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

水本智也　松本裕治
奈良先端科学技術大学院大学
{tomoya-m,matsu}@is.naist.jp

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，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）[5, 4]，CoNLL Shared Task（2013 年）[8] も開催された．

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 [9] は前置詞の誤りの訂正を行ない，Tajiri ら [11] は動詞の時制の誤りの訂正を行なった．Rozovskaya and Roth [10] 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，一般的に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 [7] [*3]の誤りの分布である．冠詞，名詞の単複，前置詞に限らず，様々なタイプの誤りを犯していることがわかる．

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 [2, 6, 1, 12, 3]．Brockett ら [2] および Mizumoto ら [6] はフレーズベース統計的機械翻訳（以下，単にフレーズベース）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（以下，単に階層的フレーズベース），Buys and Merwe [3] は統語ベース統計的機械翻訳（以下，単に統語ベース）を用いて訂正を行なっている．フレーズベースで訂正可能な誤りのタイプは [6] で議論されている．しかしながら統語ベースは，訂正する誤りタイプが限定された Shared Task で提案されたこともあり，全ての誤りタイプを対象とした際，どの誤りタイプに有効であるか議論が行なわれていない．

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

また，統語ベースとフレーズベースを用いた誤り訂正との比較も行なわれていない．そこで本稿ではフレーズベースと統語ベースで，誤り訂正性能の比較を行なう．また，フレーズベースと比較して統語ベースがどのタイプの誤りに対して有効かを調べ，議論する．

2　統計的機械翻訳を用いた誤り訂正に関する関連研究

統計的機械翻訳手法を用いた英語誤り訂正にはフレーズベースによるものと階層的フレーズベース，統語ベースによるものがある．統計的機械翻訳を用いた誤り訂正は，最初に Brockett ら [2] によって提案された．彼らはフレーズベースを用いたが，訂正する誤りのタイプを名詞の単複のみに限定していた．Mizumoto ら [6] は Brockett らの研究を受け，フレーズベースを用いて全ての誤りを対象に訂正を行ない，フレーズベースで訂正可能な誤りについて議論した．

Behera and Bhattacharyya [1] は階層的フレーズベースを用いて誤り訂正を行なった．階層的フレーズベースの場合は，通常の下記のような翻訳ルールも使用でき，[X] に入る単語によらず “has” を “have” に訂正することができる．

few has [X] → few have [X]

他の研究が F 値を用いて評価を行なっているのに対して Behera and Bhattacharyya は評価指標に BLEU を使用しており，他の研究との比較が難しい．またフレーズベースによる誤り訂正との比較もされておらず，実際にどの程度誤り訂正に有効かわからないという問題がある．

統語ベースを用いた誤り訂正には, Buys and Merwe [3] のものがある．Buys and Merwe は，統語ベースの中で

[*1] http://lang-8.com
[*2] http://www.getginger.jp
[*3] http://www.gsk.or.jp/catalog/gsk2012-a/

も String-to-Tree モデルを採用している．String-to-Tree モデルは入力側が平文で，出力側に構文を仮定した翻訳を行なう．誤り訂正では入力側が学習者の書いた文で誤りを含んでおり構文解析に失敗する可能性がある．String-to-Tree モデルは，訂正された正しい文のみを構文解析すればよいため誤りを含んだ文を構文解析する必要がなく，訂正後の文で構文を考慮した結果を出力するため，文法的に正しい訂正が期待できる．しかしながら，Buys and Merwe は CoNLL Shared Task で提案されたため，5 つの誤りを対象とした評価しか行なわれていない．

3 フレーズベースと統語ベースによる誤り訂正の比較実験

フレーズベースと統語ベースの統計的機械翻訳を用いて誤り訂正の実験を行ない，全ての誤りを対象とした際にどのような違いがあるかの調査を行なった．本稿では，フレーズベース，階層的フレーズベース，統語ベースの統計的機械翻訳の手法の比較を行なう．

