Vol.2014-NL-216 No.4
Vol.2014-SLP-101 No.4
2014/5/22

統計的機械翻訳を用いた英語文法誤り訂正の結果をリランキングすることで訂正性能の改善はできるか？

水本 智也[1,a)]　松本 裕治[1,b)]

概要：第 2 言語を学習する人が増え，コンピュータによる第 2 言語学習支援に関する研究が盛んに行なわれている．その中でも特に英語の文法誤り訂正の研究が行なわれており，文法誤り訂正の性能を競う世界規模の Shared Task が 4 年連続で開催される．学習者の犯す誤りは様々なタイプがあり，全ての誤りタイプを訂正するために，統計的機械翻訳を用いた誤り訂正が提案されている．本稿では，統計的機械翻訳による誤り訂正結果の n-best の中に，1-best の場合よりもよい訂正が含まれていることに注目する．実際の出力結果を分析することで，リランキングによる性能向上が可能であるかを議論する．

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，英語文法誤り訂正の性能を競う HOO（2011 年，2012 年）[7], [8]，CoNLL Shared Task（2013 年）[14] 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．例えば，Rozovskaya and Roth [16] は前置詞の誤りの訂正を行ない，Tajiri ら [19] は動詞の時制の誤りの訂正を行なった．Rozovskaya and Roth [17] は冠詞，名詞の単複，動詞の誤りを同時に訂正する手法を提案したが，この手法でも訂正する誤りのタイプは限定している．

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．表 1 は日本人大学生の英文エッセイに人手で誤りを訂正し，誤りタイプを付与した Konan-JIEM コーパス [12] [*3]（以下，KJ コーパス）の誤りの分布である．冠詞が最も多い誤りタイプであり，名詞の単複，前置詞，動詞の時制と続く．誤りが訂正され，誤りタイプの付与されたコーパスとして NUS Corpus of Learner English [6] もある．このコーパスはシンガポール国立大学によって作られ，CoNLL Shared Task で使用されている．文献 [6] で示される NUCLE の誤りタイプ別の数は，*wrong collocation/idiom/preposition* が 7,312 個であり最も多く，*local redundancies* が 6,390 個，*article or determiner* が 6,004 個，*noun number* が 3,995 個と続く．2 つの学習者コーパスからわかるように，第 2 言語学習者は様々なタイプの誤りを犯すことがわかる．

表 1　Konan-JIEM コーパスにおける誤りの分布

タイプ	割合 (%)	タイプ	割合 (%)
冠詞	19.23	動詞その他	4.09
名詞の単複	13.88	副詞	3.59
前置詞	13.56	接続詞	2.04
動詞の時制	8.77	語順	1.34
名詞の語彙選択	7.04	名詞その他	1.30
動詞の語彙選択	6.90	助動詞	0.88
代名詞	6.62	語彙選択その他	0.74
動詞の人称・数の不一致	5.25	関係詞	0.42
形容詞	4.30	疑問詞	0.04

そこで，誤りを限定せず訂正を行なう手法として統計的機械翻訳を用いるものが提案されている [1], [2], [3], [10], [20]. Brockett ら [2] および Mizumoto ら [10] はフレーズベース統計的機械翻訳で訂正を行なっており，Behera and Bhattacharyya [1] は階層的フレーズベース統計的機械翻訳，Buys and Merwe [3] は統語ベース統計的機械翻訳を用いて訂正を行なった．しかし，これらの手法の直接的な性能の比較は行われていなかった．

水本ら [11] は，フレーズベース，階層的フレーズベー

[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
[*3] http://www.gsk.or.jp/catalog/gsk2012-a/

ス，統語ベース (String-to-Tree モデル) の誤り訂正性能を KJ コーパスを評価コーパスとして用いて比較を行なった．実験の結果，フレーズベースが最も高い性能，次に階層的フレーズベース，最後が統語ベースという結果であった．また，文献 [11] では，考察として次の 2 つのことが示されている．

1. 10-best の出力の中に，1-best よりも良い訂正結果が含まれている
2. SMT の手法ごとにそれぞれで訂正可能な誤りタイプは異なる

先行研究の考察を受け，本稿では統計的機械翻訳の手法で誤り訂正を行なって出力した実際の結果を見ながら考察を行ない，リランキングが可能かどうか検討する．本稿では，フレーズベース機械翻訳で誤り訂正を行ない，その出力を用いて考察を行なう．

