学習者英語のための綴り誤り訂正手法と綴り誤り分析への応用

永田　亮† 高村　大也†† GrahamNeubig†††
† 甲南大学知能情報学部 †† 東京工業大学 ††† Carnegie Mellon University

1. はじめに

綴り誤りは，学習者の英文の大きな特徴の一つである．学習者の英文では，綴り誤りは頻出するだけでなく，母語の影響を受けることが知られている．そのような綴り誤りは，学習者の英文を対象とする言語処理システム（例えば，エッセイの自動添削）の性能を低下させる要因となる [13]．また，綴り誤りは，言語学的な分析にも影響を与えることがある．例えば，綴り誤りにより，単語の種類数を精度良く推定することさえ困難となる [7]．

綴り誤り訂正は，これらの問題を解決する有効な手段である．従来手法では，編集距離 [9] に基づいて綴り誤りを訂正することが多い．より高度な手法として，noisy channel model の利用 [1] と周辺文脈の利用 [2], [3] が提案されている．前者は，学習者の誤り傾向を直接モデル化できるというメリットがある一方，綴り誤り情報が付与された訓練データが必要となるというデメリットもある．学習者のレベルや母語に適応したモデルを得るためには，対象とする学習者のグループごとに専用の訓練データが必要となり更にコストが高くなる．後者は，文脈に応じた綴り誤り訂正が可能となるが，学習者の英文に頻出するノイズ（文法誤りや綴り誤りなど）から影響を受けるという問題がある．

以上の問題以外に，学習者の英文を対象にした綴り誤り訂正の研究には未解決の問題が存在する．高性能な綴り誤り訂正を実現するために，母語へのドメイン適応が必要かどうか研究者の間で合意が得られていない [4]．上述の通り，綴り誤りは母語の影響を受けるため，多くの研究者が母語へのドメイン適応を提案している．一方で，母語に依存しない一般的な言語モデルに基づいた手法の性能は，ドメイン適応した手法と同程度であるという報告 [3] もある．

以上の問題を解決するため，本稿では，新たな綴り誤り訂正手法を提案する．提案手法は，綴り誤りの情報が付与されていない学習者コーパスから動的に訂正モデルを構築する．そのため，様々な母語に応じた訂正モデルを自動的に構築することが可能となる．また，綴り誤り周辺のノイズからの影響を低減する工夫も行う．

更に，訂正性能の向上に加えて，本稿では，提案手法を利用した綴り誤りの分析についても報告する．学習者コーパスから自動構築した訂正モデルを利用して，綴り誤りにおける，母語間の差異を分析する．この分析および性能評価実験を通じて，母語に応じたドメイン適応が必要かどうかという未解決の問題も議論する．

2. 提案手法

提案手法で対象とするのは，非単語綴り誤り（注 1）である．提案手法では，非単語綴り誤り訂正に編集距離を利用する．具体的には，辞書に登録されていない綴りに対して，編集距離が最小となる辞書中の単語を訂正候補とする．ただし，この単純な方法では，明らかに不十分である．例えば，綴り誤り “mather” に対して，編集距離 1 となる単語は “bather”, “father”, “gather”, “mother” など多数あり，この中から適切な綴りを選び出さなければならない．

そこで，提案手法では単語の分散表現を通じて意味的な情報を利用して綴り誤り訂正を行う．分散表現は，学習者コーパス（綴り誤り訂正対象の文章を含めてもよい）から得る．その結果，綴り誤りを含めて，学習者コーパス中の各単語が内積空間上で表現される．綴り誤りは，対応する正しい綴りの単語と同じような文脈で使用されると予想される．従って，両者は内積空間上で近い位置に現れるとも予想される．図 1 は，正に，この状況を表している；綴り誤り “mather” は，正しい綴り “mother” の最近傍に，その他の候補は，意味の近さに応じてプロットされている．このような空間が得られれば，編集距離と組み合わせることで最適な訂正候補を選び出せるというのが基本的な考え方である．

提案手法では，単語の分散表現を得るために，continuous bag-of-words model [10] を利用する．入力となる学習者コーパスに対して，文分割，トークン分割，小文字化の前処理を行い，その結果から分散表現を得る．得られた分散表現を利用して，綴り誤り候補と訂正候補間の意味的類似度をベクトル間の余弦類似度として計算する．最終的に，従来手法 [3] にならい，(1) 余弦類似度，(2) 編集距離に基づいた類似度，Double Metaphone [12] に基づいた類似度，(3) 単語の相対頻度，(4) 訂正対象文書のほかの場所に出現するかどうかを 2 値で表したものの重み付き和をとる（注 2）．この値が最大となる綴りを訂正結果とする．