統語ベースは先行研究で用いられた String-to-Tree モデルに加えて Tree-to-Tree モデルでも実験を行なう．Tree-to-Tree モデルでは，訂正された文だけを構文解析するのではなく，学習者の書いた文に対しても構文解析を行なう．学習者の文は正しく構文解析できないために，結果的に訂正できないという可能性も考えられるが，学習者の文を解析してできた構文と正しい文を解析した構文とで対応をとることができれば訂正できる．そのため，本稿では Tree-to-Tree モデルとも比較を行なう．

3.1 各モデルで訂正できる誤りの予想

実際に実験を行なう前に，本稿で用いる統計的機械翻訳モデルの特徴から各モデルでどのような誤りに向いているかを議論する．結論から述べると，各モデルで訂正できる誤りはコーパスでタグが付けられた誤りタイプで分けることは難しく，訂正に必要な手がかりがどこにあるかで分類されると考える．フレーズベースで訂正できる誤りは [6] で言及されているように，局所的な単語列だけで訂正できる “冠詞”，“前置詞”，“形容詞” などの誤りに有効であると考える．しかしながら，“冠詞”，“前置詞”，“形容詞” 誤りの中にも，局所的な単語列で訂正できない誤り，すなわち訂正の手がかりが訂正対象の単語と離れている誤りが存在する．例えば，“a big Snoopy dolls” の “dolls” を “doll” に訂正するには，前方に “a” があることを知る必要がある．階層構造を用いることで離れた単語を考慮することができるため，階層的フレーズベースや統語ベースでは，このような誤りを訂正できると考える．またフレーズベースで訂正が難しい誤りとして単語の入れ替えを行なう “語順” 誤りがある．これに対して，階層構造を考慮できる階層的フレーズベース，統語ベースでは，句を入れ替える変換が容易に行なえるため，語順入れ替えの誤りの訂正が可能であると考えられる．

3.2 評価尺度

評価尺度として，自動評価尺度を使用し，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 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 であり，トータルスコアに対する適合率 = 1/3，再現率 = 1/2 になる．

3.3 実験に使用したツールとデータ

統計的機械翻訳のツールとして，cicada 0.30 [*5] を使用した．cicada はフレーズベース，階層的フレーズベースに加え，統語ベースを実装している．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 1 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．構文解析のツールとして，Stanford parser 3.2.0 [*8] を用いた．

$$TNR = \frac{true\ negative\ の総数}{(true\ negative\ の総数 + false\ positive\ の総数)} \quad (1)$$

トレーニングデータとして Lang-8 Learner Corpora v2.0 [*9] 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，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

[*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://nlp.stanford.edu/software/lex-parser.shtml
[*9] 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

表 2　各誤りごとの統計的機械翻訳による誤り訂正の結果．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．

	フレーズベース			階層的フレーズベース			String-to-Tree			Tree-to-Tree		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
冠詞	.452	.705	**.551**	.395	.677	.499	.395	.668	.497	.413	.720	.525
名詞の単複	.370	.854	**.516**	.350	.756	.478	.323	.533	.402	.329	.784	.463
前置詞	.358	.627	.456	.408	.586	**.481**	.369	.525	.433	.231	.473	.310
動詞の時制	.182	.352	.240	.149	.295	.198	.124	.239	.163	.183	.443	**.259**
名詞の語彙選択	.175	.500	.260	.197	.460	**.276**	.220	.341	.268	.113	.462	.181
動詞の語彙選択	.224	.423	.293	.233	.425	**.301**	.237	.329	.276	.145	.404	.213
代名詞	.163	.387	.230	.219	.451	**.295**	.188	.356	.246	.113	.333	.168
動詞の人称・数の不一致	.378	.561	**.451**	.337	.593	.429	.347	.508	.413	.248	.542	.340
形容詞	.453	.750	**.565**	.439	.642	.521	.449	.548	.494	.429	.636	.512
動詞その他	.456	.620	.525	.371	.479	.418	.383	.451	.414	.453	.739	**.562**
副詞	.254	.450	.324	.329	.535	**.407**	.188	.356	.246	.278	.611	.383
接続詞	.255	.875	**.394**	.241	.813	.371	.184	.529	.273	.148	.727	.246
語順	.133	.069	.091	.412	.212	**.280**	.333	.188	.240	.067	.037	.048
名詞その他	.407	.550	.468	.600	.750	**.667**	.464	.619	.531	.231	.375	.286
助動詞	.000	.000	.000	.158	.429	**.231**	.111	.286	.160	.100	.400	.160
語彙選択その他	.000	.000	.000	.214	.333	.261	.500	.538	.519	.278	.611	**.383**
関係詞	.111	.250	.154	.250	.333	**.286**	.222	.400	.286	.250	.333	**.286**
疑問詞	.000	.000	.000	1.00	1.00	**1.00**	.000	.000	.000	.000	.000	.000
トータル	.309 (.680)	.327 (.582)	**.318** (.627)	.273 (.691)	.326 (.607)	.297 (**.646**)	.310 (.493)	.219 (.577)	.257 (.532)	.292 (.676)	.272 (.528)	.282 (.593)

