Vol.2012-NL-209 No.5
2012/11/22

# 大規模英語学習者コーパスを用いた英作文の文法誤り訂正の課題分析

水本 智也[1,a)]　林部 祐太[1,b)]　小町 守[1,c)]　永田 昌明[2,d)]　松本 裕治[1,e)]

**概要**：英語学習者の書く作文には様々な種類の文法誤りが含まれている．英語学習者の文法誤りの自動訂正に取り組んだ先行研究は，訂正する誤りの種類を数種類に限定して取り組んできた．文法誤りの中には，ヒューリスティックスを用いたルールで訂正できるものもあれば，長距離の依存関係や選択選好を考慮した統計的なモデルを用いないと訂正が難しいものもある．しかしながら，学習者の書いたテキストに対するアノテーションは時間がかかるため，最近になるまで一般に入手できる大規模な学習者コーパスは存在していなかった．そのため，英語学習者の文法誤り訂正で学習者コーパスのサイズがどのように影響するかは分かっていない．そこで，本稿では，学習者の誤りが訂正された大規模な学習者コーパスを用いてフレーズベース統計的機械翻訳の手法によって誤り訂正を行ない，学習者コーパスのサイズを変化させ，学習者コーパスのサイズがどのタイプの文法誤りに影響があるかを調べた．

# An Analysis of Problems in Grammatical Error Correction of ESL Writings Using a Large Learner Corpus of English

***Abstract:*** English as a Second Language (ESL) learners' writings contain various grammatical errors. Previous research on automatic error correction for ESL learners' grammatical errors deals with restricted types of learners' errors. In grammatical errors, some errors can be corrected by rules using heuristics, while others are difficult to correct without statistical model using native corpora and/or learner corpora. However, since error annotation to learners' text is time-consuming, it was not until recently that large scale learner corpora become publicly available. As a result, little is known about the effect of learner corpus size in ESL grammatical error correction. Thus, in this paper, we build an error correction system with phrase-based statistical machine translation (SMT) technique trained on a large scale error-tagged learner corpus to see the effect of learner corpus size for each type of grammatical errors. We show that phrase-based SMT approach is effective in correcting frequent errors that can be identified by local context, and that it is difficult for phrase-based SMT to correct errors that need long range contextual information.

## 1. はじめに

英語学習者の書く作文には多くの種類の誤りが含まれている．近年，学習者の英語に対するコーパスアノテーションが盛んになり，学習者の書いた作文の文法誤りの詳細な解析が可能になりつつある．たとえば，Konan-JIEM Learner Corpus（以下，KJ コーパスとする．）[*1] は日本人大学生によって書かれた英語のエッセイから構成される，誤りが付与されたコーパスのひとつである．表 1 は KJ コーパスの誤りを種類ごとに分類したものである [*2]．最も頻出する誤りは冠詞に関する誤りであり，そのあと名詞の単複，前置詞，動詞の時制の誤りと続いている．誤りの割合が多い誤りの種類は文中でよく出現するものであり驚くべきことではないが，学習者コーパスに多くの異なる種類の誤りが含まれていることは注意する必要がある．

これまで，英語学習者の犯す誤りに関して多くの自動誤り訂正の研究が行なわれてきた．しかしながら，これらの第 2 言語学習に関するほとんどの研究は学習者の誤りをひ

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
[a)] tomoya-m@is.naist.jp
[b)] yuta-h@is.naist.jp
[c)] komachi@is.naist.jp
[d)] nagata.masaaki@lab.ntt.co.jp
[e)] matsu@is.naist.jp

[*1] http://www.gsk.or.jp/catalog/GSK2012-A/catalog.html

[*2] KJ コーパスではスペリング誤りはアノテートの対象から除外されている．

© 2012 Information Processing Society of Japan

とつもしくは数種類に限定して誤り訂正を試みてきた．例えば，Rozovskaya and Roth [16] は前置詞の誤り，Liu ら [11] は動詞選択に関する誤り，Tajiri ら [18] は動詞の時制，Lee and Seneff [10] は動詞の語形に関する誤り（動詞の一致，動詞の時制），Dahlmeier and Ng [3] は前置詞と冠詞の誤り，Park and Levy [15] はスペリング誤り，冠詞，前置詞，語形（動詞の一致，動詞の時制）を対象として訂正を行なった．最近では，Swanson and Yamangil [17] が Cambridge Learner Corpus の訂正されている全ての種類の誤りに対して詳細な分析を行なった．しかしながら，彼らのタスクは他の研究とは異なっており，彼らの目的は学習者の書いた文とその添削文が与えられたときに誤りを検出して誤りの種類を選択するものである．

