであれば豊富に利用可能である．更に，場合によっては，品詞解析の対象となる文書自身も含めた事前学習が有効な場面もある．すなわち，事前学習用のコーパスに，解析対象の文書（当然，品詞の情報はない）を加えて単語モジュールの事前学習を行う．これにより，解析対象の文書に出現する単語が考慮され，更なる解析精度の向上が期待できる．ただし，事前学習には時間を要するため，高い解析速度が求められる場面では用いることができない．しかしながら，資格試験などの自動採点やコーパスの言語学的分析など，学習者英語を扱う研究ではある程度時間をかけられる場合も多い．

文字列モジュールは，単語内の文字列に基づいて単語の情報をベクトルに変換する．文字列に基づくことで，単語モジュールを補完できる可能性がある．低頻度語に対しては，

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

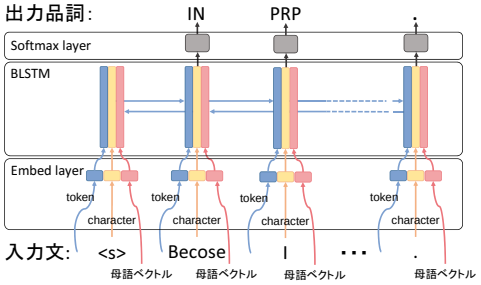

図 1： 品詞解析のためのモデル.

単語の分散表現がうまく学習できないことが多い．極端な場合，未知の綴り誤りのように訓練データに一度も出現しない単語に対しては，単語の分散表現を得ることができない．一方で，単語内の各文字に基づけば，未知の文字が含まれない限り，そのような単語についてもベクトルへの変換が可能となる．具体的には，図 2 に示すように，まず，単語内の各文字を分散表現に変換し，BLSTM(注 2)に順次入力することで，最終的なベクトルを得る．なお，文字の分散表現も事前学習を行う．また，単語モジュールとは異なり，大文字の小文字への変換は行わない．

母語モジュールは，母語の情報を分散表現に変換し品詞解析に利用するモジュールである．類似した母語は，類似した分散表現に変換されると期待される．このことにより，ある言語（例：フランス語）話者の訓練データの情報を類似した言語（例：スペイン語）話者が書いた英文の品詞解析に効果的に利用できる可能性がある．母語の分散表現は，従来研究の成果として利用可能なものが存在する．例えば，Malaviya ら [8] は，機械翻訳を通じて母語（正確には各言語）の分散表現を学習する手法を提案している．その成果として 1000 言語以上に対する分散表現が公開されている．本稿では，この言語の分散表現を母語モジュールに利用する．なお，この分散表現は訓練中固定する．代わりに，Embed layer と上層の BLSTM の間に線形変換のための全結合層を一層設ける．

以上の三つのモジュールの出力を連結して，上層の BLSTM への入力とする．連結された入力ベクトルは，品詞解析対象文中の一つの単語に対応する．単語モジュールと文字列モジュールにより，単語自身の情報と単語内の文字列の両方を考慮して単語の情報を符号化していると解釈できる．更に，母語モジュールにより，母語を考慮した単語の情報の符号化をしていると解釈できる．これらの情報は，BLSTM に送られ，母語適応した品詞解析が実現される．

3. 性能評価

2. で述べた品詞解析モデルの性能評価を行った．まずは，母語の情報なしに，どの程度学習者英語に適応できるかを評

(注 2)：上層の BLSTM とは別の BLSTM である.

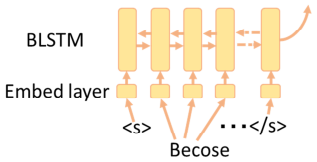

図 2： 文字列モジュール：<s> と </s> は，それぞれ語頭，語末を表す特殊な記号である.

価した．具体的には，2. の品詞解析モデルから母語モジュールを除いたものを用いた．このモデルを Penn Treebank Wall Street Journal (WSJ (注 3)) のセクション 00–18 で訓練した．ハイパパラメータの選択は，Treebank of Learner English (TLE [1]) の開発データを用いて行った(注 4)．また，単語／文字の分散表現の訓練には次の母語話者コーパスと学習者コーパスを用いた：EWT(注 5), 独自に収集した英語教材，ICLE [5], ETS(注 6), NICE(注 7), Lang-8 Learner Corpora(注 8)．分散表現の学習には Word2vec(注 9)を用いた．以上のように，訓練されたモデルの精度を Konan-JIEM Learner Corpus 第 5 版 [10](以降，KJ と表記) で評価した．KJ には表層の品詞と文脈の品詞の二種類が収録されているが，後述の実験と条件をそろえるために表層の品詞を対象にして評価を行った．

