スペル誤りに着目した学習者の英語に対する品詞タグ付け・チャンキングの性能調査

水本智也
東北大学
tomoya-m@ecei.tohoku.ac.jp

永田亮
甲南大学
nagata-nlp @ hyogo-u.ac.jp.

1 はじめに

文法誤り訂正や小論文自動採点のような学習者の英語を対象とした自然言語処理タスクにおいて，品詞タグ付けやチャンキングは重要な役割を果たす．これらの基礎解析技術は上で挙げたタスクを解くために必要な言語学的特徴を抽出するために使われる．例えば，文法誤り訂正の性能を競う CoNLL Shared Task 2014 [7] では，参加した 12 チーム中 10 チームが品詞タグ付け，チャンキングのどちらか，もしくはその両方を使用している．

品詞タグ付けやチャンキングのような基礎解析技術は，学習者の英語に対する言語学的分析のためにも活用されている．例えば，Aarts は学習者の英語における特徴的な品詞のパターンを発見するために品詞タグ付けを使用している [1]．Nagata [5] らは品詞タグ付けによって得られた品詞系列が母語を識別するために有効であることを示した．

基礎解析技術の解析ミスが自然言語処理システムや言語分析の性能を下げている [4, 10]．例えば，与えられた文に対する名詞句の同定の失敗は，冠詞や名詞の単数複数のような誤りに対する訂正の失敗を引き起こす．また，品詞タグ付けやチャンキングの失敗は，上で言及したような言語学的分析の研究でよく用いられる品詞やチャンクの出現の単純なカウントやそれらの分布の推定に悪影響を与える．

英語学習者の文を対象とした応用タスクで品詞タグ付けやチャンキングが重要であるにも関わらず，学習者の文に対する品詞タグ付けやチャンキングの性能を評価した研究は少ない．文献 [2, 6, 9] は学習者の英語に対する品詞タグ付けの性能を報告しており，ネイティブの文に対する品詞タグ付けと学習者の英語に対する品詞タグ付けの間に性能差があることを示した．しかしながら，これらの先行研究では品詞タグ付けの誤りの原因について詳細な議論はされていない．誤りの原因に対する詳細な調査は，品詞タグ付けやチャンキングの性能の改善につながると期待できる．更に品詞タグ付けやチャンキングの性能が改善することで誤り訂正のような関連タスクにおける性能改善も期待ができる．また学習者の英語に対するチャンキングの性能を報告した研究は，我々が知る限りこれまでない．[*1]

品詞タグ付けやチャンキングが失敗する原因として未知語の問題がある．学習者の英文中では，未知語となる代表的なものにスペル誤りがある．スペル誤りは学習者の英語において数多く出現する誤りのひとつである．そこで本稿では，スペル誤りに注目して品詞タグ付け・チャンキングの性能を調査する．具体的にはスペル誤りの性能調査を以下の 3 つに分けて調査する．

- スペル誤りがどの程度性能低下に影響しているか
- どのようなスペル誤りが性能に影響を与えるか
- スペルチェッカーによるスペル訂正の効果

詳細については次節で述べる．

2 スペル誤りに着目した性能調査

本節で，本稿で行なうスペル誤りに着目した品詞タグ付け・チャンキングの性能調査について述べる．

■スペル誤りがどの程度性能低下に影響しているか 学習者の英語はスペル誤りが 3.4% 含まれていると言われている [3]．未知語の単語に対して品詞タグ付けやチャンキングが全て失敗すると仮定すると，スペル誤りによって 3.4% は精度が低下していることになる．実際は前後の単語からスペル誤りの品詞やチャンクを推測することができるため，3.4% 精度が低下することはないが，どの程度当てられているかは明らかになっていない．一方，前後の単語からスペル誤りの品詞やチャンクを推定できることを考えると，逆にスペル誤りが影響して前後の単語の品詞やチャンクの推定が失敗する可能性がある

[*1] 構文解析がチャンキングを兼ねるように見えるが，チャンキングはミニマルなフレーズを対象とする．多くの誤り訂正の研究ではチャンキングまでを用いている．

と考えることができる．本稿では，学習者の文に対して品詞タグ付け・チャンキングしたものと，スペル誤りを人手で修正したものに品詞タグ付け・チャンキングしたものを比較することで，スペル誤りによってどの程度性能が低下しているかを明らかにする．また，スペル誤りの単語やその前後の単語に対する品詞タグ付け・チャンキングの正解数を見ることで，スペル誤りが前後の単語の品詞タグ付け・チャンキングに与える影響を確かめる．