文法誤りの中には，動詞の一致の誤りのようにヒューリスティックを用いた単純なルールで訂正可能なものもあれば，前置詞誤りのようにネイティブコーパスや学習者コーパスからトレーニングした統計的なモデルを用いないと訂正が難しいものもある．しかし，最近になるまで，文法誤り訂正のための大規模な学習者コーパスは広く入手できなかった．そのため，英語学習者の文法誤り訂正で学習者コーパスのサイズがどのように影響するかはほとんど分かっていない．

本稿では，学習者の誤りが訂正された大規模な学習者コーパスを使って全ての誤りを対象とした誤り訂正の実験を行ない，また，学習者コーパスのサイズを変化させ，学習者コーパスのサイズが文法誤り訂正に及ぼす影響を調べた．本稿では，フレーズベース統計的機械翻訳を用いた誤り訂正システムを用いた．また，Web から誤りが訂正された大規模な英語学習者コーパスを作成した．そして，誤り種類別ごとに分類して誤り訂正結果の分析を行ない，大規模な学習者コーパスを用いた統計的機械翻訳のアプローチの長所，短所について議論を行なう．

本稿の主な貢献は以下の 2 つである．

- 我々の知る限り，全ての種類の誤りを訂正するために大規模な学習者コーパスを使用することを試みたのは初めてである．
- 学習者コーパスのサイズの影響をフレーズベース統計的機械翻訳のアプローチを使って調べ，その長所と短所を示す．

以下，2 節で文法誤り訂正の先行研究について簡潔に述べる．3 節で文法誤り訂正システムと大規模な誤りが訂正された学習者コーパスについて説明する．4 節で実験結果を示し，フレーズベース統計的機械翻訳を用いた誤り訂正システムにおける学習者コーパスのサイズとそれぞれの誤り種類の関係についての議論を行なう．

## 2. 関連研究

学習者の英語に対する誤り訂正に関する研究は数多くあるが，さまざまなタイプの文法誤りを対象とした研究はほとんどない．

最初は Brockett ら [2] が提案したフレーズベース統計的機械翻訳を使った誤り訂正モデルである．統計的機械翻訳のモデル自体は全てのタイプの誤りを扱うことができるが，その当時は大規模な学習者コーパスがなかったため，彼らは人工データを使って名詞の単複の誤りだけで評価を行なった．我々はフレーズベース統計的機械翻訳で実世界の大規模な学習者コーパスを使うことを最初に試みた．4 節で，フレーズベース統計的機械翻訳はスパースなデータに弱いことを示す．

第 2 に，Park and Levy [15] は大規模なアノテートされてない英語学習者コーパスを用いて，ノイジーチャンネルモデルによっていくつかの種類の誤りの訂正を行なった．我々は多数のネイティブスピーカによって誤りがアノテートされた大規模学習者コーパスを用いている点で異なる．加えて，彼らはスペリング，冠詞，前置詞，語形の誤りを対象としているのに対して，我々は統計的機械翻訳の手法を使うことで誤りのタイプを限定することなく訂正を行なった，

3 番目は，Han ら [8] による大規模な誤りが付与された英語学習者コーパスを使った前置詞誤り訂正システムである．彼らは学習者コーパスとネイティブコーパスでトレーニングした前置詞に対する最大エントロピーベースの訂正モデルを構築した．我々は誤りが付与された大規模な英語学習者コーパスを生かし，様々な誤りのタイプを扱い，学習者コーパスの全てを使うためにフレーズベース統計的機械翻訳を使う．

最近では，Dahlmeier and Ng [4] はスペリング，冠詞，前置詞，句読点，名詞の単複に対してビームサーチデコーダを使った手法を提案した．彼らは彼らの提案した識別モデルが数百文でトレーニングされた統計的機械翻訳のベースラインよりもかなり良い結果を達成したと報告している．後で詳しく述べるが，小規模な学習者コーパスで統計的機械翻訳システムを学習した場合で前置詞誤り訂正において似た傾向を観測した．しかしながら，本稿では，データのスパース性の問題を解決するために Web から抽出した誤りがアノテートされた大規模なコーパスを使用した．