2. フレーズベース統計的機械翻訳による誤り訂正

本稿では，誤りのタイプを限定せずに誤り訂正を行なうためにフレーズベース統計的機械翻訳の手法 [9] を用いる．文法誤り訂正にフレーズベース統計的機械翻訳を用いた先行研究には Brockett ら [2]，Mizumoto ら [10] がある．Brockett ら [2] はフレーズベース統計的機械翻訳を使って英語学習者の誤り訂正を行なったが，彼らは名詞の加算・不加算の誤りのみを対象としていた．本稿では，文献 [10] と同様にフレーズベース機械翻訳を用いて，全ての誤りのタイプを対象に訂正を行なう．

対数線形モデルを使った統計的機械翻訳 [15] 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 はターゲット側（訂正後の文）であり，f がソース側（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．

3. 統計的機械翻訳を用いた誤り訂正実験

本節では，文献 [11] で行われたフレーベース統計的機械翻訳による誤り訂正を行ない，その結果を示す．

3.1 評価尺度

評価は人手評価ではなく，自動で評価を行ない，評価尺度として単語単位による再現率，適合率および F 値を用いた．各誤りにおける再現率と適合率は評価用コーパスにアノテートされた誤りタイプをもとに true positive, false positive, false negative の数を算出して計算した．true positive はシステムが訂正を行ない正解だった箇所，false positive はシステムが訂正を行なったが訂正する必要がなかった箇所もしくは訂正が必要だったがシステムが訂正を間違えた箇所，false negative はシステムは訂正を行なわなかったが訂正が必要だった箇所である．

注意すべき点は，評価用コーパスでタグが付いてない箇所を添削した場合でも，各誤りの適合率には影響しないことである [*4]．**図 1** を使って評価の方法を説明する．この例では，システムが前置詞の 1 つ目の “to” を削除しているが，この “to” は元々誤りタグはつけられていない．これが何の誤りであるかは評価システムでは同定できないため，前置詞誤りの適合率に影響はしない．そのため，この例では，前置詞誤りに対する適合率 = 1/2，再現率 = 1/2 となり，トータルスコアに対する適合率 = 1/3，再現率 = 1/2 になる．

3.2 実験に使用したツールとデータ

フレーズベース統計的機械翻訳のツールとして，cicada 0.30 [*5]を使用した．言語モデルには expgram 0.20 [*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 値に式 2 の True negative rate (TNR) をかけたもの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．これは誤り訂正のアプリケーションでは，システムが間違って訂正することを避けるほうが適切であると考えたためである．

$$TNR = \frac{true\ negative}{(true\ negative + false\ positive)} \qquad (2)$$

トレーニングデータとして Lang-8 Learner Corpora v2.0 [*8]を使用した．このコーパスは語学学習 SNS Lang-8 からクローリングして集められたコーパスである．Lang-8 では，第 2 言語学習者が学習している言語で書いた作文を SNS に投稿すると，Lang-8 に登録しているその学習言語を母語とするユーザが添削をしてくれる．そのため，学習者の書いた文とその文に対してネイティブが添削を行なった文が対になったデータとなっている．本稿では Lang-8

[*4] トータルのスコアはタグが付いていない箇所の訂正結果も含めて計算している．
[*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] http://cl.naist.jp/nldata/lang-8/

学習者	He talked to me _ his life of Kyoto, and he took me _ Kyoto university.
正解	He talked to me about his life in Kyoto and he took me to Kyoto university.
システム	He talked _ me _ his life on Kyoto, and he took me to Kyoto university.

図 1 評価方法を説明するための例

Learner Corpora から日本人学習者が書いた英語の作文のみを用いた．学習者の書いた文に対して大きく変更を伴う添削をされている場合は，添削者のコメントが含まれている可能性がある．学習者の書いた文と訂正された文の編集距離を動的計画法で計算し，単語の挿入数，削除数ともに 5 以下のものだけ抽出した．この結果，630,117 文が抽出され，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]のドライラン用に配られた 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．

3.3 実験結果

実験結果を表 2 に示す．10-best の訂正結果を出力して，その中で最も F 値が高いものを選んだ際のスコア（オラクル）を括弧の中に示す．“語彙選択その他”，“疑問詞” 以外全ての誤りタイプで 10-best 出力時のオラクルのスコアが，1-best の場合よりも高くなることがわかる．表 3 にオラクルの出現順位とその総数を示す．10-best 出力時に，1 番目にオラクルが出現している数は 1,320 個であり，半数以上を占める．出現順位が低くなるにつれ，オラクルの出現数も減っている．

4. リランキング可能性に対する考察

本節では，3 節で行なった実験をもとに，統計的機械翻訳による誤り訂正がリランキングによって，性能向上できるかについて考察する．リランキングの利点は，フレーズベース統計的機械翻訳では使えない特徴を使うことができることである．例えば，フレーズベース統計的機械翻訳では，品詞や構文解析の結果が使えないが，リランキングを行なう際にはそれらを用いることができる．

