学修番号 19860642

修士論文

文法誤り訂正の参照文を用いない自動評価の人手評価への最適化

吉村　綾馬

2021 年 2 月 19 日

東京都立大学大学院
システムデザイン研究科 情報科学域

吉村　綾馬

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　（副指導教員）

高間 康史 教授　（副指導教員）

文法誤り訂正の参照文を用いない自動評価の人手評価への最適化*

吉村　綾馬

修論要旨

文法誤り訂正は，主に言語学習者の書いた文法的に誤っている文（入力文）を文法的に正しい文（訂正文）に編集するタスクである．自動評価はコストをかけずにシステムを定量評価できるため，信頼できる自動評価手法の構築は研究および開発の発展に有用である．自動評価は訂正システムの出力文を入力文や人手で訂正した文（参照文）などを用いて評価を行う．訂正の仕方は一つではなく複数の訂正が考えられるため，自動評価は難しいタスクである．文法誤り訂正の自動評価は，参照文を用いる手法と用いない手法に大別できる．前者は，可能な参照文を網羅することが難しいため，参照文に含まれない表現に対してはそれが適切な訂正であっても不当に低い評価を与えるという問題がある．後者にはこの問題がなく，特に浅野ら（2018）は文法性・流暢性・意味保存性の各自動評価モデルの評価を統合することで参照文を用いる自動評価手法よりも人手評価との高い相関を達成した．文法性は訂正文が文法的に正しいかという観点である．流暢性は訂正文が母語話者にとってどの程度自然な文かという観点であり，文法性と区別されて重要性が示されている．意味保存性は入力文と訂正文がどの程度意味が同じであるかという観点であり，文法的な文でも入力文と異なる意味になる訂正は不適切なため重要な観点である．このように 3 項目で評価を行うことは，自動評価の解釈性を高めることができるため重要である．しかし，これらの各自動評価モデルは訂正文に対する各項目の人手評価に対してそれぞれ最適化されておらず，改善の余地が残されている．

そこで本研究では，人手評価との相関が高く，多様な訂正を正しく評価できる自動評価手法を構築するために，浅野ら（2018）の拡張として，文法性・流暢性・意味保存性の各自動評価モデルを各項目の人手評価に対して直接最適化する手法を

*東京都立大学大学院 システムデザイン研究科 情報科学域 修士論文, 学修番号 19860642, 2021 年 2 月 19 日.

提案する．具体的には，各項目の評価モデルとして，少量のデータで目的タスクに最適化できる事前学習された文符号化器 Bidirectional Encoder Representations from Transformers (BERT) を用い，各項目の人手評価値付きデータセットで再学習を行うことで各評価モデルの最適化を行う．また，学習者が書いた文や，機械翻訳の逆翻訳による擬似誤り文に対して文法性や流暢性をつけた既存のデータは存在するが，文法誤り訂正の自動評価の理想的な設定である，訂正文に対する各項目の人手評価値付きデータセットは存在しない．そのため，我々はクラウドソーシングを用いて，代表的な 5 種類の文法誤り訂正システムの訂正文に対して文法性・流暢性・意味保存性の人手評価を付与し，データセットの作成を行う．実験では，人手評価との相関および Methodology for Automatic Evaluation of GEC Evaluation (MAEGE) によって自動評価手法をメタ評価する．実験の結果，両方のメタ評価において我々の自動評価手法が従来の自動評価手法よりも適切な評価ができることを示した．また，各項目に対応する既存のデータセットを用いて訓練した自動評価モデルとの比較から，システムの訂正文に対する人手評価を用いて BERT を再学習することの有効性が明らかになった．分析の結果，参照文を用いない手法が多くのエラータイプの訂正に対して減点しているのに対して，提案手法は全てのエラータイプの訂正に対して加点していることがわかった．

本研究の主な貢献は以下の 4 つである．

- 文法誤り訂正の自動評価において，事前学習された文符号化器を用いて人手評価に直接最適化する手法を提案した．
- 文法誤り訂正における自動評価手法の学習のための，訂正システムの訂正文に対して文法性・流暢性・意味保存性の 3 項目の評価値を付与したデータセットを作成した．
- 人手評価との相関に基づくメタ評価および MAEGE によるメタ評価の結果，提案手法は既存手法よりも適切な評価が行えていることを示した．
- 分析の結果，従来手法に比べて，提案手法は調査可能な全てのエラータイプの訂正に対して加点できていることを示した．

本論文の構成は次の通りである．第 1 章では本研究の背景，提案，貢献について述べる．第 2 章では文法誤り訂正の自動評価手法の関連研究および既存のデータ

セットについて述べ，自動評価手法の評価について説明する．第 3 章では提案手法である，事前学習された文符号化器を用いた自動評価手法について述べる．第 4 章では訂正文に対する人手評価値付きデータセットの構築について述べる．第 5 章では提案手法の評価実験の設定について述べる．第 6 章では実験結果およびその考察について述べる．第 7 章で評価事例およびエラータイプ別評価の分析を行う．最後に第 8 章で本研究のまとめを述べる．

Optimization of Reference-less Evaluation Metric of Grammatical Error Correction for Manual Evaluations*

Ryoma Yoshimura

Abstract

Grammatical error correction (GEC) is a task of editing a grammatically incorrect sentence written by learner into a grammatically correct sentence. The construction of a reliable automatic evaluation metric is useful for the developmental cycles. In automatic evaluation, the output sentence of the GEC system is evaluated using an input sentence and a reference corrected manually. Automatic evaluation is a difficult task because there can be multiple corrections. Automatic evaluation metric of GEC can be roughly divided into methods that use reference-based metric and reference-less metric. The reference-based metrics are commonly used for automatic evaluation in the GEC task. However, these metrics penalize sentences whose words or phrases are not included in the reference, even if they are correct expressions because it is difficult to cover all possible references. In contrast, reference-less metrics do not suffer from this limitation. Among them, Asano et al. (2018) achieved a higher correlation with manual evaluations than reference-based metrics by integrating sub-metrics from the three perspectives of grammaticality, fluency, and meaning preservation. Grammaticality is the perspective of whether the corrected sentence is grammatically correct. Fluency is perspective of how natural the corrected sentence is to the native speaker and is distinguished from grammaticality in terms of its importance. Meaning preservation is the perspective of

*Master's Thesis, Department of Computer Science, Graduate School of System Design, Tokyo Metropolitan University, Student ID 19860642, February 19, 2021.

how well the meaning is preserved between the input sentence and the corrected sentence. This is an important perspective because it is inappropriate to correct a grammatical sentence that has a different meaning from the input sentence. Evaluating with these three items is important because it can improve the interpretability of automatic evaluation. However, each of these metrics has not been optimized for manual evaluation of each perspective; thus there is still room for improvement.