以上の処理により，綴り誤り訂正結果が得られるが，この結果を利用して更に訂正性能を向上させることを考える．す

(注 1)：非単語綴り誤りとは，英語の綴りとしては存在しない綴りのことを指す．一方で，綴り自体は存在するが与えられた文脈では誤りとなる文脈依存誤り（例：“Their is a house.”）もある．
(注 2)：文献 [3] 同様に，重みは開発データを用いて経験的に求める．開発データがない場合は，(1) ～ (3) を掛け合わせたものを利用しても良い．

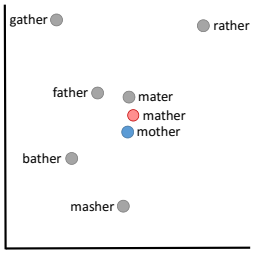

図 1： 内積空間上の綴り誤りと訂正候補.

なわち，誤り訂正結果を訓練データとして別の綴り誤り訂正モデルを得ることを考える．ここで，重要なのは，新たな綴り誤り訂正モデルに，分散表現に基づいた手法とは異なった性質を持たせなければならないということである（さもなければ，同じ訂正結果をコピーするだけになるであろう）．

この条件を満たすために，本稿では，noisy channel model を応用する（以降，このモデルのことを訂正モデルと呼ぶ）．すぐ後に見るように，提案する訂正モデルでは，文字に基づいて誤り訂正を行うため，分散表現に基づいた手法とは異なった訂正傾向をもつ．

訂正モデルを定式化するために，次の記号を定義する．綴り誤りとその訂正候補をそれぞれ w_e と w_c で表す．また，w_e，w_c 中の i 番目の文字（列）を，それぞれ $w_{e,i}$ と $w_{c,i}$ で表す．このとき，訂正モデルを

$$\hat{w}_c = \arg\max_{w_c} \Pr(w_c) \prod_i \Pr(w_{e,i}|w_{c,i}) \qquad (1)$$

で定義する．式 (1) の $\Pr(w_c)$ は，訂正候補の出現確率である．この確率は，母語話者コーパス及び綴り誤り訂正対象コーパス（すなわち学習者の書いた英文）で，それぞれ推定し，線形補完することで求める．学習者コーパスから推定された確率を用いるのは，正しい綴りは，綴り誤りが出現した同じ文章に出現しやすいという傾向をモデルに取り入れるためである．式 (1) の $\Pr(w_{e,i}|w_{c,i})$ は，ある文字（列）がどの文字（列）に誤りやすいかを表す条件付き確率である．従って，これを綴り誤りモデルと呼ぶことにする．綴り誤りモデルは，文字レベルで学習者の誤り傾向をモデル化する．その結果，分散表現に基づいた手法とは異なった訂正となることが期待できる．本稿では，次の四種類の条件付き確率を線形補完したものを綴り誤りモデルとして提案する：(a) 文字列対文字列の誤り確率；(b) 出現位置（語頭，語中，語末）を考慮した文字列対文字列の誤り確率；(c) 文字クラス（母音，子音，数字，その他）対文字列の誤り確率：(d) 出現位置を考慮した文字クラス対文字列の誤り確率.

以上の手順で得られた訂正モデルを用いて綴り誤り訂正を行う．訂正モデルでは，訂正に対する確信度（確率）が得られるため，確信度がある閾値以上の場合のみ訂正を行う．

3. 評 価 実 験

実験対象として，Konan-JIEM （KJ） corpus [11] と独自に収集した日本人英語学習者コーパス（以下，JSC と略記）を用いた．書き手の母語は，いずれの場合も日本語である．本評価実験用に，綴り誤り情報を人手で付与した．その中で，アルファベットのみからなる非単語綴り誤りを訂正対象とした．JSC の 20%を開発データとして各種パラメタの推定に利用し，残りを評価用とした．各コーパスのサイズは，それぞれ，JSC 開発 13,888 トークン（対象綴り誤り 204），JSC 評価 53,067 トークン（同 942），KJ 30,586 トークン（同 438）である．

これらのコーパスを対象にして，提案手法の性能を再現率，適合率，F 値で評価した．綴り辞書として，独自に開発した英単語および頻度が記載されている辞書（95,696 語）を用いた．分散表現の学習には，word2vec(注 3) を用いた．訓練データとして，評価対象に加え CEEJUS(注 4)，ETS Corpus of Non-Native Written English（ETS）(注 5)，NICE(注 6)，ICLE [6]，ICNALE [8]，Lang-8 (注 7) それぞれの日本人英語学習者サブコーパスを利用した．また，訂正モデルは，それぞれの評価対象から構築した．