## 3. 大規模学習者コーパスを使ったフレーズベース統計的機械翻訳による文法誤り訂正

### 3.1 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 [12]，Ehsan and Faili [7] 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，

© 2012 Information Processing Society of Japan

表 1　KJ 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 |

彼らは名詞の加算・不加算の誤りのみを対象としていた．Mizumoto ら [12] は学習者の犯す誤りを限定せずに誤り訂正を行なったが，彼らは英語ではなく日本語を対象としていた．Ehsan and Faili [7] は統計的機械翻訳のフレームワークを英語とペルシャ語の誤り訂正に適用したが，人工的に作成したコーパスを使用していた．

対数線形モデルを使った統計的機械翻訳 [13] 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，$\lambda_m$ 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から学習することができる．

### 3.2　大規模英語学習者コーパス作成のクラウドソーシング

統計的機械翻訳を使った誤り訂正システムのトレーニングのために，言語学習 SNS Lang-8 [*3] のデータを用いた．言語学習者が自分の作文を Lang-8 のサイトに投稿すると，Lang-8 をやっ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．Lang-8 から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大規模なデータを手に入れることができる．Mizumoto ら [12] は Web から学習者コーパスを構築するアプローチを最初に提案したが，我々は日本語ではなく英語のコーパスの構築を行なった点で異なっている．また，Tajiri ら [18] とも異なり，我々は英語学習者の母語を特定するためにユーザのメタデータを使用した．なぜならば，実験でテストコーパスに用いたコーパス (KJ Corpus) は日本人大学生によって書かれたものであり，本稿では母語による誤り訂正の影響を見ることは対象としておらず，同じ種類のデータを使用したかったためである．

2010 年 12 月までの Lang-8 のブログエントリをクローリングを行ない獲得した[*4]．日本人英語学習者の書いた Lang-8 の作文を統計的機械翻訳による誤り訂正システムの翻訳モデルと言語モデルを学習するために使用した．日本語を母語とする英語学習者の書いた英語の作文は 509,106 文対であった．しかしながら，学習者の書いた文が大きく添削されている場合は統計的機械翻訳でアライメントがとりにくく，結果として精度を下げる要因となるためノイジーな文のフィルタリングが必要となる．そこで，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 単語以下のものだけに限定し [*5]，実験には 391,699 文対を使用した．

## 4.　文法誤り訂正における学習者コーパスのサイズによる影響の調査実験

大規模学習者コーパスを使ったフレーズベース統計的機械翻訳による英語文法誤り訂正の実験を行なった．コーパスサイズの違いによる影響を見るために，Lang-8 コーパス（大規模学習者コーパス）を使いサイズを変化させたシステムと KJ コーパス（小規模学習者コーパス）を使ったシステムとの比較を行なった．誤り訂正のアプローチの影響をさらに詳しく見るために，識別モデルの手法として最大エントロピーモデルの手法と統計的機械翻訳ベースモデルの手法を用いて前置詞誤り訂正タスクで実験を行なった．

[*3] http://lang-8.com/

[*4] http://cl.naist.jp/nldata/lang-8/lang-8-url-201012.txt.gz

[*5] Mizumoto らが日本語でフィルタリングを行なった際に，削除数，挿入数 5 文字以下のものを使用していたため，本稿でもその値を参考とし 5 単語以下に限定して用いた．

© 2012 Information Processing Society of Japan

### 4.1 ツールと実験に使用したデータ

フレーズベース統計的機械翻訳を用いた全ての誤り訂正システムでは，Moses 2010-08-13 [*6]をデコーダ，GIZA++ 1.0.5 [*7]をアライメントのツールとして利用した．フレーズ抽出は grow-diag-final-and [14] ヒューリスティックを用いた．Lang-8 コーパスの全データを使用した際に抽出されたフレーズ数は 1,050,070 (245MB) であった．Lang-8 コーパスの訂正済みのテキストで学習した単語 3-gram を言語モデルとして用いた．