Therefore, to construct an automatic evaluation metric that is highly correlated with the manual evaluation and can correctly evaluate various corrections, we propose a method that directly optimizes each metric of grammaticality, fluency, and meaning preservation against the manual evaluation of each perspective as an extension of Asano et al. (2018). Specifically, we use Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language model that can be optimized for the target task with a small amount of data, as the metics for each perspective, and optimize each metic by fine-tuning it using a dataset with manual evaluation. Although there are existing datasets with grammaticality and fluency manual evaluation for learner-written sentences or round-trip translation from machine translation, there is no dataset with manual evaluation for each perspective in the corrected sentence, which is a ideal setting for automatic evaluation of GEC. Therefore, we use crowdsourcing to create a dataset of grammaticality, fluency, and meaning preservation of the corrected sentences of five typical GEC systems.

In the experiments, we meta-evaluate the proposed metric by evaluating a correlation with manual evaluation and using the Methodology for Automatic Evaluation of GEC Evaluation (MAEGE). The experimental results show that the our proposed metric improves the correlation with manual evaluation in both system- and sentence-level meta-evaluation. The experimental results show that the our proposed metric can evaluate more appropriate than the conventional automatic evaluation method in both meta-evaluations. The comparison with the metric trained on the existing dataset and trained on our

created dataset shows the effectiveness of fine-tuning BERT using manual evaluation of the corrected sentences. The analysis shows that the proposed metric rewards corrections of all error types, while the reference-based metrics penalize many error types of corrections.

The main contributions of our paper are as follows.

- For automatic evaluation of GEC, we proposed a metric that directly optimizes for manual evaluation using a pre-trained language model.
- We created a dataset of corrected sentences of GEC system with three manual evaluations: grammaticality, fluency, and meaning preservation.
- The results of meta-evaluation based on the correlation with manual evaluation and meta-evaluation by MAEGE show that the proposed metric can evaluate more appropriately than the existing metrics.
- The analysis results show that the proposed metric rewards the correction of all possible error types better than the existing methods.

This thesis comprises as follows. In Chapter 1, we describe the background, proposal, and contribution of this research. In Chapter 2, we explain related GEC metrics and existing datasets, and the meta-evaluation of the automatic evaluation metric. In Chapter 3, we propose a metric using a pre-trained language model. In Chapter 4, we construct a dataset with manual evaluation values for corrected sentences. In Chapter 5, we setup the experiment settings for evaluating the proposed method. In Chapter 6, we describe the experimental results and their discussion. In Chapter 7, we perform case analysis and the evaluation by error type. Finally, in Chapter 8, we summarize this thesis.

目次

図目次

第 1 章　はじめに

文法誤り訂正は，主に言語学習者の書いた文法的に誤っている文（入力文）を文法的に正しい文（訂正文）に編集するタスクである．自動評価はコストをかけずにシステムを定量評価できるため，信頼できる自動評価手法の構築は研究および開発の発展に有用である．文法誤り訂正の自動評価は図 1.1 のように訂正システムの出力文を入力文や人手で訂正した文（参照文）などを用いて訂正文の評価を行う．訂正の仕方は一つではなく複数の訂正が考えられるため，自動評価は難しいタスクである．文法誤り訂正の自動評価は，参照文を用いる手法 [1, 2] と用いない手法 [3, 4] に大別できる．前者は，可能な参照文を網羅することが難しい [5] ため，参照文に含まれない表現に対してはそれが適切な訂正であっても不当に低い評価を与えるという問題がある．後者にはこの問題がなく，特に浅野ら [4] は文法性・流暢性・意味保存性の各自動評価モデルの評価を統合することで参照文を用いる自動評価手法よりも人手評価との高い相関を達成した．文法性は訂正文が文法的に正しいかという観点である．流暢性は訂正文が母語話者にとってどの程度自然な文かという観点であり，文法性と区別されて重要性が示されている [6]．意味保存性は入力文と訂正文がどの程度意味が同じであるかという観点であり，文法的な文でも入力文と異なる意味になる訂正は不適切なため重要な観点である．このように 3 項目で評価を行うことは，自動評価の解釈性を高めることができるため重要である．しかし，これらの各自動評価モデルは訂正文に対する各項目の人手評価に対してそれぞれ最適化されておらず，改善の余地がある．