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 Learner Corpora から日本人学習者が書いた作文のみを用いた．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 [*10]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．

3.4　実験結果

表 2 に各統計的機械翻訳モデルの誤りタイプ別の実験結果を示す．括弧の中の数字は，システムが出力した上位 10 個の中で最も性能が良くなる訂正（オラクル）を選んだ場合の再現率，適合率，F 値である．

トータルの F 値を見ると，フレーズベースが最も高い F 値を達成した．統語ベースはフレーズベース，階層的フレーズベースと比べると全体的に性能が低い．この原因のひとつには，Lang-8 や Konan-JIEM コーパスの訂正後の文であっても，日本人特有の固有名詞や単語があり構文解析に失敗しているからであると考える．全てのモデルの 10-best をマージして，オラクルのスコアを計算すると再現率=0.684，適合率=0.780，F 値=0.729 であった．このことからモデルによって訂正できる誤りが異なっていることが分かる．

4　考察

3.1 節で予想したような誤りタイプごとではなく，訂正の手がかりとなる単語の位置によって，どのモデルで訂正できるかが変わることがわかった．これは各モデルの出力をマージした際のオラクルの F 値が 0.729 とフレーズベースのみの場合より向上していることからも読み取れる．

以下，各モデルで訂正できた誤りを実際に見ながら考察を行なう．表 3 にフレーズベースで訂正できた誤りの例を示す．(1) は学習者がよく間違うタイプの誤りであり，トレーニングコーパスの中にも出現するため訂正可能である．(2) に関しては，日本人英語学習者は “in” と “on” の使いわけをよく間違い，“in train” の間違いもトレーニングコーパス内に出現する．それに加えて，訂正対象の単語 “in” と手がかりの “train” が近くにあるため

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

表 3　フレーズベースモデルで訂正できる誤り

	学習者	正解
(1) 冠詞&名詞の単複	I like reading a book very much.	I like reading books very much.
(2) 前置詞	One day, I was in a train.	One day, I was on a train.

表 4　階層的フレーズベース，String-to-Tree で訂正できる誤り．

	学習者	正解
(3) 冠詞	My fevarite book is _ Harry Potter series.	My fevarite book is the Harry Potter series.
(4) 名詞の単複	I can only cook the dish of egg.	I can only cook the dish of eggs.
(5) 前置詞	I like to play kendo because i can to enjoy it.	I like to play kendo because i can _ enjoy it.
(6) 語順	It took me to drive about two hours.	It took me about two hours to drive.

訂正可能である．フレーズベースでは，このように学習者がよく犯す誤りで，訂正対象の単語と手がかりとなる単語が近い場合は訂正が可能である．

表 4 に階層的フレーズベース，String-to-Tree モデルで訂正できた誤りを示す．(3) の冠詞の誤りをフレーズベースで訂正するためには，トレーニングコーパスにも “Harry Potter series” が出ている必要がある．階層的なモデルであれば “the X series” もしくは “the NP series” という情報を使うことで訂正可能である．(4) の名詞の単複，(5) の前置詞の誤りに関しても (3) と同じような理由で，階層的な構造により訂正可能である．(6) のような語順誤りに関しても階層的モデルであれば，[X1: to drive]，[X2: about two hours] のようになっていれば，[X1 X2] → [X2 X1] のようにでき，訂正可能である．語順誤りは階層的フレーズベース，String-to-Tree モデルが他の 2 つのモデルよりも訂正できた．Tree-to-Tree モデルは学習者の文の構文解析に失敗してしまったため，訂正できなかったと考える．