■どのようなスペル誤りが性能に影響を与えるか　学習者の英語にはさまざまなタイプのスペル誤りが出現する [9]．もっとも一般的で出現数が多いスペル誤りは誤字 (e.g. *studing/studying) である．学習者の英語中には他にもさまざまなタイプのスペル誤りがあり，例えば，同音異義 (e.g. *see/sea)，分割 (e.g. *home town/hometown)，結合 (e.g. *airconditioner/air conditioner)，屈折・派生 (e.g. *smell/smelly) のようなものがある．[*2]誤字や結合のように未知語となるものもあれば，同音異義や分割のように既知の単語になるものもある．未知語のスペル誤りであれば前後の単語から品詞やチャンクを推測できる可能性がある一方で，既知語で品詞が変わる（上の同音異義）のような場合は解析に失敗すると考えられる．本稿では実際にスペル誤りの事例を比較することで調査する．

スペル誤りの中には，品詞を決定するために有効な情報を保持しているものがある．例えば，上の誤字の誤りであれば単語の接尾辞 “ing” があるため動詞の進行形/動名詞ということが推測できる．このような品詞を決定するために有効である接辞情報がどの程度スペル誤りに有効であるかも調査する．この調査のために，接辞を考慮しない品詞タグ付けシステムと接辞を考慮した品詞タグ付けシステムを比較する．

■スペルチェッカーの効果　従来の誤り訂正／検出の研究では，スペル誤りによる未知語の問題に対処するために前処理としてスペルチェッカーを用いてスペル誤りを訂正している．しかしながら，上で述べたようにスペル誤り自体やその前後の単語に対する品詞タグ付けやチャンキングの性能はわかっていない．そのため，これまで前処理で使われてきたスペルチェッカーが品詞タグ付けやチャンキングにどの程度有効であるかは不明である．スペルミスの中で，未知語となるものはスペルチェッカーを用いることで正しいスペルに訂正できる可能性があり，品詞タグ付け・チャンキングも成功する可能性がある．本稿ではスペルチェッカーを用いない場合と用いた場合の品詞タグ付け・チャンキングの性能を比較することで，スペルチェッカーの効果を明らかにする．

表 1

KJ コーパスに対するスペルチェッカーの性能

#TP	#FP	#FN	Precision	Recall	F-socre
409	197	120	67.49	77.32	72.07

3　実験設定

品詞タグ付けとチャンキングのモデルを学習するために，英語教材の読解問題に独自に品詞とチャンク情報をアノテーションした．このコーパスは 16,375 文，213,017 のトークンから成る．

品詞タグ付けとチャンキングの性能を評価するために Konan-JIEM コーパス（KJ コーパス） [6] を使用した．KJ コーパスは日本人大学生によって書かれたエッセイから構成され，3,260 文，30,1517 トークンから成る．本稿で対象とするスペル誤りの数は 654 個であった．未知語となるスペル誤りが 487 個，既知語となるスペル誤りが 167 個であった．

品詞タグ付けとチャンキングは系列ラベル付け問題として定式化し，条件付き確率場 (CRF) を用いた．CRF は品詞タグ付けやチャンキング問題を解く際に用いられる一般的な方法のひとつである．品詞タグ付けとチャンキングのモデルを構築するために，CRF のツールとして CRF++[*3]を使用した．CRF++ のパラメータはデフォルトのまま用いた．品詞タグ付けのための素性として広く使われているの素性 [8] を使用した．素性は表層，記号・数字・大文字の出現，単語の原型，接辞から構成される．チャンキングに対する素性も一般的に使用されている素性を使用した．チャンキングの素性は表層と単語の原型と品詞から構築される．

スペルチェッカーは独自に開発した．母語によるスペル誤りの影響を捉えるため，Noisy Channel Model に基づいたスペルチェッカーを構築した．表 1 は KJ コーパスに対するスペルチェッカーの性能を示す．この結果は Sakaguchi ら [9] のスペルチェッカーと比べても高い値である．

評価指標として Accuracy (正解数 / コーパス中の全トークン数) を用いた．上記に加え，スペル誤りが前後の単語に与える影響を調べるため，スペル誤りの単語とその前後の単語に対する品詞タグ付け・チャンキングの正解数を示す．

[*2] 本稿では，分割と結合のスペル誤りは扱わない．

[*3] https://taku910.github.io/crfpp/

4 学習者の英語に対する品詞タグ付け実験と考察

2 節で述べたことを確かめるために実験を行なう．以下の 5 つの手法を比較する:

1. 表層，原型，数字・記号・大文字の出現の素性で学習した品詞タグ付けシステム（**Baseline**）
2. Baseline に接辞の素性を追加したシステム (**Baseline+Affix**)
3. スペルチェッカーでスペル誤りを訂正した文に対して Baseline で品詞タグ付け (**Baseline+SpellChecker**)
4. スペルチェッカーでスペル誤りを訂正した文に対して Baseline+Affix で品詞タグ付け (**Baseline+Affix+SpellChecker**)
5. スペル誤りを全て正解に訂正した文に対して Baseline+Affix で品詞タグ付け (**Baseline+Affix+SpellGold**)

表 2 に品詞タグ付けの実験結果を示す．“Baseline+Affix” と “Baseline+Affix+SpellGold” を比較すると，“Baseline+Affix+SpellGold” の方が 0.23% 高い．このことから，スペル誤りは品詞タグ付けの性能を 0.23% 下げていると言える．“Baseline” と “Baseline+Affix” を比較すると，“Baseline+Affix” の方が 1.3% 精度が高い．また, “Baseline+Affix” と “Baseline+SpellChecker” を比較しても “Baseline+Affix” の精度が高い．これらの結果から学習者の文に品詞タグ付けする場合に接辞の情報が重要であることがわかる．“Baseline+Affix” と “Baseline+Affix+SpellChecker” では，0.06% の差しかなく，スペルチェッカーを使わなくても接辞情報だけで十分であると言える．

表 3 にスペル誤りとその前後の単語に対する品詞タグ付けの正解数を示す．“Baseline” と品詞タグ付けでもっとも精度が高かった “Baseline+Affix+SpellGold” を比較すると，スペル誤りの単語自体の正解数は増えている．一方，その前後の単語の正解数はほとんど差がなく，スペル誤りは前後の単語の品詞を当てるのに影響を与えないことがわかる．

■どのようなスペル誤りが影響を与えるかの分析　まず接辞情報で捉えられるスペル誤りを見るため “Baseline” と “Baseline+Affix” の比較から行なう．スペル誤りの単語の正解数を比較すると “Baseline” が 344 個，“Baseline+Affix” は 465 個である．接辞情報を使うことで約 120 個のスペル誤りの単語に対して正しい品詞を付与できた．“Baseline” で品詞付与に失敗し，“Baseline+Affix” で品詞付与できた例を挙げる．

表 2　品詞タグ付けの実験結果

手法	Accuracy
Baseline	93.97
Baseline+Affix	95.31
Baseline+SpellChecker	94.21
Baseline+Affix+SpellChecker	95.37
Baseline+Affix+SpellGold	95.54

表 3　スペル誤りの単語とその前後の単語に対する品詞タグ付けの正解数．手法 *Best* は “Baseline+Affix+SpellGold” の結果を示す．

手法	スペル誤り単語の正解数	前の単語の正解数	後の単語の正解数
Baseline	344	540	590
Baseline+Affix	465	542	598
Best	528	547	596

(1) a. Winter is decolated/動詞過去 ...
　　b. Accoding/動名詞現在分詞 to ...
　　c. ... big Snoopy dools/名詞複数 ...

接辞情報を使うことで，decolated（正解は decorated）は *ed* を元に動詞過去，Accoding（正解は According）は *ing* を元に動名詞現在分詞，dools（正解は dolls）は *s* を元にそれぞれ正しい品詞を当てることができたと考える．

次に “Baseline+Affix” と “Baseline+Affix+SpellGold” の出力を分析する．“Baseline+Affix” では品詞タグ付けに失敗したが，“Baseline+Affix+SpellGold” で品詞タグ付けに成功したものは 105 個であった．このうち，未知語のスペル誤りが 78 個であり，既知語のスペル誤りが 27 個であり，どちらの誤りも約 84% は正しく品詞が付けられている．さらにこの 105 個を失敗の原因から 5 つに分類した．もっとも多かったものはスペルミスによって未知語となり失敗しているものが 54 個 (51.4%) あった（例：evey）．品詞タグ付けの素性として用いてた接辞によって品詞が決まっているものが 21 個 (20.0%)（例：whiting/動詞現在分詞・動名詞），スペルミスによって違う単語になっているものが 17 個 (16.1%)（例：thought → though），品詞タグ付けの大文字が含まれるかどうかの素性による失敗が 10 個（例：Exsample/固有名詞），日本語の単語をローマ字にしているために起こっている失敗が 3 個あった．

■スペルチェッカーの効果の分析　接辞情報では正しい品詞を当てられなかったが，スペルチェッカーを使うこ

とで正しい品詞が付与できたものを分析する．スペルチェッカーを使うことで正しい品詞を付与できたスペル誤りの単語は 74 個であった．スペルチェッカーを使うことで正しい品詞が当てられなくなったスペル誤りの単語も 49 個あった．スペルチェッカーを使うことで正しい品詞を付与できるようになったものは，