本研究では，人手評価との相関が高く，多様な訂正を正しく評価できる自動評

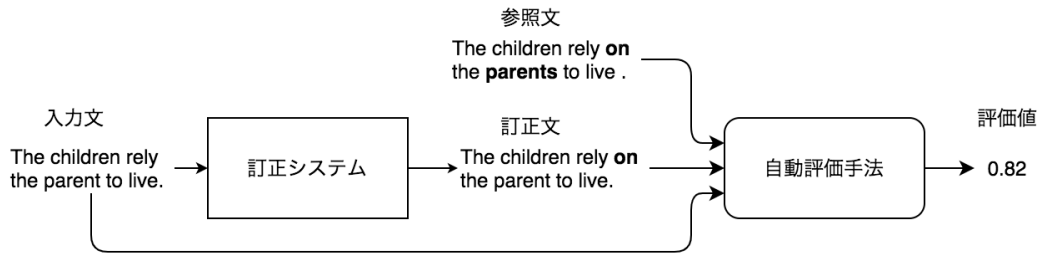

図 1.1: 文法誤り訂正の自動評価

価手法を構築するために，浅野らの拡張として，文法性・流暢性・意味保存性の各自動評価モデルを各項目の人手評価に対して直接最適化する手法を提案する．具体的には，各項目の評価モデルとして，少量のデータで目的タスクに最適化できる事前学習された文符号化器 Bidirectional Encoder Representations from Transformers (BERT) [7] を用い，各項目の人手評価値付きデータセットで再学習を行うことで各評価モデルの最適化を行う．また，学習者が書いた文や，機械翻訳の逆翻訳による擬似誤り文に対して文法性や流暢性をつけた既存のデータは存在するが，文法誤り訂正の自動評価の理想的な設定である，訂正文に対する各項目の人手評価値付きデータセットは存在しない．そのため，我々はクラウドソーシングを用いて，代表的な 5 種類の文法誤り訂正システムの訂正文に対して文法性・流暢性・意味保存性の人手評価を付与し，データセットの作成を行う．

実験では，人手評価との相関および MAEGE [8] によって，自動評価手法をメタ評価する．実験の結果，両方のメタ評価において我々の自動評価手法が従来の自動評価手法よりも適切な評価ができることを示した．また，各項目に対応する既存のデータセットを用いて訓練した自動評価モデルとの比較から，システムの訂正文に対する人手評価を用いて BERT を再学習することの有効性が明らかになった．分析の結果，参照文を用いない手法が多くのエラータイプの訂正に対して減点しているのに対して，提案手法は全てのエラータイプの訂正に対して加点していることがわかった．

本研究の主な貢献は以下の 4 つである．

- 文法誤り訂正の自動評価において，事前学習された文符号化器を用いて人手評価に直接最適化する手法を提案した．
- 文法誤り訂正における自動評価手法の学習のための，訂正システムの訂正文に対して文法性・流暢性・意味保存性の 3 項目の評価値を付与したデータセットを作成した．
- 人手評価との相関に基づくメタ評価および MAEGE によるメタ評価の結果，提案手法は既存手法よりも適切な評価が行えていることを示した．
- 分析の結果，従来手法に比べて，提案手法は調査可能な全てのエラータイプの訂正に対して加点できていることを示した．

本論文の構成は次の通りである．第 1 章では本研究の背景，提案，貢献について述べる．第 2 章では文法誤り訂正の自動評価手法の関連研究および既存のデータセットについて述べ，自動評価手法の評価について説明する．第 3 章では提案手法である，事前学習された文符号化器を用いた自動評価手法について述べる．第 4 章では訂正文に対する人手評価値付きデータセットの構築について述べる．第 5 章では提案手法の評価実験の設定について述べる．第 6 章では実験結果およびその考察について述べる．第 7 章で評価事例およびエラータイプ別評価の分析を行う．最後に第 8 章で本研究のまとめを述べる．

第 2 章　関連研究

2.1　文法誤り訂正の自動評価手法

2.1.1　参照文を用いる手法

初期の研究 [9] では，訂正した単語単位の編集に対して適合率・再現率・F 値を評価していた．Dahlmeie らはフレーズ単位で，適合率，再現率および，適合率を重視する F 値 ($F_{0.5}$) の 3 つの値を評価する Max Match (M^2) を提案し，より正確な自動評価が可能となった [1]．Felice らは悪い訂正と訂正を行わない場合の評価値がどちらも 0 になるなどの M^2 の問題点に対処する I-measure を提案した [10]．I-measure は単語レベルのアライメントに基づく重み付き精度によって計算され，入力文が悪化すると負の値，改善すると正の値となる．Napoles らは機械翻訳の自動評価で使われる BLEU [11] を文法誤り訂正のために改善した GLEU を提案した [2]．BLEU は訂正文と参照文を用いるが，GLEU は訂正文と参照文に加えて入力文も考慮する．Napoles らの調査によると，参照文を用いる手法の中では，GLEU が人手評価との最高の相関を持つ [3]．これらの評価指標は参照文を必要とするため，参照文が全ての正しい文を網羅できない問題に対処できない．しかし，可能な参照文を網羅することは現実的ではない [5] ため，近年は参照文を用いない自動評価手法が注目されている．

2.1.2　参照文を用いない手法

Napoles らは参照文を用いない文法誤り訂正の自動評価手法を初めて提案した [3]．文法誤り検出ツールに基づく手法と言語モデルなどの言語学的素性に基づく手法が提案され，前者が GLEU と同等の性能を持つことが示された．既存の文法誤り検出ツールである e-rater® や language-tool を用いた手法と，Heilman らのミススペリング数や言語モデルの評価値や未知語の数などの言語学的素性を用いた手法 [12] で実験を行い，e-rater® を用いた手法が GLEU とほぼ同等の性能であることを示した．浅野らは文法性・流暢性・意味保存性の 3 つの自動評価モデルに基づく参照文を用いない自動評価手法を提案し，人手評価との最高の相関を達成した [4]．

文法性は GUG データセット [12] で訓練したロジスティック回帰，流暢性は RNN 言語モデル，意味保存性は METEOR [13] を用いて評価し，各評価値の重み付き和を最終的な評価とした．各評価値の重みは JFLEG データセット [14] を用いてチューニングを行い，システムレベル・文レベルで Grundkiewicz らの評価データ [15] を用いてメタ評価を行なっている．浅野ら [4] の流暢性と意味保存性の自動評価モデルは人手評価に対して最適化されておらず，特に意味保存性は人手評価との相関が低い．本研究では浅野ら [4] の拡張として，文法性・流暢性・意味保存性の各自動評価モデルを各項目の人手評価に対して最適化を行う．各項目の評価モデルを人手評価に最適化した効果を検証するために，最終的な評価値の計算方法や設定は浅野ら [4] と同様にして実験を行う．