比較のために，Conditional Random Field (CRF) に基づいた手法も実装した．用いた素性は次のとおりである：単語の表層形，原型，接頭辞，接尾辞，特殊な文字（数字，大文字，記号それぞれ）が存在するかどうかを表す 2 値．文脈窓幅は，対象単語の前後五単語とした．また，一次のマルコフモデルを仮定した．ハイパパラメータは TLE の開発データを用いて決定した．また，綴り誤りから受ける影響を調査するため，綴り誤りを訂正した KJ とオリジナルの KJ の両方に対して性能を評価した．

表 1 に結果を示す．表 1 から，品詞付き学習者コーパスを使用せずとも，深層学習に基づいた品詞解析モデルが学習者英語にうまく適応していることがわかる．興味深いことに，

(注 3)：Marcus, Mitchell, et al. Treebank-3. LDC, 1999.
(注 4)：BLSTM のドロップアウトレートは 0.5 とした．各層のユニット数は，分散表現の次元数に従う．ただし，母語モジュールを使用するモデルについては，Malaviya ら [8] より提供されている 512 次元の言語の分散表現を固定で使用し，線形変換のための全結合層のユニット数は 200 とした．最適化アルゴリズムは Adam を使用し，パラメータをステップサイズ 0.01，一次のモーメント 0.9，二次のモーメント 0.999 とした．各モデル，20 エポックの学習を行い，開発データに対する解析精度が最も高いものを用いた.
(注 5)：Bies, Ann, et al. English Web Treebank. LDC, 2012.
(注 6)：Blanchard, Daniel, et al. ETS Corpus of Non-Native Written English. LDC, 2014.
(注 7)：http://sgr.gsid.nagoya-u.ac.jp/wordpress/?page_id=695
(注 8)：http://cl.naist.jp/nldata/lang-8/
(注 9)：https://github.com/dav/word2vec. 各ハイパパラメータは次の通りとした：分散表現の次元：200（単語），50（文字）；文脈窓幅：5；低頻度語に対する閾値：5 未満.

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

表 1： KJ に対する品詞解析精度.

モデル	精度
本モデル（母語モジュールなし）	0.950
CRF（綴り誤り訂正あり）	0.942
CRF（綴り誤り訂正なし）	0.940

正しい綴りの情報が与えられた CRF の性能よりも更に性能がよい；本モデルと CRF（綴り誤り訂正あり）の解析精度差は有意水準 1%で有意である（母比率の差の検定）．このことは，母語モジュールなしモデルが綴り誤り以外にも学習者英語特有の現象にうまく対処できていることを示唆する．

TLE [1] に対しても性能評価を行った．Berzak らの報告 [1] に合わせて，EWT（母語話者コーパス）と TLE（学習者コーパス），それぞれの訓練データを用いて訓練を行った．この評価実験では，母語モジュールあり／なしの二種類のモデルを評価した．また，学習者コーパスを用いずに(注 10)EWT のみで訓練したモデルの性能も評価した．それ以外の条件は，KJ の際と同一とした．なお， TLE では，十種類の母語話者（中国語，フランス語，ドイツ語，イタリア語，日本語，韓国語，ポルトガル語，スペイン語，ロシア語，トルコ語）の英文が含まれる．また，Penn Treebank POS と Universal POS の両方の情報を収録するが，上述の実験との整合性を考え，前者のみを対象とした．

表 2 に，結果を示す．表 2 には，同じ訓練，開発，評価データを用いた Turbo Tagger の性能(注 11)も示している．表 2 より，TLE に対しても本モデルは同様に高い解析精度を示すことがわかる；本モデル（母語モジュールあり，学習者コーパスあり）と Turbo Tagger（学習者コーパスあり）の性能差は 1%の有意水準で有意である（母比率の差の検定）．また，品詞付き学習者コーパスが与えられない場合でも Turbo Tagger より高い性能を達成する．同一の訓練データ（品詞付きコーパス）を用いているにも関わらず，Turbo Tagger より本モデルの性能が高いということは，単語／文字の分散表現が効果的であると解釈できる．

一方，我々の予想に反して，母語の情報のあり／なしによる性能差はほとんどない．言い換えれば，母語モジュールのように，母語に直接的に適応させる方法と（母語の情報なしの）品詞付き学習者コーパスを訓練データに加えるという単純な方法との間には性能差はみられない．

4. 解析結果の分析と考察

品詞付き学習者コーパスなしでも深層学習ベースの品詞解析モデルが高い性能を示す理由を明らかにするため，KJ に対する解析結果の分析を行った．具体的には，母語モジュールなしの深層学習ベースのモデル（以降，本モデルと表記する），CRF に基づいた手法（綴り誤り訂正あり／なし）の解析結果の差分を目視で確認することで分析した．

(注 10)：したがって，母語モジュールも用いていない．

(注 11)：性能値は，文献 [1] から引用した．

表 2： TLE の評価データに対する品詞解析精度.

