# 対訳コーパスへの擬似文法誤りの挿入による翻訳誤り訂正データの構築

大嶽匡俊 [1]　宮尾祐介 [1]
[1] 東京大学
{otake,yusuke}@is.s.u-tokyo.ac.jp

## 概要

本論文では、対訳コーパスに**擬似的な文法誤り**を挿入して**翻訳誤り訂正 (TEC)** のデータを生成する方法を検討し、このデータを TEC の事前学習に用いてその有効性を検証する。文法誤り訂正は文中の文法誤りを計算機で訂正するタスクであり、その一分野として翻訳を通じた第二言語学習を想定する TEC が提案されている。TEC の有効性は先行研究によってある程度示されたものの、TEC ではデータの不足が深刻である。そこで我々は先行研究の再現を試みると共に、擬似的な文法誤りを対訳コーパスに加えることで TEC の擬似データを大幅に拡張した。また、これを事前学習に用いる場合と用いない場合を比較することで有効性を確認した。

## 1　はじめに

計算機が文章を訂正する文法誤り訂正 (Grammatical Error Correction; GEC) は急速に成長しており、Grammarly[1)] など実社会で使われるようにもなった。しかし、英作文の勉強をしている学生らは未だ計算機の助けを十分に借りられていない。この要因の一つとして、現在の誤り訂正システムが意味の誤りや複雑な誤りを訂正できないこと考えられる。例えば Cao ら [1] は、「I am leaving in Tokyo.」という文章を訂正する際に、leaving を living に直すか、in を for に直すかという曖昧性があり、誤りがこのように複合的であると現在の誤り訂正システムでは訂正が難しいことを指摘している。そこで、日本の学生たちが「和文英訳」と呼ばれる日本語を英語に直す翻訳を通して英語を学んでいることに着目する。この時、誤り訂正システムは元の和文を参照することでより良い訂正を行える可能性がある。以下、本論文では翻訳の誤りを訂正するこのタスクを「**翻訳誤り訂正 (Translation Error Correction; TEC)**」と呼ぶ。

本論文では、TEC の擬似データを生成しその有効性を検証する。学習者への TEC は Cao ら [1] 以降、研究が十分にされていない。その一因が、データ量の不足である。機械学習を用いてこの問題を解く場合十分な学習データが必要だが、日本語文 (母国語文)、誤り文、訂正文の 3 つ組全てを含むデータはほとんど存在しない。Cao ら [1] は学習者コーパスと呼ばれる GEC 用のコーパスから機械翻訳を用いて擬似的な TEC 用のデータセットを構築したが、これは機械学習を用いた自然言語処理の問題設定としては量が不十分である。そこで我々は、翻訳タスクのデータセットである対訳コーパスと擬似的に文法誤りを生成する手法を組み合わせ、新たな TEC の擬似データを生成しその効果を検証した。その結果、擬似データを用いることで安定的に精度が向上することや、GEC で効果的だとされている擬似誤り文の生成方法 [2] が TEC でも有効に働くことが確かめられた。また、Cao らの提案手法を GEC で一般的に使用されているデータセットに適用し、日本語文を誤り訂正に用いることの有効性を再度確認した。

## 2　関連研究

本研究では対訳コーパスに擬似的な文法誤りを挿入して生成した TEC のデータの有効性を検証する。本研究に必要なモデル、誤り文、データの生成手法の研究を本節で説明する。

### 2.1　Cao らの先行研究

1 節で取り上げた通り、Cao ら [1] の研究が本研究の直接的な先行研究である。TEC では日本語文、英語の誤り文、それを訂正した英語の訂正文の 3 つのデータが必要だが、Cao らは学習者コーパスの訂正された英語を機械翻訳にかけることによって日本語文を生成し、TEC のデータセットとした。生成元の