2.2 各項目の評価モデルの学習に使用できる既存のデータセット

文法性に関しては，浅野ら [4] が文法性に関する自動評価モデルを訓練するために使用している，GUG データセット* [12] がある．GUG データセットは，学習者の書いた文に対して人手による文法性の評価値が付与されている．流暢性に関しては，British National Corpus および Wikipedia の英文を Google 翻訳で折り返し翻訳した擬似誤り文に対して人手による流暢性（Acceptability）の評価値が付与されたデータセット† [16] が公開されている．文法性，流暢性に関しては上述のデータセットが存在するが，意味保存性に関しては，誤り文を含む文対に対して人手評価が付与されたデータセットは存在しない．

これらのデータセットは，学習者が書いた文や機械翻訳が生成した文に対して人手評価を付与している．本研究では，これらのテキストは文法誤り訂正システムの訂正文とは異なる性質を持つと考え，実際の訂正文に対して文法性・流暢性・意味保存性の自動評価を新たに収集し，自動評価モデルを訓練する．

*https://github.com/EducationalTestingService/gug-data

†https://clasp.gu.se/about/people/shalom-lappin/smog/experiments-and-datasets

2.3 自動評価手法の評価

自動評価手法の評価（メタ評価）は，人手評価と高い相関を持つ自動評価ができることを検証するのが最も直観的であり [17]，一般的には人手評価との相関係数を用いて評価される．実際に，機械翻訳の自動評価手法の Shared task である WMT 2019 Metrics Shared Task [18] では，複数の翻訳システムの出力文に対する人手評価と自動評価の相関を用いて，システムレベルと文レベルでのメタ評価を行なっている．

文法誤り訂正における自動評価手法のメタ評価では，複数の訂正システムに対する人手のランキングとの相関を測るメタ評価が行われている [15, 6, 3, 4]．浅野ら [4] はシステムレベルのメタ評価だけでなく文レベルでのメタ評価も行なっている．本研究でも同様に，システムレベルと文レベルの両方のメタ評価を行う．Choshen らは人手のランキングを用いるメタ評価とは別に，人手の訂正から構築した束に基づく文法誤り訂正のメタ評価手法（MAEGE）を提案した [8]．人手の訂正を用いることで，人手によるランキングを用いるメタ評価の評価者間および評価者内の一致が低い問題に対処している．さらに，システム出力に対する人手ランキングに依存することによる一部の有効なエラータイプの訂正を適切に評価できない問題に対して，人手による訂正から構築した束全体を評価に用いることで対処している．本研究ではより適切なメタ評価を行うため，人手評価との相関によるメタ評価だけでなく，人手の訂正に基づく MAEGE によるメタ評価も行う．人手評価との相関よるメタ評価の詳細を 5.2.1 節に，MAEGE によるメタ評価の詳細を 5.2.2 節に示す．

第 3 章　提案手法: 事前学習された文符号器を用いた自動評価モデル

文法性・流暢性・意味保存性の人手評価値付きデータセットを用いて，各項目の評価モデルの最適化を行う．各項目の評価モデルに，訂正文あるいは入力文と訂正文の対から人手評価値を推定する回帰モデルとして, 事前学習された文符号化器である BERT [7] を用いることを提案する．BERT は，大規模な生コーパスを用いて Self-Attention Network に基づくモデル [19] の事前学習を行う．事前学習は，生コーパスの一部のマスクされた単語を当てる学習と，2 文が隣接しているかどうかを 2 値分類する学習の 2 つを同時に行う．BERT はそれまでの単方向でしか学習していなかった事前学習モデルとは異なり，双方向での学習が可能となり，文全体を考慮した表現を獲得することができる．事前学習済みの BERT は，対象タスクに適した出力層に変更し，モデル全体を少量のデータセットで再学習させることで様々な自然言語処理タスクで高精度の予測を行うことができる．本研究では人手評価値の予測を行うため，図 3.1 のように出力層を多層パーセプトロンの回帰モデルに変更して再学習を行う．BERT は文および文対の符号化ができる．文の符号化では，先頭に [CLS] トークンを追加して BERT に入力する（図 3.1）．最終隠れ層の C が文の分散表現を示し，回帰モデルの入力として使用される．文対の符号化で

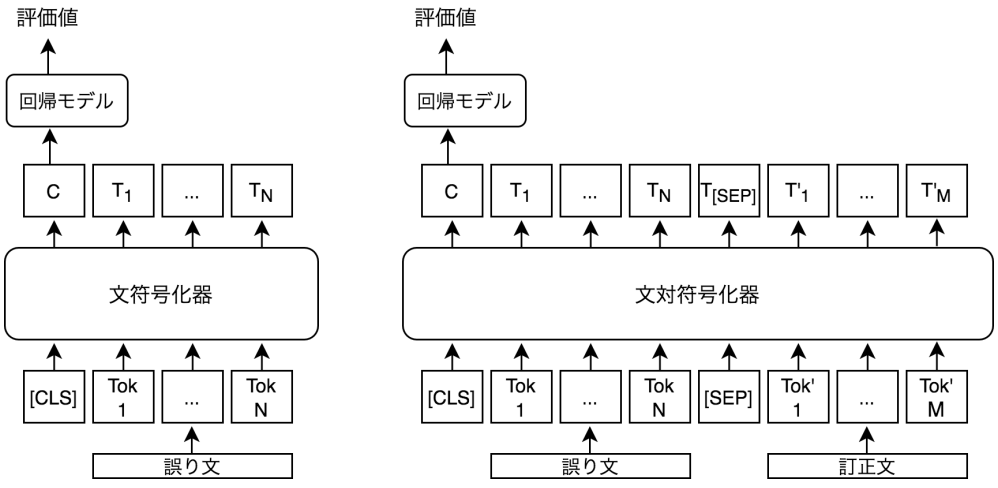

図 3.1: BERT による文（左）および文対（右）のモデリング

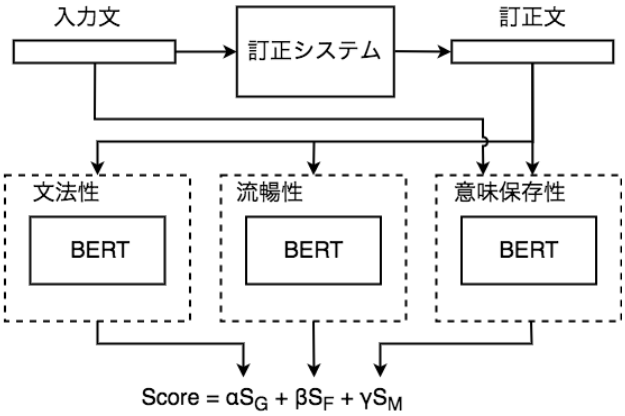

図 3.2: 評価手法の全体像（α，β，γ は各評価値の重み）

は，先頭に [CLS] トークンを付与して，誤り文と訂正文の間に [SEP] トークンを付与して BERT に入力する（図 3.1）．同様に [CLS] トークンに対応する最終隠れ層の C が文対の分散表現を示し，回帰モデルの入力として使用される．再学習を行う際は，回帰モデルだけでなく文符号化器も同時に再学習される．