モデル	精度
本モデル（母語モジュールあり，学習者コーパスあり）	0.964
本モデル（母語モジュールなし，学習者コーパスあり）	0.963
Turbo Tagger（学習者コーパスあり）[1]	0.958
本モデル（学習者コーパスなし）	0.951
Turbo Tagger（学習者コーパスなし）[1]	0.943

まず，予想通り，綴り誤りの影響を低減できていることが確認できた．綴り誤り訂正なしの CRF では，綴り誤りのある単語に適切でない品詞を割り当てることが散見された: (例：famouse/NN，thier/JJR，shoud/VBD，goot/NN，foward/NN)．この例のように，綴り誤りがある単語では，基本的には名詞（NN）もしくは接頭辞，接尾辞により他の品詞に誤って解析される．一方，本モデルでは，正しく解析できることが多い．上述の例は，全て正しく “famouse/JJ”，“thier/PRP$”，“shoud/MD”，“goot/JJ”，“foward/RB” と解析された．これらの例は，単語の分散表現が得られた綴り誤りの例である．このことは，品詞情報が付与されていない通常の学習者コーパスから得られた単語の分散表現が，綴り誤りのある単語の品詞解析に有効であることを示す．

第二に，文字列モジュールの効果も確認できた．低頻度であるため分散表現が得られなかった綴り誤りでも正しく解析できる例が確認された．これは，文字列に基づいた単語の情報（すなわち，文字列モジュール）が期待通りの効果を発揮していると分析できる．具体例として，“ranchi/NN”，“eatig/VBG”，“dilicuse/JJ”，“beutihure/JJ”(注 12)などが確認された．これらの綴り誤りは頻度が 5 未満であり，単語の分散表現は得られなかった．にもかかわらず，本モデルでは正しく解析できた．一方，接頭辞，接尾辞の情報は品詞に関する情報を与えない（もしくは間違った情報を与える）ため，綴り誤り訂正なしの CRF では解析に失敗した(注 13)．以上のことから，単純に接尾辞，接頭辞の情報を用いるより，本モデルのように，文字の分散表現を通じた BLSTM による単語情報の符号化のほうが，綴り誤りの影響を低減させるといえる．しかしながら，綴り誤りが重度(注 14)である場合，解析に失敗することも観察され，更なる改善が望まれる．例えば，“plople” や “resentuly”（正しくは “people” と “recently”）がある．

第三に，母語話者コーパスと学習者コーパスにおける品詞

(注 12)：正しい綴りは，それぞれ，“lunch”，“eating”，“delicious”，“beautiful” である．

(注 13)：ただし，接頭辞，接尾辞の情報を用いないと，解析に失敗する例も多数あるため，全体の性能は 1%程度低下する．

(注 14)：編集距離が大きい，もしくは文字の置換確率が低いという観点で重度であるという意味である．

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

の分布の差異もある程度吸収できることが判明した．例えば，“like” という単語は，少なくとも動詞と前置詞の二種類の品詞で使用されるが，訓練に用いた母語話者コーパス（WSJ）では，前置詞としての使用が多く，全体の 82%が該当した．一方，学習者コーパス（KJ）では，前置詞としての使用は 5%で，大部分（94%）については動詞としての使用であった．このように，品詞の分布に差異があると解析ミスを引き起こすことがある．実際，KJ には，304 回 “like” が出現したが，本モデルと綴り誤り訂正ありの CRF，それぞれの解析性能は 0.927 と 0.635 と大きな差がある．本モデルでは，単語の分散表現を通じて，“like” は動詞の文脈で使用されることが大多数であることが学習されるため解析精度が高いと分析できる．同様な例として，“fight” における動詞と名詞の使い分けも確認できた．

