Vol.2014-NL-217 No.15
2014/7/4

複単語表現を考慮した英語文法誤り訂正

水本 智也[1,a)]　松本 裕治[1,b)]

概要：複単語表現は重要な言語学的な情報として認識されており，複単語表現の獲得に関する研究が行なわれている．その一方で，自然言語処理の応用タスクにおいて，複単語表現はあまり用いられてはいない．英語の第二言語学習者もネイティブと同様に，作文中で複単語表現を使用する．しかしながら，文法誤り訂正タスクにおいて複単語表現は考慮されてこなかった．本稿では，複単語表現を使った英語文法誤り訂正の手法を提案する．提案する手法は複単語表現を直接的に英語文法誤り訂正に適用しているが，実験の結果，複単語表現が英語文法誤り訂正に有効であることがわかった．

1. はじめに

一般の人が気軽に使える Web 上の言語学習支援サービスが増えている．例えば，学習している言語の作文を SNS 上で相互に添削しあう Lang-8 [*1]や英文チェッカー GINGER [*2]などが公開されている．また，第 2 言語学習支援に関する研究も盛んに行なわれており，特に英語学習者の書いた作文の文法誤り訂正が盛んである．過去には，英語文法誤り訂正の性能を競う HOO（2011 年，2012 年）[4], [5], CoNLL Shared Task（2013 年）[9] も開催された．2014 年も英語文法誤り訂正を対象として CoNLL Shared Task が開催される予定である．

英語文法誤り訂正の研究では，誤りのタイプを 1 つもしくは数種類に限定して誤り訂正を行なうことが一般的である．しかしながら，第 2 言語学習者の犯す誤りのタイプは様々である [6]．全てのタイプの誤りを扱うために，フレーズベース統計的機械翻訳を用いた英語誤り訂正手法が提案されている [2], [6]．フレーズベース統計的機械翻訳は翻訳単位として連続した単語の列であるフレーズを用いる．しかし，フレーズは教師なしの方法で抽出されるため，“a lot of” のような複単語表現は 1 つのフレーズとして扱えない可能性もある．機械翻訳の分野では，フレーズベース統計的機械翻訳において複単語表現を考慮することで高い性能を達成している [3], [11].

本稿では，複単語表現を考慮した英語文法誤り訂正の手法を提案する．正確に言うと，Carpuat and Diab [3] が提案した複単語表現を考慮した統計的機械翻訳の手法を英語文法誤り訂正に適応した．彼らは入力側（英語）の文中で複単語表現を 1 つの単語としてまとめて扱った．異なる 2 言語間の翻訳を行なう典型的な機械翻訳と異なり，英語文法誤り訂正は入力側の文がエラーを含む可能性がある．その問題に対して，本稿では以下の 2 つの手法を提案した．1 つ目は，入力側と出力側の文両方で複単語表現を 1 つの単語としてまとめて扱い，2 つ目は出力側の文だけで複単語表現を 1 つの単語としてまとめて扱う手法である．

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] tomoya-m@is.naist.jp
[b)] matsu@is.naist.jp

[*1] http://lang-8.com
[*2] http://www.getginger.jp

2. 関連研究

英語文法誤り訂正の研究は近年盛んに行なわれている．文法誤り訂正は大きく 2 つに分けることができる．1 つ目は，1 つもしくは数種類の誤りタイプを対象として誤りの訂正を行なうものである [12], [13], [17]．2 つ目は，全ての誤りタイプを対象として誤り訂正を行なうものである [6]．1 つ目では，Support Vector Machine のような分類器を用いて，誤り訂正を分類問題に落とし込んで解く．2 つ目では，統計的機械翻訳の手法を用いて，誤った文から正解文に翻訳する形で定式化されている．文法誤り訂正の多くの先行研究で用いられてきた素性は，単語，品詞，単語間の構文情報のみであり，複単語表現のような 2 単語（もしくはそれ以上の）連続した単語列で意味をもつ表現に関する素性は用いられていない．

複単語表現に関する言語資源を開発する研究や文中の複単語表現を同定する研究が多く行なわれている [15], [16]．また，数は多くないが，自然言語処理の応用タスク，例えば統計的機械翻訳 [3], [11]，情報検索 [8]，意見抽出 [1] でも複単語表現は用いられている．

我々の研究は，複単語表現を用いた統計的機械翻

表 1 The rate of overlap of multi-word expressions from Penn Treebank section of OntoNotes and Lang-8 Learner Corpora

top number	rate of overlap
10	30.0%
20	45.0%
30	46.7%
40	57.5%
50	54.0%
70	57.1%
120	66.7%
170	66.5%

訳 [3], [11] の研究と類似している．しかし，我々のタスクでは入力側の文が学習者の書いた文で誤りを含むため，入力側の文での複単語表現の同定に失敗する可能性がある．