1) https://grammarly.com/

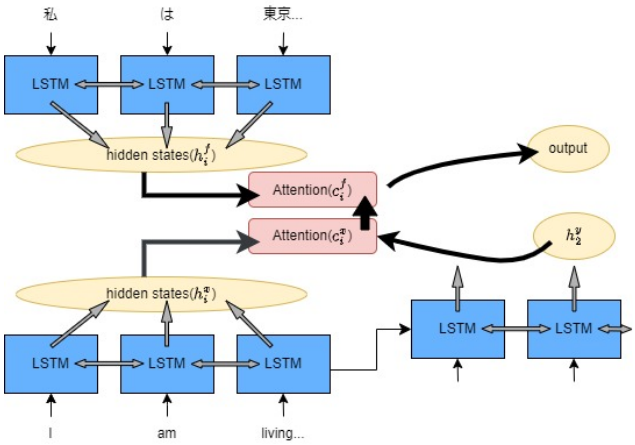


**図 1**　Cao ら [1] が提案した TEC モデル

学習者コーパスには Lang-8 [3] が用いられ、これを訓練、開発、評価データに分割して用いた。

また、Cao らは日本語文と誤り文の二つの情報を統合するモデルを提案し、誤り訂正における日本語文の有効性を評価した。評価スコアには BLEU スコアを GEC 用に改変した GLEU スコア [4] が用いられ、日本語文を用いると誤り文だけを用いた時より GLEU スコアが高くなることが示された。

**マルチ階層注意機構モデル**　Cao ら [1] はマルチ階層注意機構モデルと呼ばれるモデルを TEC 用に提案した（図 1）。このモデルでは、それぞれの入力文に対し双方向 LSTM を用いて隠れ状態を計算した上で、decoder の隠れ状態と英語文間で注意機構を適用し文脈ベクトル $c_i^x$ を計算する。その文脈ベクトル $c_i^x$ と decoder の隠れ状態の和 $h_i^y$ と日本語文間に再び注意機構を適用し日本語文からの文脈ベクトル $c_i^f$ を計算する。最後に、計算されたそれぞれの文脈ベクトルと decoder の隠れ状態から出力する単語を計算する。

## 2.2　その他の先行研究

機械翻訳の出力を人間が編集する post-editing の流れを汲んで、プロの翻訳家の誤りを訂正するタスクとしても TEC が研究されている [5]。また、菊池ら [6] は誤り訂正ではなく分類問題として問題を設計し研究を行っている。

## 2.3　擬似文法誤りの生成手法

文法誤りを機械的に生成しデータを拡張しようという試みは、GEC の一つのタスクとして盛んに研究されてきた。

**逆翻訳を用いた誤り文生成**　GEC における逆翻訳とは、学習者コーパスを用いて訂正文を入力、誤り文を出力とするモデルを構築し、単言語コーパスに適用して擬似誤りデータを生成する手法である。Xie ら [7] はノイズ付き逆翻訳と呼ばれるビームサーチ中にノイズを加える手法を提案した。

**ルールベースの誤り文生成**　定められたアルゴリズムに基づき誤り文を生成する手法である。もっとも一般的なものは文字や単語を確率的に削除、挿入、置換などするシンプルな手法で、GEC の Shared Task の優勝システムにも用いられた [8]。また、近い意味を持つ別の単語に置き換える試みや、品詞タグ付けを利用して品詞ごとにルールを定める等、より多様なルールを定めた手法も用いられている [9]。

## 2.4　データセット

本研究の実験には、対訳コーパスから生成したデータセットと、学習者コーパスから生成したデータセットの二つを作成し用いる。この二つのデータセットの生成元となる対訳コーパスと学習者コーパスについて本小節で説明する。

### 2.4.1　日英対訳コーパス

翻訳タスクに用いられる対訳コーパスはこれまで多数構築されてきたが、本研究では Reuters Corpora [10] と JparaCrawl [11] を用いる。GEC の研究からの知見として清野ら [2] が Wikipedia から作られたコーパスと新聞記事から作られたコーパスを比較し、後者の方が性能が安定することを確認している。Reuters Corpora はニュースコーパスで、新聞記事から作られたコーパスと似た性質を持つと考えられる。JparaCrawl はウェブをクロールして作られた大規模日英対訳コーパスである。しかし、文章の質は Reuters Corpora より悪いことが想定される。