次に最大エントロピー法モデル [1] を多クラス分類器として用いて前置詞誤り訂正システムを構築した．最大エントロピーモデルのツールとして Maximum Entropy Modeling Tookit [*8]をデフォルトパラメータで使用した．素性は [19], [6] で挙げられている単語の表層，品詞，WordNet，構文，言語モデル素性を用いた．品詞と構文素性の抽出は Stanford Parser 2.0.2 [*9]を用いた．このシステムは CLC FCE データセット [20] でトレーニング，テストを行ない，再現率：18.44，適合率：34.88，F 値：24.12 を達成し，HOO 2012 Shared Task [5] の前置詞誤り訂正タスクで 13 システム中 4 番目の成績であった．

テストデータとして KJ コーパスを使用した．KJ コーパスは 170 エッセイ，2,411 文からなる．KJ コーパスを使用した実験を行なうとき，5-fold cross validation を行ない，トレーニングデータとテストデータに分割して実験を行なった．

### 4.2 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 KJ コーパスにアノテートされた誤りタイプタグをもとに true positive, false positive, false negative を算出して計算した．そのため，KJ 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 [*10]．表 2 を使って評価の仕方を説明する．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ため，前置詞誤りの適合率に影響はしない．そのため，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．実際の誤り別の適合率はこの数字より少し下がるが，トレーニングコーパスを変えた場合でも両方とも同じように精度が下がるため，結果の優劣にはほとんど影響はない．

[*6] http://www.statmt.org/moses/
[*7] http://code.google.com/p/giza-pp/
[*8] https://github.com/lzhang10/maxent
[*9] http://nlp.stanford.edu/software/lex-parser.shtml
[*10] 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．

### 4.3 実験結果

表 3 に異なるコーパスでの各誤りタイプに対する誤り訂正結果を示す．KJ コーパスでトレーニングした統計的機械翻訳システムと Lang-8 コーパスを KJ コーパスとほぼ同サイズの 2,000 文を利用した場合と全てのデータ使用してトレーニングした場合とを比較した．F 値, 適合率, 再現率はコーパスサイズを大きくすることで高くなることが分かる．また，コーパスサイズを大きくすると，再現率よりも適合率のほうが上がる傾向があることがわかる．

表 4 に Lang-8 でコーパスサイズを 2K, 10K，20K，100K，200K，300K，全てのデータ (390K) と変化させた場合の各誤りに対する F 値の変化を示す．次の節で詳しく述べるが，学習者コーパスのサイズを変えたとき，誤りタイプを 2 つに分けることができる．

表 5 は前置詞誤り訂正の実験結果を示す．驚くことではないが，Lang-8 コーパスで学習した統計的機械翻訳のシステムが他の 2 つのシステムよりも明らかに性能が良いことは注目すべきことである．最大エントロピーモデルは同じ小さな規模のコーパスで学習した場合は統計的機械翻訳よりも良くなっている．最大エントロピーモデルの実装の都合上，計算量が膨大となったため，Lang-8 コーパスを用いた実験は行なわなかった．

### 4.4 考察

誤りは 2 つのタイプに分けることができ，(1) コーパスサイズが大きくなると訂正が良くなる誤りと (2) コーパスサイズとあまり関係がない誤りである．最初のタイプの誤りは冠詞，前置詞，名詞の語彙選択，動詞の語彙選択，形容詞，名詞その他である．一方，2 つ目のタイプの誤りは名詞の単複，動詞の時制，動詞の人称・数の不一致，副詞，接続詞，語順，助動詞，関係詞，疑問詞である．データが増えると精度が向上する誤りは（特に再現率，適合率両方上がっているもの），さらにデータを増やすことでフレーズベース統計的機械翻訳でも精度の向上を期待できる．一方で，データを増やしても精度が向上しないものに関してはフレーズベース統計的機械翻訳で訂正するのは難しい誤りと言える。コーパスのデータサイズを増やすことで，大きく F 値が向上している冠詞，名詞の語彙選択とデータサイズを増やしても F 値に差がほとんど見られない名詞の単複，動詞の時制，動詞の人称・数の不一致について実例を見ながら考察を行なう．

表 6 にコーパスサイズを大きくした場合に F 値が向上した冠詞と名詞の語彙選択の例を示す．この 2 つの例は両方とも，KJ コーパスでは訂正できなかったが，Lang-8 のコーパスを使うことによって訂正可能になった例である．コーパスサイズを大きくすることで，多くの誤りのフレーズとその訂正フレーズを得ることができるため，フレーズベース統計的機械翻訳で Lang-8 コーパスを使用することで訂