文法性・流暢性・意味保存性の項目ごとに BERT の再学習を行い，各項目の人手評価に最適化した自動評価モデルを構築する．再学習には各項目の人手評価値が付与されたデータセットを用いる．ただし，図 3.2 に示すように，文法性および流暢性については訂正文のみから人手評価を推定し，意味保存性については入力文と訂正文の対から人手評価を推定する．文法性および流暢性は訂正文のみから評価できるが，意味保存性は入力文と訂正文の 2 つを用いないと評価できないためである．

訂正文の最終的な評価値は，浅野ら [4] と同様に各項目の評価値の線形和で求める．

$$\mathrm{Score} = \alpha \cdot \mathrm{Score_G} + \beta \cdot \mathrm{Score_F} + \gamma \cdot \mathrm{Score_M}. \tag{3.01}$$

ここで，$\mathrm{Score_G}$，$\mathrm{Score_F}$，$\mathrm{Score_M}$ はそれぞれ文法性，流暢性，意味保存性の評価値であり，各評価値は 0 から 1 になるように正規化を行う．重み α，β，γ は $\alpha+\beta+\gamma=1$ であり，いずれも負の値を取らない．各重みの決め方については 5.2 節で説明する．

第 4 章　訂正文に対する人手評価値付きデータセットの構築

文法誤り訂正の自動評価に理想的なデータで自動評価モデルの学習を行うために，訂正システムの訂正文に対して各項目の人手評価が付与されたデータセットを構築する．文法誤り訂正の検証用に一般的に利用されている CoNLL 2013*のテストデータである 1,381 文を複数の訂正システムで訂正した訂正文に対して，文法性・流暢性・意味保存性の人手評価を収集する．†CoNLL 2013 のテストデータは，英語を母語としないシンガポール国立大学の学生 25 名によって書かれたエッセイをもとに作られている．監視技術と高齢化についての 2 つのトピックについて書かれており，学生の習熟度は比較的高い．

4.1　訂正文生成のための文法誤り訂正システム

多様な訂正文に対して人手評価を収集するために，各入力文を代表的な以下の 4 種類の文法誤り訂正システムおよびシステムの学習時に公開されている中で最高性能のシステム [20]‡によって訂正し，アノテーションを実施する．

■SMT　：統計的機械翻訳 (Statistical Machine Translation) に基づくモデル．Grundkiewicz ら [21] の実装§を用いた．言語モデルおよび単語クラス言語モデルの訓練には，公開されているデータ¶のうち part00 から part02 までの 30 億文を用いた．KenLM‖ [22] を用いて言語モデルの学習を行なった．単語クラスの学習には Word2Vec [23] を用いた．

*https://www.comp.nus.edu.sg/~nlp/conll13st.html

†一般的に文法誤り訂正の評価に用いられている CoNLL 2014 や JFLEG を自動評価モデルの学習データとして使用するのは不適切なため使用は避けた．

‡2019 年 7 月時点

§https://github.com/grammatical/baselines-emnlp2016/tree/master/train-2018

¶http://data.statmt.org/romang/gec-emnlp16/sim/

‖https://kheafield.com/code/kenlm/

表 4.1: CoNLL 2013 における各システムの $F_{0.5}$ 値

システム	本研究	先行研究
SMT	34.20	32.30 [27]
RNN	35.50	33.76 [27]
CNN	33.04	31.96 [26]
SAN	35.57	36.20 [27]
SAN+Copy	40.89	- -

■RNN ：Recurrent Neural Network に基づく系列変換モデル．実装には fairseq** [24] を用いた．4 層の Bi-directional LSTM を用い，単語埋め込みの次元数は 512，バッチサイズは 32 とし，その他の設定は Luong ら [25] に従った．

■CNN ：Convolutional Neural Network に基づく系列変換モデル．実装には fairseq** [24] を用いた．符号化器および復号器の次元数は 512 とし，その他の設定は Chollampatt ら [26] に従った．

■SAN ：Self-Attention Network に基づく系列変換モデル．実装には fairseq** [24] を用いた．設定は Vaswani ら [19] に従った．

■SAN+Copy ：文法誤り訂正の入力文の単語が訂正されない割合が高いという性質を考慮して，訂正する必要がない部分は入力文から単語を直接コピーして出力できるようなコピー機構を SAN に追加したモデル．データセットの作成時に公開されているモデルの中で最高性能のモデルであった．Zhao ら [20] の実装††を用いた．

各モデルの訓練のために，SAN+Copy は著者らが公開している訓練用データ（Lang-8, NUCLE, FCE）を，その他のモデルは Bryant らの訓練用データ‡‡（Lang-8, NUCLE, FCE, W&I）[28] を用いた．ただし，100 トークン以上の文を削除し

**https://github.com/pytorch/fairseq

††https://github.com/zhawe01/fairseq-gec

‡‡https://www.cl.cam.ac.uk/research/nl/bea2019st/#data

た後に，Byte Pair Encoding[§§] [29] を用い，結合回数を 30,000 として単語のサブワード化を行なった．各モデルを CoNLL 2013 のデータで 2 分割交差検証によって評価した結果および類似の設定で実験を行っている先行研究の結果の比較を表 4.1 に示す．ただし，先行研究は 2 分割交差検証ではなく，NUCLE データの一部を検証データとして CoNLL 2013 の評価を行なっている．先行研究の報告と同等の性能を確認できたので，2 分割交差検証のテストデータに対する訂正文に人手評価を付与する．