統計的機械翻訳を用いた英語文法誤り訂正において，リランキングを行なって性能を向上するという研究はないが，他のタスクにおいてはリランキングを用いて性能を向上させる研究が行なわれている．例えば，構文解析 [5] [4] や機械翻訳 [18] などでリランキングを行なう研究が行なわれており．Maximum-Entropy や Perceptron といった機械学習の手法が応用されている．本稿では，リランキングに用いる手法ではなく，リランキングに有効な特徴（素性）に注目して考察を行なう．

[*9] https://sites.google.com/site/edcw2012/

表 2 KJ コーパスに対する統計的機械翻訳による誤り訂正の誤りタイプごとの結果．括弧の中は訂正システムが 10-best 出力した場合のオラクルのスコアである．

	再現率	適合率	F 値
冠詞	.452 (.798)	.705 (.925)	.551 (.857)
名詞の単複	.370 (.717)	.854 (.950)	**.516** (.817)
前置詞	.358 (.556)	.627 (.811)	.456 (.660)
動詞の時制	.182 (.530)	.352 (.739)	.240 (.618)
名詞の語彙選択	.175 (.261)	.500 (.677)	.260 (.377)
動詞の語彙選択	.224 (.308)	.423 (.608)	.293 (.409)
代名詞	.163 (.436)	.387 (.783)	.230 (.560)
動詞の人称・数の不一致	.378 (.730)	.561 (.875)	.451 (.796)
形容詞	.453 (.550)	.750 (.822)	.565 (.659)
動詞その他	.456 (.692)	.620 (.844)	..525 (.761)
副詞	.254 (.462)	.450 (.720)	.324 (.563)
接続詞	.255 (.352)	.875 (.864)	.394 (.500)
語順	.133 (.250)	.069 (.160)	.091 (.195)
名詞その他	.407 (.536)	.550 (.652)	.468 (.588)
助動詞	.000 (.368)	.000 (.636)	.000 (.467)
語彙選択その他	.000 (.000)	.000 (.000)	.000 (.000)
関係詞	.111 (.182)	.250 (.667)	.154 (.286)
疑問詞	.000 (.000)	.000 (.000)	.000 (.000)
トータル	.309 (.680)	.327 (.582)	.318 (.627)

表 4 冠詞を間違って訂正した例

システムの出力順位	出力文
学習者，正解	Moreover, music is necessary for me.
1	Moreover, the music is necessary for me.
2 (オラクル)	Moreover, music is necessary for me.

表 5 時制誤りの例

システムの出力順位	出力文
学習者	I wondered they make me study reading and whiting then.
正解	I wondered they made me study reading and whiting then.
1	I wondered if they make me study reading and whiting then.
6 (オラクル)	I wondered if they made me study reading and whiting then.
学習者	I don't read a book for a long time.
正解	I haven't read a book for a long time.
1	I don't read a book for a long time.
2 (オラクル)	I haven't read a book for a long time.

表 3 オラクルの出現順位と総数

出現順位	総数
1	1,320
2	381
3	227
4	120
5	100
6	80
7	57
8	52
9	46
10	28

4.1 実例による考察

表 4 に冠詞誤りを間違って訂正した例を示す．学習者の書いた文で正解であったが，“music” に “the” を付けてしまっている．リランキングするために必要な特徴が 1 文内にないため難しい例である．冠詞誤りは，おおざっぱに分類すると “a”，“the” と “冠詞を付けない” の 3 つのため，10-best を出力した際にオラクルが含まれていることが多い誤りである．しかしながら，例のように 1 文ではリランキングすることが難しい例も多い．

表 5 に時制誤りに関する例を示す．1 つ目の例では，システムが出力した 1-best の結果は “made” にしなければならないところが “make” のままである．周辺の単語しか見ることのできないフレーズベース機械翻訳では訂正が難しいが，リランキングを行なう際は “wondered” や文の最後にある “then” を考慮できるように素性を設計すれば訂正可能になると考える．2 つ目の例に関しても，完了系にしなければいけない箇所を現在形にしてしまっている例である．文の最後にある “for a long time” を素性として使うことができれば，訂正可能になると考えられる．

表 6 に冠詞，名詞の単複の例を示す．“are” と “a funny book” が一致していないため誤りである．このような誤りに関しても，be 動詞と名詞をセットにした素性を作ることで，リランキング可能であると考えられる．

4.2 リランキングの素性設計に関する考察

4.1 節での実例を示し，リランキング可能であるか検討を行なった．素性設計を工夫すればリランキング可能であると言えるが，実際にその素性を作れるとは限らない．訂正後の文であっても誤りを含む可能性がある．誤りが含まれていると品詞付与や構文解析に失敗してしまう可能性があり，リランキングに有効な素性を作ることができないためである．