© 2012 Information Processing Society of Japan

**表 2** 評価方法を説明するための例

| | |
|---|---|
| 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. |

**表 3** 各誤りごとの統計的機械翻訳による誤り訂正の結果（再現率・適合率・F 値）．太字はあるシステムが他のシステムの結果より 0.1 ポイント以上高いものを表す．

| トレーニングコーパス | KJ コーパス | | | Lang-8 コーパス (2K) | | | Lang-8 コーパス (390K) | | |
|---|---|---|---|---|---|---|---|---|---|
| | 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 | 0.187 | 0.531 | 0.277 | 0.187 | 0.571 | 0.282 | **0.359** | **0.761** | **0.488** |
| 名詞の単複 | 0.207 | 0.603 | 0.308 | 0.136 | 0.671 | 0.226 | 0.199 | **0.710** | 0.311 |
| 前置詞 | 0.137 | 0.375 | 0.201 | 0.092 | 0.319 | 0.143 | **0.262** | **0.585** | **0.361** |
| 動詞の時制 | 0.102 | 0.170 | 0.128 | 0.043 | 0.088 | 0.058 | 0.080 | 0.149 | 0.104 |
| 名詞の語彙選択 | 0.035 | 0.114 | 0.054 | 0.033 | 0.152 | 0.054 | **0.182** | **0.443** | **0.258** |
| 動詞の語彙選択 | 0.070 | 0.161 | 0.098 | 0.065 | 0.200 | 0.098 | **0.192** | **0.324** | **0.241** |
| 代名詞 | 0.075 | 0.220 | 0.112 | 0.040 | 0.143 | 0.063 | 0.150 | **0.367** | **0.213** |
| 動詞の人称・数の不一致 | 0.236 | **0.604** | 0.340 | 0.125 | 0.483 | 0.199 | 0.228 | 0.469 | 0.307 |
| 形容詞 | 0.151 | 0.326 | 0.206 | 0.056 | 0.286 | 0.094 | **0.389** | **0.522** | **0.446** |
| 動詞その他 | 0.089 | 0.139 | 0.109 | 0.147 | 0.333 | 0.204 | **0.286** | **0.419** | **0.340** |
| 副詞 | 0.265 | 0.450 | 0.333 | 0.214 | 0.429 | 0.286 | 0.292 | 0.432 | 0.349 |
| 接続詞 | 0.100 | 0.417 | 0.161 | 0.091 | 0.714 | 0.161 | 0.115 | **0.546** | 0.190 |
| 語順 | 0.500 | 0.025 | 0.048 | 0.667 | 0.050 | 0.093 | **0.750** | 0.075 | 0.136 |
| 名詞その他 | 0.182 | 0.222 | 0.200 | 0.143 | 0.167 | 0.154 | **0.571** | **0.429** | **0.490** |
| 助動詞 | 0.056 | 0.167 | 0.083 | 0.100 | 0.400 | 0.160 | 0.100 | **0.400** | 0.160 |
| 語彙選択その他 | 0.167 | 0.200 | 0.182 | 0.000 | 0.000 | 0.000 | **0.357** | **0.455** | **0.400** |
| 関係詞 | 0.111 | 0.250 | 0.154 | 0.182 | 0.667 | 0.286 | 0.091 | **0.500** | 0.154 |
| 疑問詞 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| トータル | 0.149 | 0.147 | 0.148 | 0.113 | 0.205 | 0.146 | 0.247 | **0.275** | **0.260** |

正できるようになると考えられる．

表 7 はコーパスサイズを大きくしても F 値に差がなかった名詞の単複，動詞の時制，動詞の人称・数の不一致の例である．名詞の単複の 1 つ目の例は学習者に一般的な表現の 1 つであるため，フレーズベース統計的機械翻訳で Lang-8 コーパスを使って訂正することができた．一方，名詞の単複 2 つ目の例はフレーズベース統計的機械翻訳で Lang-8 コーパスを使って訂正できなかった例であり，これは “dools” と “a big” の間に固有名詞 “snoopy” が挿入されている．しかしながら，“doll” と “a” の間に少し距離があり，間に “snoopy” という固有名詞が入っているため訂正が困難であり，Brockett ら [2] の名詞の単複に関する人工データを使う手法でも訂正することは難しい．この問題を解くためには，品詞を使って一般化を行なったり，係り受け関係を見る必要がある．