4.2 訂正文に対する文法性・流暢性・意味保存性のアノテーション

CoNLL2013 の 1,381 文を前節の各システムで訂正し，重複する訂正文を除いた合計 4,223 文に対して文法性・流暢性・意味保存性の人手評価値を付与する．文法性および流暢性は訂正文のみで評価を行い，意味保存性は入力文と訂正文から評価を行うようにした．

■文法性 ：訂正文の文法的な正しさを評価する．Heilman ら [12] の 5 段階評価（4. 完璧，3. 分かりやすい，2. 理解できる，1. 理解できない，0. その他）に従って評価した．

■流暢性 ：訂正文の自然さを評価する．Lau ら [16] の 4 段階評価（4. 非常に自然，3. やや自然，2. やや不自然，1. 非常に不自然）に従って評価した．

■意味保存性 ：入力文と訂正文の間の意味内容の等価性を評価する．Xu ら [30] の 5 段階評価（4. 同一，3. わずかに異なる，2. 異なる，1. 大幅に異なる，0. その他）に従って評価した．

Amazon Mechanical Turk[¶¶]を用いて，1 文あたり 5 人の評価者を募集した．データセットの質を担保するために，US 在住者のうち，質の高い回答を行う Master 資格を保有し，過去のタスク承認率 99% 以上かつ承認数 50 以上の評価者を採用した．また，入力文や訂正文を読まずに回答する評価者を拒否するために，

[§§] https://github.com/rsennrich/subword-nmt
[¶¶] https://www.mturk.com/

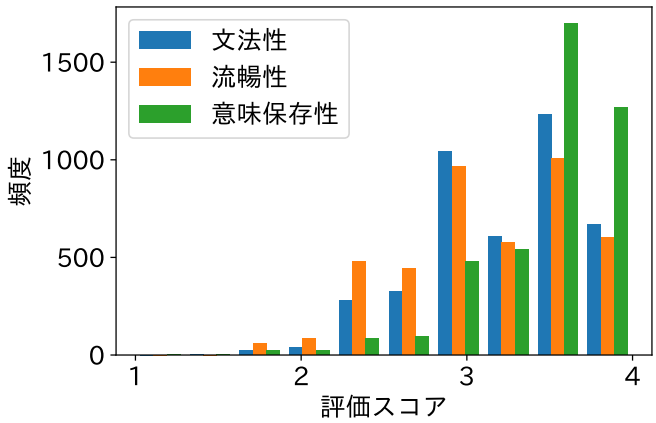

図 4.1: 各項目のヒストグラム

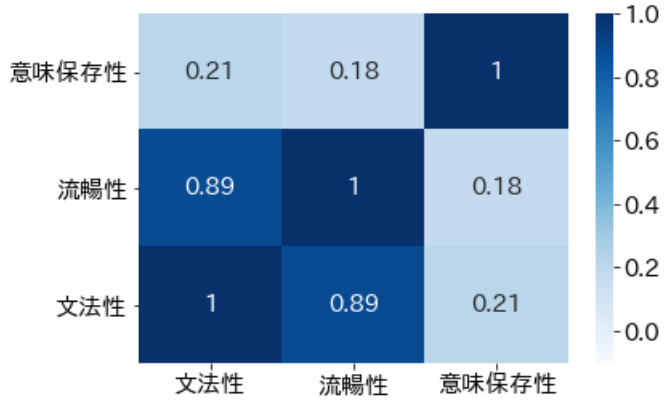

図 4.2: 各項目間の相関行列

全ての項目に 4 を回答させるようなダミーの設問を設置した．さらに，全ての項目に同じ回答をする評価者や，極端に回答時間が速い評価者も拒否した．最終的に，3 人以上の評価者が「0. その他」***を回答した文を除き，合計 4,221 文の人手評価値付きデータセットを作成した．評価者 1 人あたりおおよそ時給 7.25 ドルと見積もって作業を分割して依頼し，合計約 10 万円でデータセットを作成した．

人手評価値のヒストグラムを図 4.1 に示す．2 以下の評価は全体に少なく，特に意味保存性の項目は多くが 3 以上の評価を得た．各項目間の相関行列を図 4.2 に

***不完全な文や意味不明な文.

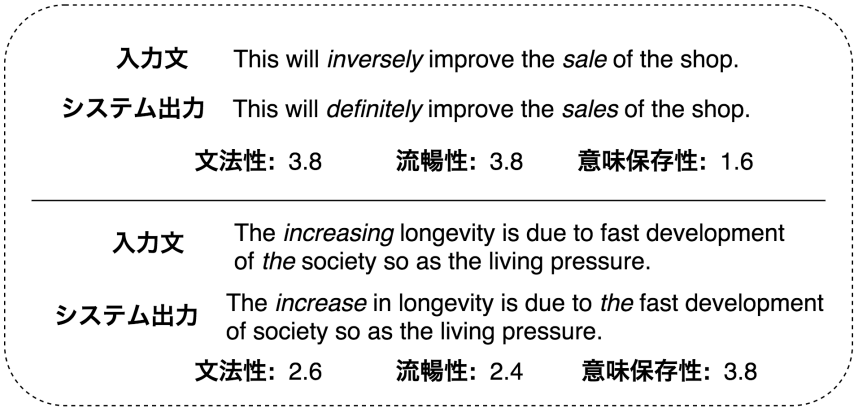

図 4.3: 実際の評価例

示す．各要素の値はピアソンの積率相関係数を示す．文法性と流暢性との相関は高く，意味保存性は他の 2 項目間との相関が低い．流暢性の観点は文法性の観点も含まれているため高い相関になったと考えられる．

図 4.3 に実際のアノテーション例を示す．上の評価例では，文法性および流暢性の評価は高いが，副詞の “inversely” が “definitely” になっているため意味保存性は低い評価となっている．下の例では文法性および流暢性の評価は低いが，形の変化や冠詞の有無の違いであり，全体としては意味はほとんど変わらないため意味保存性は高い評価となっている．