3. 複単語表現

複単語表現は，単語境界（もしくはスペース）を越える表現で特異な解釈をもつ表現と定義されている [14]．本稿では，副詞，接続詞，冠詞，前置詞，前置詞句，代名詞といった機能的表現で，連続した単語で構成される（ギャップを許さない）複単語表現をメインに扱う．

3.1 ネイティブコーパスと学習者コーパスに含まれる複単語表現

英語の第二言語学習者もネイティブと同じように，作文中で多くの複単語表現を使用する．ネイティブと英語の第二言語学習者の使用する複単語表現を比べるため，ネイティブコーパスと学習者コーパスを準備した．Shigeto et al. [16] による複単語表現のデータセット，OntoNotes Release 4.0[*3]の Penn Treebank セクションに複単語表現がアノテートされたものをネイティブコーパスとして用いた．Lang-8 Learner Copora[*4]を学習者コーパスとして使用し，5.1 節で説明するツールで自動で複単語表現をつけたものを用いた．

表 1 は OntoNotes の Penn Treebank セクションと Lang-8 Learner Corpora に含まれる上位 N 件に含まれる複単語表現を取ってきた際に同じ複単語表現が含まれる割合を示す．この 2 つのコーパスは違うドメインであるが，学習者が使う複単語表現とネイティブが使う複単語表現は約 60%は同じということが分かる．

複単語表現の出現頻度はおおよそジップの法則に従っている．学習者コーパスに含まれる複単語表現のトークンを数えると，上位 70 件の複単語表現で全体のおおよそ 50%，上位 120 件の複単語表現で全体のおおよそ 80%，上位 170 件の複単語表現で全体のおおよそ 90%をカバーしている．

[*3] https://catalog.ldc.upenn.edu/LDC2011T03
[*4] http://cl.naist.jp/nldata/lang-8/

3.2 複単語表現を文法誤り訂正に用いる利点

複単語表現を文法誤り訂正に用いる利点は 2 つある．1 つ目の利点は複単語表現の一部の単語を，別の単語に訂正してしまうことを防ぐことができる点である．これを説明するために以下の例を考える．

He ate sweets, for example ice and cake.

この例文は文法誤りを含んでいないため，文法誤り訂正システムはこれを訂正する必要がない．しかしながら，システムは単語 “example” を以下のように書き換えてしまう可能性がある．

He ate sweets, for examples ice and cake.

これはシステムが，“for example” が複単語表現であることを知らないためである．

2 つ目の利点は，複単語表現を用いた場合，システムが長いコンテクストを考慮可能になる点である．例えば，以下の例文で考えると，

I have a lot of red apple.

複単語表現を考慮しなければ，誤り訂正システムは 3 グラムで，“I have a”，“have a lot”，“a lot of”，“lot of red”，“of red apple” となり，“a lot of” と “apple” の関係を直接考慮することができない．複単語表現を考慮した場合は，“a lot of red apple” が 3 グラムに含まれ，“a lot of” と “apple” の関係を直接考慮できる．

4. 複単語表現を用いた文法誤り訂正手法

この節では，複単語表現を使った文法誤り訂正の手法について説明する．我々は，文法誤り訂正の手法として統計的機械翻訳を用い，特にフレーズベース統計的機械翻訳の手法を用いた．

4.1 フレーズベース統計的機械翻訳による文法誤り訂正

フレーズベース統計的機械翻訳を利用した文法誤り訂正は Brockett et al. [2] が最初に提案した．彼らはフレーズベース統計的機械翻訳を文法誤り訂正に利用したが，彼らが扱った誤りは “名詞の単複” の誤りだけであった．Mizumoto et al. [6] もフレーズベース統計的機械翻訳の手法を文法誤り訂正に用いたが，彼らは Brockett et al. と違い全ての誤りタイプを対象として誤り訂正を行なった．

対数線形モデルを使った統計的機械翻訳 [10] の式は次のように定義される．

$$\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e,f) \qquad (1)$$

ここで e は出力側（訂正後の文）であり，f が入力側（学習者の書いた訂正前の文）である．$h_m(e,f)$ は M 個の素性関数であり，λ_m が各素性関数に対する重みである．この式はソース側の文 f に対して，素性関数の重み付き線形和を最大化するターゲット側の文 e を探せばいいことを意味

している．素性関数には，翻訳モデルや言語モデルなどが用いられる．翻訳モデルは一般にフレーズ間の翻訳確率に分解された $P(f|e)$ という条件付き確率の形で表される．言語モデルは一般に $P(e)$ という確率の形で表され，n-gram 言語モデルが広く用いられている．また，翻訳モデルは添削前後の文で 1 対 1 対応のとれた学習者コーパスから学習し，言語モデルはターゲット側言語の生コーパス（添削後の文）から学習することができる．