動詞の時制の 1 つ目の例もフレーズベースの統計的機械翻訳で Lang-8 のコーパスを使って訂正できた例である．システムが訂正できた 1 つの理由は，ローカルの情報だけで訂正が可能で，かつ小規模の学習者コーパスにもよく出現するような一般的な誤りであったためである．2 つ目の例は，複雑な文でシステムが動詞の時制の一致を見つけられなかったものである．動詞の時制の誤りはフレーズベース統計的機械翻訳の手法で訂正することは難しく，Tajiri ら [18] が提案したような広い文脈を考慮できる手法が必要である．

動詞の一致の誤りの例の 1 つ目はフレーズベースの統計的機械翻訳で Lang-8 のコーパスを使って訂正できたものである．これは “Flowers is” を “Flowers are” に訂正するフレーズがよく出現し，言語モデルでも “Flowers are” の方が “Flowers is” よりも確率が高いためである．2 つ目の例は，学習者コーパスで出てこないパターンで，かつシステムが “reading” と “are” の係り受け関係をとらえることができず，システムが動詞の一致誤りを訂正できなかったものである．この問題を解くには，係り受け構造を考慮し，動詞の主語が何であるかを知る必要がある．KJ コーパス第 3 版では，主語と動詞が係っているかどうかの情報が付与されており，これを活用して誤り訂正を行なうことが考えられる．

© 2012 Information Processing Society of Japan

**表 4** トレーニングに使用する学習者コーパスのサイズを変化させた場合の統計的機械翻訳による誤り訂正の実験結果（F 値）．アスタリスクは Lang-8 コーパスを用いた場合と KJ コーパスを用いた場合の結果とで統計的有意差があることを表す ($p < 0.01$)．

| トレーニングコーパス | KJ コーパス | Lang-8 コーパス | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2K | 10K | 20K | 100K | 200K | 300K | 390K |
| 冠詞 | 0.277 | 0.282 | *0.390 | *0.420 | *0.443 | *0.459 | *0.475 | *0.488 |
| 名詞の単複 | 0.308 | 0.226 | 0.214 | 0.238 | 0.270 | 0.300 | 0.319 | 0.311 |
| 前置詞 | 0.201 | 0.143 | 0.192 | 0.226 | *0.333 | *0.336 | *0.344 | *0.362 |
| 動詞の時制 | 0.128 | 0.058 | 0.066 | 0.058 | 0.081 | 0.096 | 0.089 | 0.104 |
| 名詞の語彙選択 | 0.054 | 0.054 | 0.124 | 0.133 | *0.189 | *0.216 | *0.250 | *0.258 |
| 動詞の語彙選択 | 0.098 | 0.098 | 0.087 | 0.138 | *0.196 | *0.232 | *0.232 | *0.241 |
| 代名詞 | 0.112 | 0.063 | 0.131 | 0.150 | 0.177 | 0.195 | 0.213 | 0.213 |
| 動詞の人称・数の不一致 | 0.340 | 0.197 | 0.224 | 0.248 | 0.260 | 0.284 | 0.307 | 0.307 |
| 形容詞 | 0.206 | 0.094 | 0.165 | 0.219 | *0.413 | *0.426 | *0.426 | *0.446 |
| 動詞その他 | 0.109 | 0.204 | 0.240 | 0.311 | 0.291 | *0.340 | 0.308 | 0.340 |
| 副詞 | 0.333 | 0.286 | 0.286 | 0.302 | 0.333 | 0.349 | 0.349 | 0.349 |
| 接続詞 | 0.161 | 0.161 | 0.161 | 0.191 | 0.161 | 0.191 | 0.191 | 0.191 |
| 語順 | 0.048 | 0.093 | 0.093 | 0.091 | 0.091 | 0.091 | 0.091 | 0.136 |
| 名詞その他 | 0.200 | 0.154 | 0.286 | 0.286 | *0.531 | *0.490 | *0.490 | *0.490 |
| 助動詞 | 0.083 | 0.160 | 0.160 | 0.083 | 0.083 | 0.160 | 0.160 | 0.160 |
| 語彙選択その他 | 0.182 | 0.000 | 0.095 | 0.095 | 0.400 | 0.400 | 0.400 | 0.400 |
| 関係詞 | 0.154 | 0.285 | 0.154 | 0.154 | 0.154 | 0.154 | 0.154 | 0.154 |
| 疑問詞 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| トータル | 0.148 | 0.146 | 0.180 | 0.200 | 0.239 | 0.247 | 0.254 | 0.260 |