第 5 章　実験設定

提案手法の有効性を検証するために、人手評価との相関によるメタ評価および MAEGE によるメタ評価を行い，既存手法と比較を行う．さらに，本研究で作成した，訂正システムに対して人手評価値を付与したデータセットで BERT を再学習することの有効性を検証するために、2.2 節で説明した既存のデータセットで BERT を再学習した場合との比較も行う．各項目の自動評価モデル単体のメタ評価を行なった後，3 章で説明した各自動評価モデルを組み合わせた手法のメタ評価を行う．

5.1　BERT の再学習

4 章で構築したデータセットを訓練/検証/評価に 3,376/422/423 で分割し，BERT の再学習に使用する．比較手法における BERT の再学習のために，文法性および流暢性については 2.2 節で説明したデータセットを使用する．意味保存性については，2 文間の意味的類似度を $[0.0, 5.0]$ の連続値で評価した Semantic Textual Similarity タスク [31] のデータセット*を使用する．文法性と流暢性に用いるデータセットは誤った文に対して評価が付与されているのに対し，意味保存性では誤りを含まない文対に対して評価が付与されている．事前学習済みの BERT (BERT-BASE-CASED)†を，それぞれのデータセットを用いて再学習する．ハイパーパラメータは検証データを用いて，最大文長は 128，256，バッチサイズは 8，16，学習率は 2e-5，3e-5，5e-5，エポック数は 1 から 10 の組み合わせからグリッドサーチで決定した．その他の学習設定は使用した事前学習済みモデルの事前学習時のものと同じである．検証データでスピアマンの順位相関係数が最大となるモデルを選択した．

*http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark
†https://github.com/huggingface/transformers

5.2 メタ評価実験

5.2.1 人手評価との相関によるメタ評価

■システムレベルのメタ評価 システムレベルの評価では，図 5.1 のように，システムレベルの自動評価値と人手評価値の相関を測ることでメタ評価を行う．システムレベルの自動評価値は，各訂正文の自動評価値をシステムごとに平均して用いた．浅野ら [4] と同様に，Grundkiewicz ら [15] が公開しているデータを用いる．このデータは 12 システムによる CoNLL2014 のテストデータの訂正文に人手で 1 から 5 のランク付けを行なったものである．CoNLL2014 のテストデータは CoNLL2013 のテストデータと同様に，シンガポール国立大学の学生 25 名の書いたエッセイをもとに作られているが，トピックが異なり，遺伝子検査とソーシャルメディアの 2 つのトピックについて書かれている．さらに Grundkiewicz ら [15] はそのランクからレーティングアルゴリズムである True Skill [32] を用いてシステムレベルでの人手評価を計算した．浅野ら [4] と比較を行うために，Grundkiewicz ら [15] の Table 3c[‡] の人手評価を用いて計算した．相関係数にはピアソンの積率相関係数とスピアマンの順位相関係数を用いた．式 (3.01) の重み α，β，γ は浅野ら [4] と同様に JFLEG データセット [14] の人手評価を用いて調整した．JFLEG データセットを用いたのは，単純には浅野ら [4] の結果と比較するためであるが，評価データや学習に使った CoNLL データと，システムの数・学習者の習熟度・編集率などの性質が大きく異なる JFLEG データを使って重みを決めることでデータ依存性の問題を議論できるからである．各重み α，β，γ は 0.01 刻みで，$\alpha + \beta + \gamma = 1$, を

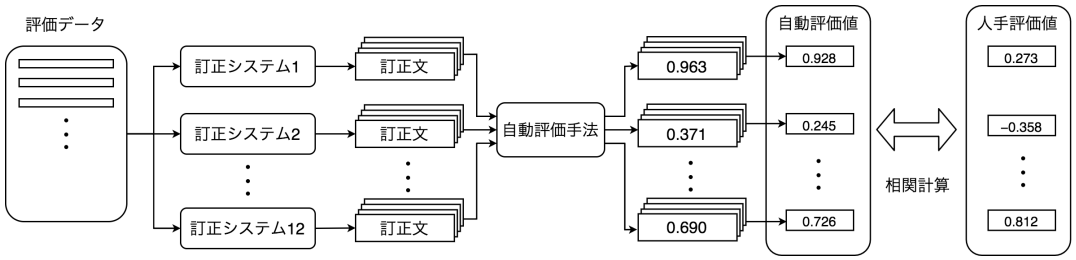

図 5.1: システムレベルでのメタ評価

[‡]Table 3c には，12 システムの名前および True Skill によって計算された人手評価が記載されている．

満たす組み合わせからグリッドサーチを行い，ピアソンの積率相関係数を最大化するように決定した．

■文レベルのメタ評価　文レベルのメタ評価では，各訂正文の自動評価値を人手評価値と直接比較する．浅野ら [4] と同様に，Grundkiewicz ら [15] のデータセットにおける正解率およびケンドールの順位相関係数によって，任意の 2 つの訂正文の優劣判定調査を行う．同一文に対するシステム出力の評価が異なるペアの評価を行い，その正解率（Accuracy）(式 (5.21)) と WMT2017 [33] で使われているケンドールの順位相関係数 τ（Kendall's τ） (式 (5.22)) を用いてメタ評価を行う．一部の文に対しては複数人が評価を行なっているが，それらは別事例とみなし，14,822 ペアに対して評価を行なった．

$$Accuracy = \frac{\text{適切に評価したペア数}}{\text{人手評価の大小が異なるペア数}}. \tag{5.21}$$

$$Kendall's\ \tau = \frac{\text{適切に評価したペア数} - \text{逆に評価したペア数}}{\text{人手評価の大小が異なるペア数}}. \tag{5.22}$$

式 (3.01) の重み α，β，γ は Grundkiewicz ら [15] のデータセットを 1:9 の割合で検証用と評価用に分割し，検証用データにおけるケンドールの順位相関係数を最大化するように調整した．グリッドサーチの探索範囲はシステムレベルの評価と揃えた．