比較のため，編集距離に基づいた手法と言語モデルに基づいた手法の性能も評価した．編集距離に基づいた手法で最小編集距離が同じになる場合は，単語の頻度が高い候補を選択した．言語モデルに基づいた手法は，文献 [3] を参考にして，提案手法の余弦類似度を言語モデルの確率 (注 8) に置きかえたものとした．いずれの手法とも，上述の綴り辞書を用いた．

表 1 に評価結果を示す．表 1 から，提案手法（分散表現に基づいた手法，訂正モデルに基づいた手法）は，従来手法に比べ，再現率も適合率も改善していることがわかる．訂正結果を分析したところ，分散表現を用いて意味的な情報を利用して綴り誤り訂正を行うことの効果が確認できた．例えば，綴り誤り “sistem” に対して，編集距離 1 の訂正候補は，少なくとも “system” と “sister” の二種類があるが，日本人英語学習者コーパス（すなわち，分散表現の訓練データ）では，“system” の意味で使用されるほうが多数であった．その結果，余弦類似度は，“system : 0.25” と “sister : -0.05” となり

（注 3）: https://github.com/dav/word2vec. ハイパパラメタは次の通り：次元：200；文脈サイズ：9；単語の出現頻度閾値：2.
（注 4）: http://language.sakura.ne.jp/s/doc/projects/CEEAUS.pdf
（注 5）: https://catalog.ldc.upenn.edu/LDC2014T06
（注 6）: http://sgr.gsid.nagoya-u.ac.jp/wordpress/?page_id=695
（注 7）: http://cl.naist.jp/nldata/lang-8/
（注 8）: 言語モデルの訓練には Faster RNNLM（https://github.com/yandex/faster-rnnlm）を用いた（オプション “-hidden 128 -hidden-type gru -nce 20 -alpha 0.01” を用いた）．訓練データとして，AQUAINT（Graff, David. The AQUAINT Corpus of English News Text LDC2002T31, https://catalog.ldc.upenn.edu/LDC2002T31）中の New York Times の記事を用いた（同コーパスの 5,600 万行を訓練データ，残り 147,329 行を開発データとした）．その際，数字は記号 NUM に置き換え，頻度 10 以下の単語は未知語とした．また，文分割処理を行い，文末を表すピリオドは削除した．更に，確率の計算は一文全体に対して行った．

表 1： 綴り誤り訂正性能.

Target: JSC test			
Method	再現率	適合率	F 値
分散表現に基づいた手法	**0.688**	**0.641**	**0.664**
綴り誤り訂正モデル	0.680	0.638	0.658
言語モデルに基づいた手法	0.676	0.630	0.652
編集距離に基づいた手法	0.593	0.553	0.572
Target: KJ			
分散表現に基づいた手法	**0.614**	**0.486**	**0.543**
綴り誤り訂正モデル	0.569	0.463	0.521
言語モデルに基づいた手法	0.562	0.445	0.497
編集距離に基づいた手法	0.515	0.408	0.455

正しい綴り “system” を選択できていた．一方で，言語モデルに基づいた手法では，予想通り，綴り誤り周辺のノイズから影響を受ける結果となった．例えば，“sistem” は，“· · · I want to be a *sistem* enginir · · ·” という文脈で使用されており，直後のノイズ（綴り誤り）の影響を受け，“sister” を選択していた．同様に，文法誤り（例 : “But, *tuch* the computer is · · ·”）の影響もうかがえた．更に，文脈自体に曖昧性がある場合にも同手法では訂正に失敗する傾向にあった（例 : “He is really *qute*.” だけでは，“cute” か “quiet” か曖昧である）．いずれの場合も，提案手法では正しく訂正できた．ただし，分散表現に基づいた手法の性能を訂正モデルで改善することには成功しておらず，この点については今後の課題となる．

更なる分析により，学習者の英文中の綴り誤りには一つの強い傾向があることが判明した．表層が同じである非単語綴り誤りは，文脈に関係なく，ほぼ常に同じ綴りに訂正されるという傾向である（本評価実験では，99.05%がこのケースに該当）．このことは，文脈情報がなくとも正しい綴りを推測できることを示唆する．そのため，文脈情報を使用しない提案手法で，高い訂正性能を達成できたと分析できる．更に，one-sense-per-discourse [5] を考慮すると，綴り誤り訂正対象の文章が一つの内容について書かれている場合，この傾向は強まると予想される．幸い，語学学習支援では，試験問題やクラスでのライティング課題など，この条件が満たされる場面が多くある．