**表 5** 前置詞誤り訂正タスクでの比較実験の結果

| システム | トレーニングコーパス | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|
| 最大エントロピーモデル | KJ コーパス | 0.165 | 0.407 | 0.235 |
| フレーズベース統計的機械翻訳 | KJ コーパス | 0.137 | 0.375 | 0.201 |
| フレーズベース統計的機械翻訳 | Lang-8 コーパス (390K) | **0.262** | **0.585** | **0.362** |

前置詞誤り訂正に関しては，Lang-8 コーパス全てを用いた統計的機械翻訳が最大エントロピーモデルより良い性能を出したのには 2 つの理由があると考える．1 つ目は最大エントロピーモデルで用いたトレーニングデータが小規模な KJ コーパス（2,000 文）であったため，最大エントロピーモデルが高い性能を出すことができなかった点である．2 つ目は，統計的機械翻訳システムが高い性能を出す事ができたのは，KJ コーパスと Lang-8 コーパスの両方とも日本語ネイティブスピーカーによって書かれた作文であるため，データ量の増加が直接語彙選択など単語のカバー率が関係する誤りの性能向上につながったのだと考えられる．

また，同じ小規模のコーパスで学習した場合は，最大エントロピーモデルが統計的機械翻訳よりもよい性能であった．これは，KJ コーパスがとても小さくフレーズベース統計的機械翻訳では英語学習者の誤りのバリエーションを学習することができないが，識別モデルはリッチな素性を使って小さなデータからも学習することができるためであると考えられる．

## まとめと今後の課題

本稿では，フレーズ統計的機械翻訳の手法で大規模学習者コーパスを使って英語学習者の全てのタイプの誤りを対象とした文法誤り訂正を行なった．先行研究は小規模な学習者コーパスを使って限られたタイプの誤りだけを対象としていた．この問題を Web から抽出した誤りが付与された大規模なコーパスでトレーニングすることで解決した．

フレーズベース統計的機械翻訳のアプローチの性能改善にコーパスのサイズが重要であることがわかった．しかしながら，誤りのタイプによって改善の度合いが異なる．フレーズベース統計的機械翻訳はローカルの情報だけで訂正可能でよく出現する誤りに対して有効である．例えば，冠詞，前置詞，語彙選択，形容詞の誤りの訂正に大しては学習者コーパスのサイズを大きくすることが有効であり，一方，動詞の一致や動詞の時制といった誤りはコーパスのサ

© 2012 Information Processing Society of Japan

**表 6** 冠詞誤りと名詞の語彙選択誤りに対する例

| | 学習者 | 正解 |
|---|---|---|
| 冠詞 | I like a chocolate very much. | I like _ chocolate very much. |
| 名詞の語彙選択 | my cycle was injured, but i wasn't. | my bicycle was damaged, but i wasn't. |

**表 7** 名詞の単複，動詞の時制，動詞の一致誤りに対する例．アスタリスクは Lang-8 コーパス全てを用いた統計的機械翻訳システムで訂正できなかったことを表す．

| | learner | correct |
|---|---|---|
| 名詞の単複 1 | I read various type books. | I read various types of books. |
| *名詞の単複 2 | There is a big snoopy dools in my room. | There is a big snoopy doll in my room. |
| 動詞の時制 1 | If I 'll live in saitama, I must have ... | If I _ live in saitama, I must have ... |
| *動詞の時制 2 | The weather is very sunny, so we were ... | The weather was very sunny, so we were ... |
| 動詞の一致 1 | Flowers is very beautiful. | Flowers are very beautiful. |
| *動詞の一致 2 | I think, reading comics are not "reading" | I think, reading comics is not "reading" |