### 2.4.2　学習者コーパス

BEA-2019 [12] は GEC の Shared Task として開催され、これの Restricted Track ではいくつかの訓練用学習者コーパスが指定されており、現在の GEC において標準的な訓練データとなっている。BEA-2019 の Restricted Track で指定されている訓練用コーパスは Lang-8 [3]、NUCLE [13]、FCE [14]、W&I+LOCNESS [12] [15] の 4 つである。また、BEA-2019 が開催されるまでは CoNLL-2014 [16] の評価データが GEC で最も標準的な評価データであった。

# 3 提案手法

## 3.1 擬似文法誤り TEC データの評価

対訳コーパスに擬似的な文法誤りを挿入する事で擬似的な TEC データを生成し、これを事前学習に用いて生成したデータの有効性を評価する。以下、対訳コーパスと擬似的な文法誤りから生成された TEC データを「擬似文法誤り TEC データ」、学習者コーパスと機械翻訳を用いて生成された TEC データを「擬似母国語 TEC データ」と呼ぶ。

前節で挙げた Reuters Corpora [10] と JparaCrawl [11] の 2 種類の対訳コーパスと、逆翻訳 [7]、シンプルなルール [2]、多様なルール [9] の 3 つの誤り文生成手法から、6 パターンの擬似文法誤り TEC データを生成する。そして、これらのデータで事前学習をした後、擬似母国語 TEC データでファインチューニングを行い TEC モデルを作成する。また、擬似母国語 TEC データのみで学習したモデルを用意し、事前学習をしたモデルと事前学習をしていないモデルで開発データ、評価データの GLEU スコア [4] を比較し事前学習の有効性を検討する。モデルには Cao ら [1] のマルチ階層注意機構モデルを用いる。

## 3.2 先行研究の別データでの再現

我々は GEC で一般的に用いられるデータセットを擬似母国語 TEC データの作成元として、Cao ら [1] の実験の再現を行う。Cao らは Lang-8 [3] から擬似母国語 TEC データを作成し TEC の学習データや評価データとして用いた。しかしながら、Lang-8 は質の面で問題があり他のデータセットでの検証が必要だと考えられる。我々は、BEA-2019 [12] の Restricted Track で与えられたデータから擬似母国語 TEC データとして学習データと開発データを生成し利用する。また、擬似母国語 TEC データを作成するには評価データの訂正文が公開されている必要がある。そこで、評価データの訂正文の公開がない BEA-2019 の代わりに CoNLL-2014 [16] の評価データセットを評価データの生成元として用いる。

# 4 実験

我々は「擬似文法誤り TEC データを作成、擬似文法誤り TEC データで事前学習、擬似母国語 TEC データでファインチューニング、疑似母国語 TEC データで評価」という一連の実験を試みる。

**表 1** 学習に用いるデータセットの文対数

| Dataset | 文対数 |
|---|---|
| Lang-8 [3] | 1,037,561 |
| NUCLE [13] | 57,151 |
| FCE [14] | 33,236 |
| W&I+LOCNESS [12] [15] | 34,304 |

また、英語のみの入力で GEC として問題を解く場合と日本語のみの入力で翻訳として問題を解く場合についても実験を行う。これをその二つを入力とする TEC と比較し、Cao らが示した TEC タスクの有効性を再検証する。

## 4.1 利用するコーパスの詳細

擬似文法誤り TEC データの生成元となる対訳コーパスと、擬似母国語 TEC データの生成元となる学習者コーパスについて本小節で述べる。

**対訳コーパス** JparaCrawl は非常に膨大でかつ質は低いデータセットであるため、bleualign スコア [17] が 0.75 以上の文章のみを抜粋して用いた。Reuters Corpora は 56,782 文対、JparaCrawl の抜粋データは 8,300,634 文対である。