綴り誤り訂正の際に，編集距離いくつまでの単語を訂正候補とするかは重要なパラメタである．開発データで最適な値を求めたところ，誤り訂正モデル以外の手法は，いずれも編集距離 1 のときに F 値が最大となった（綴り誤り訂正モデルは編集距離 3 のときに最大となった）．また，編集距離を増加させると，F 値は低下する傾向が見られた．従って，正しい綴りから編集距離 2 以上の綴り誤りの訂正については，改善の余地が残されているといえる．

母語へのドメイン適応が必要かどうかを調査するために，訓練データを次のように変化させ分散表現に基づく手法の性能を評価した： (a) 全て日本人英語学習者コーパス（上と同じ訓練データ）； (b) (a) に日本人英語学習者以外のコーパス（ETS，ICLE，ICNALE）を加えたもの ; (c) 日本人英語学習者以外の学習者コーパス（(b) から（a）を除いたもの）; (d) (a) に母語話者コーパス（New York Times News の 2000 年の記事）を加えたもの．ただし，全ての条件とも，JSC と KJ を訓練データに含めた．

表 2： 訓練データの違いによる訂正性能の変化.

訓練データ（トークン数）	JSC	KJ
Japanese only (14 million)	**0.664**	**0.543**
All (35 million)	0.651	0.514
All except Japanese (17 million)	0.654	0.480
Japanese with Native (50 million)	0.649	0.517

表 2 に実験結果を示す．表 2 より， 書き手と母語が同じである訓練データが性能向上に寄与していることがわかる．それ以外の学習者コーパスを追加しても，効果がない，もしくは逆に若干性能が低下している．また，予想に反して，母語話者の英文データの追加は，性能を低下させる結果となった．これらの結果は，母語へのドメイン適応の効果を示唆する．ただし，英文の内容やトピックに適応している可能性もある．

4. 綴り誤り分析への応用

提案手法の綴り誤りモデルを用いて，綴り誤りを定量的に分析することが可能である．ここでは，ケーススタディとして，日本人英語学習者コーパス（JSC と KJ）とフランス人英語学習者コーパス（ETS および ICLE の当該サブコーパス）の比較を行う．両コーパスから，提案手法を用いて，綴り誤りモデルを得た．分析を効率的に行うため，二つのモデルにおける条件付き確率の比を考える．この比が大きいということは，日本人英語学習者コーパスに比べて，フランス人英語学習者コーパスでは，ある文字が別の文字に誤る可能性が高いということである（また，逆の議論も成り立つ）．

表 3 に，条件付き確率の比が高い順に上位 12 件を示す．表中の $c \to e$ は，ある文字 c が e に誤るという意味を表す（すなわち，$\Pr(e|c)$ に対応する）．なお，可読性を高めるため表 3 は，(a) 文字列対文字列の誤り確率と（b）位置を考慮した文字列対文字列の誤り確率のみを対象にしている．表中の記号ˆ, +, $は，それぞれ，語頭，語中，語末を表す（記号がないものは，位置を考慮しないモデルに対応する）．

フランス人英語学習者では，明らかに余分な文字の追加（例 : “$\phi \to$ p” ; ϕ は空文字を表す）が多い．当該コーパスを分析してみると，この誤りは同じ子音の連続で頻繁に起こることが判明した．代表例は，対応するフランス語の単語が誤って英語で使用されるケースである（例 : “développement”, “consumme” など）．これ以外にも，現在／過去分詞の語尾変化における誤りも頻出することが確認された（例 : comming, mentionned）．更に，フランス語の文法の影響もみられた．

例えば，“ϕ\$ → s” に該当する綴り誤りを調査したところ，“differents” など形容詞と名詞の一致に起因する誤り（例：differents topics）が確認できた．この種の誤りは，文法誤りとして捉えることもでき，フランス人英語学習者に対しては，綴り誤りとしてではなく，文法誤りとしてその理由と共に訂正したほうが効果的であると予想される．一方で，日本語には形容詞と名詞の一致は基本的にないため，そのようなフィードバックは必要ない．実際，著者の一人が目視で確認したところ，日本人英語学習者コーパスでは，この種の誤りは皆無であった．従って，学習者へのフィードバックという観点からも母語へのドメイン適応は重要であるといえる．

日本人英語の綴り誤りは，より文字指向である．典型例は “+l → r” であり，表 3 でも突出している．また，母音の区別にも特徴が見られる．例えば，“stady（c.f.，study）” のように,（日本人）英語で,「あ」と発音されるが “a” 以外の文字で表記される単語は綴り誤りが頻出する．また，逆の影響（“+a → u”）もみられる（例：buttle ; c.f，battle）．