4.2 複単語表現を考慮した文法誤り訂正手法

我々は 2 つの複単語表現を考慮した誤り訂正の手法を提案する．複単語表現を使った機械翻訳の先行研究 [3] では，入力側の文に含まれる複単語表現の構成単語をアンダースコアによってつなげ，単純に 1 つの単語として扱うことで複単語表現を考慮した．我々は基本的に彼らと同じ方法を文法誤り訂正に応用するが，文法誤り訂正タスクの場合，入力側の文に文法誤りが含まれているため，複単語表現の同定に失敗する可能性がある．従って，以下の 2 つの手法を提案する．

入力側，出力側の両方で複単語表現を利用

この方法では，入力側，出力側の文に含まれる複単語表現を 1 つの単語として扱う．例を示すと以下のようになる．

Source: I have a_lot_of pen.
Target: I have a_lot_of pens.

出力側のみで複単語表現を利用

この方法では，入力側の文のみに対して，文中に含まれる複単語表現を 1 つの単語として扱う．例を示すと以下のようになる．

Source: I have a lot of pen.
Target: I have a_lot_of pens.

本稿では，言語モデルと翻訳モデルの両方で複単語表現考慮したテキストで学習した．

5. 複単語表現を考慮した文法誤り訂正の実験

4 節提案した複単語表現を考慮した文法誤り訂正の手法が有効であるか調べるため実験を行なった．

5.1 実験設定

フレーズベース統計的機械翻訳のツールとして，cicada 0.3.0[*5]を使用した．単語アライメントにも cicada 0.3.0 の内部実装を用いた．言語モデルには expgram 0.2.0[*6]を使用し，5-gram 言語モデルを構築した．統計的機械翻訳のモデルのパラメータ調整には ZMERT [*7]を使用し，F 値を最適化するようにパラメータのチューニングを行なった．

複単語表現を自動で同定するために，AMALGr 1.0[*8] [15] を用いた．複単語表現を同定するツールは, Shigeto et al. [16] が Penn Treebank sections of OntoNotes Release 4.0 に複単語表現をアノテートしたデータを用いて再学習を行なった．これは，文献 [16] のアノテートが我々の目的に，より適していたからである．

トレーニングデータとして Lang-8 Learner Corpora v2.0 を使用した．本稿では Lang-8 Learner Corpora から日本人学習者が書いた英語の作文のみを用い，データに含まれるノイズを除くため，文献 [7] の方法をもちいた．この結果，629,787 文対が抽出され，これを翻訳モデルと言語モデルの構築に使用した．

テストデータおよびパラメータチューニングに用いるデベロップメントデータとして Konan-JIEM コーパスを使用した．テストデータとして，EDCW2012 [*9]のドライラン用である 170 エッセイ，2,411 文を使用した．デベロップメントデータとして，EDCW2012 のフォーマルラン用の 63 エッセイからランダムに 300 文取り出したものを使用した．

5.2 実験結果

評価指標として，適合率，再現率，F 値を用いた．ベースラインとして複単語表現を利用しないフレーズベース統計的機械翻訳を使った文法誤り訂正を用い，4.2 節で提案した 2 つの複単語表現を用いた手法と比較を行なった．それに加えて，翻訳モデルと言語モデルを構築するデータに対して，利用する複単語表現の数を変化させて実験を行なった．これは，コーパス中にほとんど出現しない複単語表現はノイズになると考えたからである．3.1 節で述べた，上位 70 件 (50%), 120 件 (80%) and 170 件 (90%) と全てを利用した場合で実験を行なった．

表 2 に実験結果を示す．複単語表現を考慮した誤り訂正の手法は，複単語表現全てを考慮した場合を除いてベースラインよりも高い F 値を達成した．また，使用する複単語表現を増やすと F 値は上がっている．

5.3 考察

実験結果を見ると，全ての複単語表現を利用すると F 値が下がっている．これは，コーパス中にほとんど出現しない複単語表現が学習時とテストの際にノイズになっているからであると考える．複単語表現を出力側のみで用いた場合の方が，入力側と出力側の両方で用いた場合よりも良い結果であった．これは学習者が複単語表現の一部を間違って書いて，システムが複単語表現の同定に失敗することがあるためだと考える．また，実験結果で適合率が下がっているが，この理由のひとつとして，学習者が誤って使用し

[*5] http://www2.nict.go.jp/univ-com/multi_trans/cicada/
[*6] http://www2.nict.go.jp/univ-com/multi_trans/expgram/
[*7] http://cs.jhu.edu/~ozaidan/zmert/
[*8] https://github.com/nschneid/pysupersensetagger
[*9] https://sites.google.com/site/edcw2012/