(2) a. pepole/名詞 → peaple/名詞複数
b. tow/名詞 → two/数詞

のようにスペルチェッカーが正しい単語に訂正できたものである．上で挙げた “スペルミスによって未知語のため失敗しているもの” に対する品詞タグ付けに成功している．スペルチェッカーを使うことで正しい品詞が当てられなくなったものを調べると，*tero*（正解は terrorist）を *to* に訂正したり，*tittle*（正解は title）を *little* に訂正したものがあり，これはスペルチェッカーの間違った訂正が原因である．

5 学習者の英語に対するチャンキングの実験

品詞タグ付けと同様に，チャンキングでもスペル誤りの単語の影響を調べるために実験を行なう．チャンキングの実験では，3 節で説明した素性を使用したものを “**Baseline**” とし，スペルチェッカーを使ったもの (**Baseline+SpellChecker**)，正しいスペルを使ったもの (**Baseline+SpellGold**) で比較する．チャンキングのモデル学習に素性として使用する品詞は，自動タグ付けしたものを用いた．

表 4 にチャンキングの実験結果を示す．“Baseline” と “Baseline+SpellChecker” を比較すると，0.03% としか差がなく，スペルチェッカーの効果はない．一方，“Baseline” と “Baseline+SpellGold” では，0.2% の差があり，品詞と同様で理想的なスペルチェッカーの場合は効果がある．しかし，チャンキングの素性として品詞タグ付けの結果を使っているため，チャンキングの場合は品詞タグ付けの性能がそのまま影響していると考えることができる．

表 5 はスペル誤りとその前後の単語に対するチャンキングの正解数である．品詞の場合と同様に，スペル誤りの単語自体の正解数は正しいスペルを使った場合に増えているが，前後に関しては大きな差がない．

表 4 チャンキングに対する実験結果

手法	精度
Baseline	94.38
Baseline+SpellChecker	94.41
Baseline+SpellGold	94.58

表 5 スペル誤りの単語とその前後の単語のチャンキングの正解数．手法 *Best* は “Baseline+SellGold” の結果を示す．

手法	スペル誤り単語の正解数	前の単語の正解数	後の単語の正解数
Baseline	532	504	565
Best	566	519	570

6 おわりに

本稿では，学習者の英語に対する品詞タグ付けとチャンキングに対する詳細な性能調査を行なった．品詞タグ付けにおいて未知語が性能に影響することが知られており，学習者の文で未知語として出現するスペル誤りに注目して調査した．学習者の文に対するチャンキングの性能評価は世界初である．スペル誤りがどの程度性能低下に影響しているか，どのようなスペル誤りが性能に影響を与えるか，品詞タグ付け・チャンキングに対するスペルチェッカーの効果について調査した．

参考文献

[1] J. Aarts and S. Granger, “Tag sequences in learner corpora: a key to interlanguage grammar and discourse,” in Learner English on Computer, ed. S. Granger, pp.132–141, Addison Wesley Longman: London and New York, 1998.

[2] Y. Berzak, J. Kenney, C. Spadine, J.X. Wang, L. Lam, K.S. Mori, S. Garza, and B. Katz, “Universal Dependencies for Learner English,” Proceedings of ACL, pp.737–746, 2016.

[3] M. Flor, Y. Futagi, M. Lopez, and M. Mulholland, “Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus,” the Second Learner Corpus Research Conference, 2013.

[4] N.R. Han, M. Chodorow, and C. Leacock, “Detecting Errors in English Article Usage by Non-Native Speakers,” Natural Language Engineering, vol.12, no.2, pp.115–129, 2006.

[5] R. Nagata and E. Whittaker, “Reconstructing an Indo-European Family Tree from Non-native English Texts,” Proceedings of ACL, pp.1137–1147, 2013.

[6] R. Nagata, E. Whittaker, and V. Sheinman, “Creating a Manually Error-tagged and shallow-parsed corpus,” Proceedings of ACL-HLT, pp.1210–1219, 2011.

[7] H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant, “The CoNLL-2014 Shared Task on Grammatical Error Correction,” Proceedings of CoNLL Shared Task, pp.1–14, 2014.

[8] A. Ratnaparkhi, “A Maximum Entropy Model for Part-Of-Speech Tagging,” Proceedings of EMNLP, pp.133–142, 1996.

[9] K. Sakaguchi, T. Mizumoto, M. Komachi, and Y. Matsumoto, “Joint English Spelling Error Correction and POS Tagging for Language Learners Writing,” Proceedings of COLING, pp.2357–2374, 2012.

[10] J.Z. Sukkarieh and J. Blackmore, “c-rater: Automatic Content Scoring for Short Constructed Responses,” Proceedings of FLAIRS, pp.290–295, 2009.