以上のように，表 3 の結果は両学習者の特徴を示すが，誤訂正の影響も見られる．特に，日本語の固有名詞における誤訂正の影響を受けている（例：地名 “Akashi” を “Takashi” に訂正）．固有名詞における訂正ミスを減らすためには，母語に特化した綴り辞書を使用することが有効である．

2. で述べたように，分析に必要となる綴り誤りモデルは，学習者コーパスから動的に構築される．分析結果は一部誤訂正の影響を受けるものの，上で示したように，学習者の英文における綴り誤りの特徴分析に十分に利用可能である．特に，自動構築した綴り誤りモデルを用いることは，任意の母語に対する分析を容易にするという大きな利点がある．このような分析を，綴り誤りモデルなしで様々な母語に対して行うことは困難であろう．

5. おわりに

本稿では，新たな綴り誤り訂正手法を提案した．提案手法の利点として，(1) 書き手の母語に適応した綴り誤りモデルを自動的に構築できる：(2) 綴り誤り周辺のノイズに強いという二点を挙げることができる．更に，本稿では，提案手法を用いて，学習者の英文における綴り誤りの特徴を分析した．この分析と評価実験から，次の知見を得た：(I) 綴り誤りと対応する正しい綴りの単語は，内積空間上で近い位置に出現する; (II) 一つの綴り誤りは，ほぼ常に同じ綴りに訂正される ; (III) 訂正性能の観点からも，学習者へのフィードバックという観点からも，母語へのドメイン適応は重要である．また，本研究の成果として，綴り誤り情報付き KJ コーパスを公開している．

謝　辞

本研究の一部は科研費若手研究（B）（19700637）の助成により実施した．

表 3：　日本人／フランス人英語学習者コーパスに特徴的な綴り誤り．

French English		Japanese English	
Ratio	$c \to e$	Ratio	$c \to e$
20.0	$+\phi \to$ p	273.0	+l → r
17.0	$\phi \to$ p	268.0	l → r
12.0	i\$ → w	104.0	+t → ϕ
7.0	d → ϕ	84.0	+u → a
6.0	ϕ\$ → t	53.0	u → a
5.0	+t→ e	43.0	+z → g
5.0	^i → u	36.0	z → g
5.0	$\phi \to$ n	30.0	^c → k
5.0	$\phi \to$ b	26.0	c → k
4.0	q\$ → k	25.0	+a → u
4.0	d\$ → s	23.0	o → l
4.0	ϕ\$ →s	23.0	e\$ → i

参考文献

[1] E. Brill and R.C. Moore, “An improved error model for noisy channel spelling correction,” Proc. of 38th ACL, pp.286–293, 2000.

[2] M. Flor, “Four types of context for automatic spelling correction,” TAL, vol.53, no.3, pp.61–99, 2014.

[3] M. Flor and Y. Futagi, “On using context for automatic correction of non-word misspellings in student essays,” Proc. of 7th BEA Workshop, pp.105–115, 2012.

[4] M. Flor and Y. Futagi, “Patterns of misspellings in L2 and L1 English: a view from the ETS spelling corpus,” Bergen Language and Linguistic Studies, vol.6, pp.107–132, 2015.

[5] W. Gale, K. Church, and D. Yarowsky, “One sense per discourse,” Proc. of 4th DARPA Speech and Natural Language Workshop, pp.233–237, 1992.

[6] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot, International Corpus of Learner English v2, Presses universitaires de Louvain, Louvain, 2009.

[7] S. Granger and M. Wynne, “Optimising measures of lexical variation in EFL learner corpora,” in Corpora Galore, pp.249–257, Rodopi, 1999.

[8] S. Ishikawa, A new horizon in learner corpus studies: The aim of the ICNALE project, pp.3–11, University of Strathclyde Publishing, Glasgow, 2011.

[9] V.I. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals,” Soviet Physics-Doklady, vol.10, no.8, pp.707–710, 1966.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” Proc. of 2013 Workshop of International Conference on Learning Representations, 2013.

[11] R. Nagata, E. Whittaker, and V. Sheinman, “Creating a manually error-tagged and shallow-parsed learner corpus,” Proc. of 49th ACL, pp.1210–1219, 2011.

[12] L. Philips, “The double metaphone search algorithm,” C/C++ Users J., vol.18, no.6, pp.38–43, 2000.

[13] J.Z. Sukkarieh and J. Blackmore, “c-rater: Automatic content scoring for short constructed responses,” Proc. of 2nd International FLAIRS Conference, pp.290–295, 2009.