表 2 文法誤り訂正の実験結果

		適合率	再現率	F 値
ベースライン (w/o 複単語表現)		0.301	0.329	0.314
入力: w/複単語表現, 出力: w/複単語表現	70 (50%)	0.273	0.378	0.317
	120 (80%)	0.300	0.349	0.322
	170 (90%)	0.279	0.382	0.323
	All	0.292	0.328	0.309
入力: w/o 複単語表現, 出力: w/複単語表現	70 (50%)	0.301	0.351	0.324
	120 (80%)	0.293	0.369	0.327
	170 (90%)	0.298	0.367	**0.329**
	All	0.313	0.294	0.304

表 3 システムの誤り訂正の出力例

学習者の文	Last month, she gave me a lot of rice and onion.
ベースライン	Last month, she gave me a lot of rice and onion.
w/複単語表現	Last month, She gave me a lot of rice and onions.

ていない “many” のような単語を “a lot of” のような複単語表現に訂正してしまうことがあるためである．

表 3 はシステムの実際の訂正例である．ベースラインシステムが訂正できていないが，複単語表現を考慮した誤り訂正システムでは訂正できた．

6. Conclusion

本稿では，複単語表現を考慮した誤り訂正の手法を提案した．提案した手法は複単語表現を直接的に英語文法誤り訂正に適用しただけだが，実験の結果，複単語表現が英語文法誤り訂正に有効であることがわかった．実験結果は全ての複単語表現を考慮した場合を除き，複単語表現を考慮したシステムがベースラインシステムよりも高い F 値を達成した．今後は，本稿で扱わなかった句動詞のような複単語表現を利用した誤り訂正を行なう予定である．

謝辞

Lang-8 のデータ使用に関して，快諾してくださった喜洋洋さんに感謝いたします．

参考文献

[1] Berend, G.: Opinion Expression Mining by Exploiting Keyphrase Extraction, *Proceedings of IJCNLP*, pp. 1162–1170 (2011).
[2] Brockett, C., Dolan, W. B. and Gamon, M.: Correcting ESL Errors Using Phrasal SMT Techniques, *Proceedings of COLING-ACL*, pp. 249–256 (2006).
[3] Carpuat, M. and Diab, M.: Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation, *Proceedings of HLT-NAACL*, pp. 242–245 (2010).
[4] Dale, R., Anisimoff, I. and Narroway, G.: HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task, *Proceedings of BEA*, pp. 54–62 (2012).
[5] Dale, R. and Kilgarriff, A.: Helping Our Own: The HOO 2011 Pilot Shared Task, *Proceedings of ENLG*, pp. 242–249 (2011).
[6] Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M. and Matsumoto, Y.: The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings, *Proceedings of COLING*, pp. 863–872 (2012).
[7] Mizumoto, T., Komachi, M., Nagata, M. and Matsumoto, Y.: Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, *Proceedings of IJCNLP*, pp. 147–155 (2011).
[8] Newman, D., Koilada, N., Lau, J. H. and Baldwin, T.: Bayesian Text Segmentation for Index Term Identification and Keyphrase Extraction, *Proceedings of COLING*, pp. 2077–2092 (2012).
[9] Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C. and Tetreault, J.: The CoNLL-2013 Shared Task on Grammatical Error Correction, *Proceedings of CoNLL Shared Task*, pp. 1–12 (2013).
[10] Och, F. J. and Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, *Proceedings of ACL*, pp. 295–302 (2002).
[11] Ren, Z., Lü, Y., Cao, J., Liu, Q. and Huang, Y.: Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions, *Proceedings of Workshop on MWE*, pp. 47–54 (2009).
[12] Rozovskaya, A. and Roth, D.: Algorithm Selection and Model Adaptation for ESL Correction Tasks, *Proceedings of ACL*, pp. 924–933 (2011).
[13] Rozovskaya, A. and Roth, D.: Joint Learning and Inference for Grammatical Error Correction, *Proceedings of EMNLP*, pp. 791–802 (2013).
[14] Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A. and Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP, *Proceedings of CICLing*, pp. 1–15 (2002).
[15] Schneider, N., Danchik, E., Dyer, C. and Smith, N. A.: Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut., *TACL*, Vol. 2, pp. 193–206 (2014).
[16] Shigeto, Y., Azuma, A., Hisamoto, S., Kondo, S., Kouse, T., Sakaguchi, K., Yoshimoto, A., Yung, F. and Matsumoto, Y.: Construction of English MWE Dictionary and its Application to POS Tagging, *Proceedings of Workshop on MWE*, pp. 139–144 (2013).
[17] Tajiri, T., Komachi, M. and Matsumoto, Y.: Tense and Aspect Error Correction for ESL Learners Using Global Context, *Proceedings of ACL*, pp. 198–202 (2012).