例として，表 7 を見る．この例では，“play”，“watch”，“eat” の関係が重要である．訂正後の文は，to play は変えていないが，“watch” を “watching” に，“eat” を “eating” に訂正している．学習者の文，正解の文であれば 3 つの動詞の関係を構文解析することで取ってくることができるが，訂正後の文ではこの 3 つ動詞の関係が並列であると解析することができず，リランキングに有効な素性の設計も難しい．例にあげた以外にも動詞の一致の誤りなどは構文解析の失敗によって，主語と動詞の関係が抽出できずにリランキングできない可能性がある．

5. おわりに

本稿では，統計的機械翻訳による誤り訂正の n-best 出力の中に，1-best の場合よりもよい訂正が含まれていることに注目し，その出力をリランキングすることが可能か検討した．フレーズベース統計的機械翻訳で誤り訂正を行ない，出力結果を見てリランキングが可能か分析した．誤り種類

表 6 冠詞，名詞単複誤りの例

システムの出力順位	出力文
学習者	If there are a funny book, I may read one.
正解	If there are funny books, I may read one.
1	If there are a funny book, I may read one.
9 (オラクル)	If there are funny books, I may read one.

表 7 素性抽出の難しい文の例

	出力文
学習者	For example, that is to play sports, watch TV and eat dinner with my friends, and so on.
正解	For example, they are to play sports, watch TV and eat dinner with my friends, and so on.
訂正後	For example, that is to play sports, watching TV, eating dinner with my friends, and so on.

によっては，1 文だけではリランキングが難しいものもあるが，リランキングでは構文解析結果なども用いて遠くの関係を見ることが可能であるため，リランキングすることで性能は改善できる．一方，誤りが残っている文で，構文解析に失敗すると，単語間の関係を捉えることができずリランキングに失敗する可能性がある．構文解析を使用せずに，誤り検出を行なう手法 [13] が提案されている．この文献で提案されている手法を用いて素性を作ることで，構文解析誤りに影響を受けずにリランキングが可能であると考える．

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] Behera, B. and Bhattacharyya, P.: Automated Grammar Correction Using Hierarchical Phrase-Based Statistical Machine Translation, *Proceedings of IJCNLP*, pp. 937–941 (2013).
[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] Buys, J. and van der Merwe, B.: A Tree Transducer Model for Grammatical Error Correction, *Proceedings of CoNLL Shared Task*, pp. 43–51 (2013).
[4] Charniak, E. and Johnson, M.: Coarse-to-fine N-best Parsing and MaxEnt Discriminative Reranking, *Proceedings of ACL*, pp. 173–180 (2005).
[5] Collins, M.: Discriminative Reranking for Natural Language Parsing, *Proceedings of ICML*, pp. 175–182 (2000).
[6] Dahlmeier, D., Ng, H. T. and Wu, S. M.: Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31 (2013).
[7] 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).
[8] Dale, R. and Kilgarriff, A.: Helping Our Own: The HOO 2011 Pilot Shared Task, *Proceedings of ENLG*, pp. 242–249 (2011).
[9] Koehn, P., Och, F. J. and Marcu, D.: Statistical Phrase-Based Translation, *Proceedings of HLT-NAACL*, pp. 48–54 (2003).
[10] 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).
[11] 水本智也，松本裕治：統計的機械翻訳に基づく英語文法誤り訂正におけるフレーズベースと統語ベースの比較と分析，第 20 回言語処理学会年次大会，pp. 258–261 (2014).
[12] Nagata, R., Whittaker, E. and Sheinman, V.: Creating a manually error-tagged and shallow-parsed learner corpus, *Proceedings of ACL-HLT*, pp. 1210–1219 (2011).
[13] 永田亮：構文解析を必要としない主語動詞一致誤り検出手法，電子情報通信学会論文誌. D, 情報・システム，Vol. 96, No. 5, pp. 1346–1355.
[14] 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).
[15] Och, F. J. and Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, *Proceedings of ACL*, pp. 295–302 (2002).
[16] Rozovskaya, A. and Roth, D.: Algorithm Selection and Model Adaptation for ESL Correction Tasks, *Proceedings of ACL*, pp. 924–933 (2011).
[17] Rozovskaya, A. and Roth, D.: Joint Learning and Inference for Grammatical Error Correction, *Proceedings of EMNLP*, pp. 791–802 (2013).
[18] Shen, L., Sarkar, A. and Och, F. J.: Discriminative Reranking for Machine Translation, *Proceedings of HLT-NAACL*, pp. 177–184 (2004).
[19] 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).
[20] Yuan, Z. and Felice, M.: Constrained Grammatical Error Correction using Statistical Machine Translation, *Proceedings of CoNLL Shared Task*, pp. 52–61 (2013).