**学習者コーパス** BEA-2019 [12] の Restricted Track で指定された学習者コーパスを本研究の学習データの生成元として用いる。各コーパスで学習に用いる文対数を表 1 にまとめる。

本研究の開発データには W&I+LOCNESS [12] [15] の開発データ 4,380 文対を用い、評価データには CoNLL-2014[16] の評価データ 1,312 文対を用いる。CoNLL-2014 の評価データの訂正文には初めに作られたデータ (test) と参加者によって修正されたデータ (test-withalt) があり、両方を用いた。

## 4.2 擬似文法誤り文の生成

本研究では 3 つの手法を用いて擬似文法誤り文を生成した。

**逆翻訳** 逆翻訳モデルには Transformer [18] を用いた。ビームサーチへのノイズ付加 [7] を行い、パラメータ $\beta$ には 8.0 を用いた。

**シンプルなルール** 清野ら [2] の実装の通り、各単語に対して ‘mask’, ‘deletion’, ‘insertion’, ‘keep’ を確率的に選択しノイズを加えた。

**多様なルール** 古山ら [9] の erarigilo[2] と reguligilo[3] を用いた。

2) https://github.com/nymwa/erarigilo
3) https://github.com/shotakoyama/reguligilo

**表 2**　事前学習あり、なしのモデルの GLEU スコア

| 手法 | 開発 | test-withalt | test |
|---|---|---|---|
| 事前学習なし | 64.5 | 39.5 | **42.8** |
| JparaCrawl (逆翻訳) | **66.6** | **42.3** | 42.7 |
| JparaCrawl (シンプル) | 66.0 | 41.3 | 41.7 |
| JparaCrawl (多様) | 66.5 | 41.8 | 41.9 |
| Reuters Corpora (逆翻訳) | 65.1 | 40.2 | 40.5 |
| Reuters Corpora (シンプル) | 64.5 | 40.1 | 40.1 |
| Reuters Corpora (多様) | 65.0 | 40.2 | 40.6 |

### 4.3　評価

Cao ら [1] のマルチ階層注意機構モデルに擬似文法誤り TEC データを事前学習させ、擬似母国語 TEC データでファインチューニングしモデルの GLEU スコアを比較する。事前学習では JparaCrawl から抜粋したものは 3 エポック、Reuters Corpora のものは 20 エポック学習する。

また、事前学習を行わずに入力に英語文のみ、日本語文のみを用いた時の GLEU スコアを算出する。これによって日本語文を用いることの有効性と日本語文からの答えの深刻なリークが無いかを確認し先行研究の再現性を確認する。

## 5　実験結果と分析

### 5.1　事前学習の評価

本研究で生成した擬似文法誤り TEC データセットで事前学習し擬似母国語 TEC データでファインチューニングしたモデルと、事前学習なしで擬似母国語 TEC データでチューニングしたモデルを GLEU スコアで比較する。スコアの比較は評価データ、開発データで行った。結果を表 2 に示す。

JparaCrawl から生成したデータで事前学習すると事前学習なしの場合より安定的にスコアが高くなることが示され、事前学習の有効性が確認された。一方で、Reuters Corpora から事前学習したものと事前学習がないものを比較すると結果にばらつきがある。これは、Reuters Corpora のデータ量に不足があるため過学習が起こり、ファインチューニング間に事前学習データへの過学習から脱せなかった可能性がある。また、誤り文生成手法について比較すると逆翻訳、多様なルール、シンプルなルールの順に精度が高く、特に逆翻訳が高いスコアであった。これは GEC での結果 [2] と整合性がある。

**表 3**　入力文を変えた時の GLEU スコア
訂正なしとは、誤り訂正のモデルにかけず誤り文のままで GLEU スコアを計算した結果である。

| 入力 | 開発 | test-withalt | test |
|---|---|---|---|
| 訂正なし | 63.9 | 38.1 | 41.2 |
| 誤り文のみ | 63.5 | 39.1 | 42.0 |
| 日本語文のみ | 11.0 | 8.93 | 9.15 |
| 誤り文+日本語文 | **64.5** | **39.5** | **42.8** |