イズを大きくしてもあまり有効ではなかった．

今後の課題の 1 つ目としてはコーパスのサイズをさらに大きくすることが考えられる．本稿で 2010 年 12 月までの Lang-8 のデータしか使用していないが，2011 年，2012 年のデータを使用することができる．そうすることで，本稿で明らかにしたデータ量を増やすと性能が向上する誤り（冠詞，前置詞，語彙選択など）をさらに訂正することができるようになると考えられる．

2 つ目は品詞情報や構文情報などを用いることである．コーパスサイズを大きくしても性能が改善しない誤りの多くは，データがスパースなためであったり，遠くにある単語との関係を考慮して訂正する必要がある誤りである．そのため，品詞を用いて単語の一般化を行なったり，係り受けなどを用いて単語と単語の関係を捉える必要がある．

3 つ目の課題は学習者の母語の違いによる影響を調べることである．本稿で行なったのは日本人英語学習者の誤り訂正だけであり，トレーニングデータに関しても日本人が書いたコーパスを用意して実験を行なった．しかし，英語学習者は日本人だけではなく，その他の国でも多くの英語学習者が存在しており，学習者コーパスの中にもさまざまな母語の学習者の書いた作文がある．そのため，トレーニングデータやテストデータの学習者の母語を変えた場合，同様な結果が得られるかはわかっていないため調査が必要である．

4 つ目としては，学習者の書いた作文の誤りの種類，誤りの理由の推定がある．これまで行なわれてきた誤り訂正の研究では，誤り訂正だけが行なわれて誤りのタイプが何であるか，誤りの原因が何であるかという研究は行なわれていない．学習者の支援を行なうためには，誤りを訂正するだけではなく，誤りを訂正するとともに学習者への誤りに関する情報のフィードバックが重要であると考える．

## 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] Berger, A. L., Pietra, V. J. D. and Pietra, S. A. D.: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71 (1996).
[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] Dahlmeier, D. and Ng, H. T.: Grammatical Error Correction with Alternating Structure Optimization, *Proceedings of ACL-HLT*, pp. 915–923 (2011).
[4] Dahlmeier, D. and Ng, H. T.: A Beam-Search Decoder for Grammatical Error Correction, *Proceedings of EMNLP*, pp. 568–578 (2012).
[5] 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).
[6] De Felice, R. and Pulman, S. G.: A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English, *Proceedings of COLING*, pp. 169–176 (2008).
[7] Ehsan, N. and Faili, H.: Grammatical and Context-Sensitive Error Correction Using a Statistical Machine Translation Framework, *Software: Practice and Experience* (2012).
[8] Han, N.-R., Tetreault, J., Lee, S.-H. and Ha, J.-Y.: Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System, *Proceedings of LREC*, pp. 763–770 (2010).
[9] Koehn, P., Och, F. J. and Marcu, D.: Statistical Phrase-Based Translation, *Proceedings of HLT-NAACL*, pp. 48–54 (2003).
[10] Lee, J. and Seneff, S.: Correcting Misuse of Verb Forms, *Proceedings of ACL-HLT*, pp. 174–182 (2008).
[11] Liu, X., Han, B. and Zhou, M.: Correcting Verb Selection Errors for ESL with the Perceptron, *Proceedings of CICLing*, pp. 411–423 (2011).
[12] 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).
[13] Och, F. J. and Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation,

© 2012 Information Processing Society of Japan

*Proceedings of ACL*, pp. 295–302 (2002).

[14] Och, F. J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51 (2003).

[15] Park, Y. A. and Levy, R.: Automated Whole Sentence Grammar Correction Using a Noisy Channel Model, *Proceedings of ACL*, pp. 934–944 (2011).

[16] Rozovskaya, A. and Roth, D.: Algorithm Selection and Model Adaptation for ESL Correction Tasks, *Proceedings of ACL*, pp. 924–933 (2011).

[17] Swanson, B. and Yamangil, E.: Correction Detection and Error Type Selection as an ESL Educational Aid, *Proceedings of NAACL: HLT*, pp. 357–361 (2012).

[18] 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).

[19] Tetreault, J., Foster, J. and Chodorow, M.: Using Parse Features for Preposition Selection and Error Detection, *Proceedings of ACL*, pp. 353–358 (2010).

[20] Yannakoudakis, H., Briscoe, T. and Medlock, B.: A New Dataset and Method for Automatically Grading ESOL Texts, *Proceedings of ACL*, pp. 180–189 (2011).

© 2012 Information Processing Society of Japan