興味深いことに，文法誤りからの影響の緩和も観測された．特に，文頭の冠詞の抜けの影響の緩和が大きい．これは次のように説明される．文頭の冠詞が抜けると，次に来る単語の先頭の文字が大文字で綴られる（例：“The flower is ···” → “Flower is ···”）．この影響を受け，CRF に基づく手法では，“Flower” を固有名詞と誤って解析する．一方，学習者コーパスでは，冠詞の抜けが頻出するため，語頭が大文字で始まる一般名詞の “Flower” も出現する．また，小文字で始まる単語の文頭での出現，逆に，大文字で始まる（固有名詞以外の）単語の文中での出現もある．更に，無冠詞単数で出現するため，見かけ上，“flower” は不可算名詞のように観測される．これらの出現が単語／文字の分散表現に反映されると，文頭の “Flower” を一般名詞として解析することが可能になる．なぜなら，不可算名詞であれば，文頭で無冠詞単数で使用可能であるからである（例：Water is abundant.）．この種の誤りは，“Town”，“Mountain” などの単語で見られた．

語順の誤りに対する効果もみられた．例えば，“*abroad/RB travel” や “*I very/RB enjoyed.” などの語順の誤りを正しく解析できていた．この理由として，分散表現を通じた BLSTM による単語列の符号化の効果を挙げることができる．分散表現と BLSTM により単語列の情報を符号化することで，語順に対してある程度寛容になると予想される．一方で，CRF に基づく手法では，これら二つの表現に対して，全く異なる素性が割り当てられるため，語順の変化に対して敏感である．

最後に，母語モジュールに効果がないことを考察する．効果がないことを再確認するために，解析対象と同じ母語が訓練データに含まれないように（すなわち，leave-one-native-language-out cross validation）して，モデルを再度訓練した．こうすることで，類似した母語の訓練データの効果がより際立ち，母語モジュールあり／なしの性能差が大きくなると予想される．しかしながら，母語モジュールあり／なし，それぞれの性能は，0.966 と 0.965 となり，やはり効果がないことが確認された．この理由としては，(i) 訓練／評価データの量が足りない，(ii) 母語によらず学習者に共通した言語現象の影響が大きい（言い換えれば，母語固有の解析ミスが少ない），(iii) 母語モジュールなしのモデルで，母語モジュールのような機構が自動的につくられている（例えば，BLSTM のユニットのいくつかで母語の推定や類似母語の判別が行われている）などが考えられる．どれが正しいか特定するには，更なる調査が必要である．

5. おわりに

本稿では，品詞解析の学習者英語への分野適応として，深層学習に基づいたモデルの検討を行った．単語，文字の分散表現，BLSTM を用いると，品詞付き学習者コーパスなしでも，効率的かつ効果的に学習者英語への分野適応が可能なことを示した．実験結果の分析により，綴り誤り，品詞の分布の違い，文法誤りに対して頑健であることを明らかにした．一方で，母語ベクトルを通じて，学習者の母語へ品詞解析を適応させる方法については期待された効果が見られなかった．今後は，この点について調査する予定である．

謝　辞

本研究の一部は，JST，さきがけ，JPMJPR1758 の支援をうけて実施したものである．

参考文献

[1] Y. Berzak, J. Kenney, C. Spadine, J.X. Wang, L. Lam, K.S. Mori, S. Garza, and B. Katz, “Universal dependencies for learner English,” Proc. of ACL, pp.737–746, 2016.
[2] C. Bryant, M. Felice, and T. Briscoe, “Automatic annotation and evaluation of error types for grammatical error correction,” Proceedings of ACL, pp.793–805, 2017.
[3] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M.D. Harris, “Automated scoring using a hybrid feature identification technique,” Proc. of ACL, pp.206–210, 1998.
[4] M. Chodorow and C. Leacock, “An unsupervised method for detecting grammatical errors,” Proc. of 1st NAACL, pp.140–147, 2000.
[5] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot, International Corpus of Learner English v2, Presses universitaires de Louvain, Louvain, 2009.
[6] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” 2015.
[7] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF,” Proc. of ACL, pp.1064–1074, 2016.
[8] C. Malaviya, G. Neubig, and P. Littell, “Learning language representations for typology prediction,” Proc. of EMNLP, pp.2529–2535, 2017.
[9] R. Nagata and E. Whittaker, “Reconstructing an Indo-European family tree from non-native English texts,” Proc. of ACL, pp.1137–1147, 2013.
[10] R. Nagata, E. Whittaker, and V. Sheinman, “Creating a manually error-tagged and shallow-parsed learner corpus,” Proc. of ACL, pp.1210–1219, 2011.

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.