### 5.2　先行研究との比較

Cao ら [1] は日本語文が GEC に有効に働くことを Lang-8 [3] を用いて検証したが、我々は GEC において標準的に用いられる BEA-2019 のデータセットと CoNLL-2014 の評価データでも日本語文が有効に働くか再度検証した。この検証には生成した擬似文法誤り TEC データは用いず、擬似母国語 TEC データのみを用いる。結果を表 3 に示す。誤り文だけを用いたものと誤り文と日本語文の二つを用いるものを比較すると、後者の方が安定的に性能が良いことが示された。日本語文のみから学習した場合のスコアは非常に低く、訂正文の深刻なリークは発生していないと推測される。Cao らの先行研究では二つの入力を用いる場合と英語のみを用いる場合で GLEU スコアが 1.13 ポイント変化しており、本研究とおおむね対応する。

## 6　おわりに

本研究では，対訳コーパスへ擬似的な文法誤りを挿入することで TEC の新しい擬似データを生成し、その有効性を検証した。その結果、この擬似データで事前学習をしたモデルの GLEU スコアは事前学習がないモデルのスコアを上回り、TEC のデータ不足を補い得ることが分かった。また、データが大規模であることや、擬似誤り文の生成手法として逆翻訳を用いることの優位性が示唆された。さらに、我々は Cao らの手法を GEC で一般的に用いられるデータセットである BEA-2019 や CoNLL-2014 に適用し、日本語文を誤り訂正に用いることの有効性を再度検証した。一方、TEC でどのようなモデルを使うべきかより検討されるべきであるように思われる。また、Cao らや我々が生成した TEC 用のデータは擬似データであり、真のデータセットでの検証も必要である。TEC の研究がこれからより広く行われることを期待したい。

## 謝辞

本研究の実装では、曹国林氏、高村大也先生、奥村学先生には多大な協力を頂きました。古山翔太氏には氏の実装をご説明頂きました。感謝致します。

## 参考文献

[1] Cao GUOLIN, Hiroya TAKAMURA, and Manabu OKUMURA. Multi-source neural grammatical error correction. **Proceedings of the Annual Conference of JSAI**, Vol. JSAI2018, pp. 4Pin123–4Pin123, 2018.

[2] Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 1236–1242, Hong Kong, China, November 2019. Association for Computational Linguistics.

[3] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In **Proc. of 5th International Joint Conference on Natural Language Processing**, pp. 147–155, 2011.

[4] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Ground truth for grammatical error correction metrics. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 588–593, Beijing, China, July 2015. Association for Computational Linguistics.

[5] Jessy Lin, Geza Kovacs, Aditya Shastry, Joern Wuebker, and John DeNero. Automatic correction of human translations. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 494–507, Seattle, United States, July 2022. Association for Computational Linguistics.

[6] Seiya Kikuchi, Taisuke Onaka, Hiroaki Funayama, Yuichiroh Matsubayashi, and Inui Kentaro. 項目採点技術に基づいた和文英訳答案の自動採点. 言語処理学会, 第 27 回年次大会, pp. 690–695, 2021.

[7] Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. Noising and denoising natural language: Diverse backtranslation for grammar correction. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 619–628, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 252–263, Florence, Italy, August 2019. Association for Computational Linguistics.

[9] Shota Koyama, Hiroya Takamura, and Naoaki Okazaki. Various errors improve neural grammatical error correction. In **Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation**, pp. 251–261, Shanghai, China, 11 2021. Association for Computational Lingustics.

[10] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In **Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics**, pp. 72–79, Sapporo, Japan, July 2003. Association for Computational Linguistics.

[11] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[12] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 52–75, Florence, Italy, August 2019. Association for Computational Linguistics.

[13] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner English: The NUS corpus of learner English. In **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 22–31, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[14] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[15] Sylviane Granger. The computer learner corpus: a versatile new source of data for sla research. In **Learner English on computer**, pp. 3–18. Routledge, 2014.

[16] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In **Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task**, pp. 1–14, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[17] Rico Sennrich and Martin Volk. Iterative, mt-based sentence alignment of parallel texts. In **Nordic Conference of Computational Linguistics**, 2011.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.