10-best のオラクルの F 値を見ると，1-best のスコアの 2 倍近くになっている．そのため統計的機械翻訳モデルの 10-best の出力をリランキングし直すことで性能の向上が期待できる．また，各モデルの出力をマージした際のオラクルのスコアが最も高くなることから，各モデルの出力を組み合わせてリランキングすることで，それぞれのモデルで訂正ができない箇所を補うことができ，1 つのモデルの場合よりも性能の向上が期待できる．

5　おわりに

本稿では，誤りのタイプを限定せずに訂正できる統計的機械翻訳を使った誤り訂正に注目した．統計的機械翻訳手法による誤り訂正は，フレーズベース，階層的フレーズベース，統語ベースを使った手法が提案されている．しかしながらフレーズベースを除く手法では，全ての誤りを対象とした場合にどのような誤りを訂正可能であるか議論は行なわれていなかった．そこでフレーズベース，階層的フレーズベース，統語ベースの手法を用いて，全ての誤りを対象に訂正し，比較を行なった．

実験の結果，1-best のトータルスコアはフレーズベースが最も F 値が高かった．誤りタイプごとに見ても，フレーズベースの性能が統語ベースのものより高いものが多い．しかしながら，フレーズベースで訂正することのできない誤りを他のモデルでは訂正できる場合があった．そのため，各統計的機械翻訳のそれぞれの出力から最も訂正できている 1 文を選ぶことで訂正性能の向上ができると考える．今後は，各統計的機械翻訳モデルの訂正結果から訂正として尤もらしいものを選択するタスクに取り組む予定である．

謝辞

Lang-8 のデータの使用に関して，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] B. Behera and P. Bhattacharyya, “Automated Grammar Correction Using Hierarchical Phrase-Based Statistical Machine Translation,” Proceedings of IJCNLP, pp.937–941, 2013.
[2] C. Brockett, W.B. Dolan, and M. Gamon, “Correcting ESL Errors Using Phrasal SMT Techniques,” Proceedings of COLING-ACL, pp.249–256, 2006.
[3] J. Buys and B. van der Merwe, “A Tree Transducer Model for Grammatical Error Correction,” Proceedings of CoNLL Shared Task, pp.43–51, 2013.
[4] R. Dale, I. Anisimoff, and G. Narroway, “HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task,” Proceedings of BEA, pp.54–62, 2012.
[5] R. Dale and A. Kilgarriff, “Helping Our Own: The HOO 2011 Pilot Shared Task,” Proceedings of ENLG, pp.242–249, 2011.
[6] T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto, “The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings,” Proceedings of COLING, pp.863–872, 2012.
[7] R. Nagata, E. Whittaker, and V. Sheinman, “Creating a manually error-tagged and shallow-parsed learner corpus,” Proceedings of ACL-HLT, pp.1210–1219, 2011.
[8] H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, “The CoNLL-2013 Shared Task on Grammatical Error Correction,” Proceedings of CoNLL Shared Task, pp.1–12, 2013.
[9] A. Rozovskaya and D. Roth, “Algorithm Selection and Model Adaptation for ESL Correction Tasks,” Proceedings of ACL, pp.924–933, 2011.
[10] A. Rozovskaya and D. Roth, “Joint Learning and Inference for Grammatical Error Correction,” Proceedings of EMNLP, pp.791–802, 2013.
[11] T. Tajiri, M. Komachi, and Y. Matsumoto, “Tense and Aspect Error Correction for ESL Learners Using Global Context,” Proceedings of ACL, pp.198–202, 2012.
[12] Z. Yuan and M. Felice, “Constrained Grammatical Error Correction using Statistical Machine Translation,” Proceedings of CoNLL Shared Task, pp.52–61, 2013.