5.2.2　MAEGE によるメタ評価

Choshen ら [8] は，人手評価との相関を用いるメタ評価に存在する，評価者間および評価者内の一致が低い問題や，一部のエラータイプに対する訂正を適切に評価できない問題に対処するために，人手による訂正を用いた自動評価手法のメタ評価手法（MAEGE）を提案した．人手による訂正アノテーションが付与されたコーパスを用い，各文に対して図 5.2 に示すような人手の訂正を元とする束を構築する．全ての訂正が全体的な品質に等しく作用すると仮定して，訂正の適用数で順序づけが行われる．束を適用した誤り文セットを自動評価手法で評価を行い，定義され

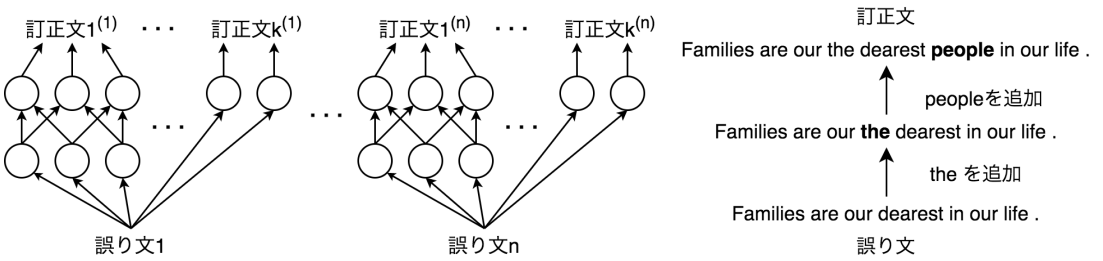

図 5.2: 各誤り文に対する人手の訂正を元とした束（左図）と実際の例（右図）．各有向エッジは人手の訂正の適用を表し，訂正文 k は誤り文に対して k 人目の訂正者の全ての訂正を適用した文である．

た順序関係をゴールドのランキングとして比較を行い評価を行う．MAEGE では，コーパスレベルと文レベルのメタ評価を行うことができる．詳細な説明は Choshen らの論文 [8] を参照せよ．

本研究の実験では，CoNLL 2014 [34] のテストデータを使用する．実験は公開されている実装[§]を使用したが，束を構築する際に入力文または訂正を適用した結果が空文にならないように訂正を加えた．コーパスレベルのメタ評価では Choshen ら [8] と同様の設定で，ピアソンの積率相関係数およびスピアマンの順位相関係数を用いて評価した．文レベルのメタ評価では Choshen ら [8] と同様の設定で，ケンドールの順位相関係数 τ [¶]およびピアソンの積率相関係数で評価した．各メタ評価における各項目の重みは 5.2.1 節の文レベルのメタ評価で決めた値を用いた．Choshen ら [8] では，2 人のアノテーションのうち 1 人が訂正を行なっていない文は削除している．それに加えて，上述した実装に加えた条件によって，元の 1,312 文から 808 文が選択される．その 808 文に対する合計 1,704 のアノテーションから合計 6,634 文が生成され，評価に使用された．それぞれ 3 つのランダムシードで実験を行い，平均値を報告する．

[§]`https://github.com/borgr/EoE`

[¶]通常用いられる，全順序集合に対して定義されたものではなく，半順序集合に合わせて Choshen ら [8] が再定義したものを用いている．

5.3 ベースライン手法

本実験では，3 つの自動評価手法と提案手法を比較する．参照文を用いる自動評価手法としては，2.1.1 節で説明したうち，文法誤り訂正の自動評価で一般的に使用されている M^2 [1] および GLEU [2] を用いる．参照文を用いる手法では，性能を最大にして比較を行うために，公式の 2 つの参照文に加えて Bryant ら [5] の追加の 8 つの参照文と Sakaguchi ら [6] の追加の 8 つの参照文を加えた 18 文を用いる．参照文を用いない自動評価手法としては，先行研究である浅野ら [4] の手法を用いる．

第 6 章　実験結果

各実験結果の表における，BERT w/ existing data は既存のデータで BERT を再学習した場合，BERT w/ our data は作成したデータで BERT の再学習を行なった場合の結果を表す.

6.1　人手評価との相関によるメタ評価

表 6.1 に文法性・流暢性・意味保存性の各項目の自動評価モデルの，作成した評価データを用いたメタ評価の結果を示す．浅野ら [4] のシステムレベルにおける結果は再実装したものを用いた．浅野ら [4] の手法と比較して，既存のデータで BERT を再学習する意味保存性の評価モデルを除いて，提案手法が高い相関を達成している．既存のデータで BERT を再学習する意味保存性の評価モデルの相関が高くないのは，再学習に用いたデータが文法的な誤りを含まない文対に対して人手評価が付与されていたからであると考えられる．再学習に用いるデータセットの違いでは，我々が作成したデータで再学習した各評価モデルが最高の人手評価との相関と

表 6.1: 作成したデータセットにおける各項目の自動評価手法のメタ評価.

	項目	ピアソン	スピアマン
浅野ら	文法性	0.342	0.358
	流暢性	0.220	0.238
	意味保存性	0.593	0.504
BERT w/ existing data	文法性	0.608	0.624
	流暢性	0.545	0.548
	意味保存性	0.570	0.355
BERT w/ our data	文法性	**0.700**	**0.719**
	流暢性	**0.676**	**0.696**
	意味保存性	**0.639**	**0.619**

表 6.2: Grundkiewicz ら [15] のデータセットにおける各項目の自動評価手法のメタ評価.

		システムレベル		文レベル	
	項目	ピアソン	スピアマン	正解率	ケンドール
浅野ら	文法性	0.759	0.835	0.641	0.283
	流暢性	0.864	0.819	0.707	0.415
	意味保存性	**0.198**	**−0.192**	0.189	**0.059**
BERT w/ existing data	文法性	0.966	0.967	0.735	0.483
	流暢性	0.865	0.742	0.714	0.443
	意味保存性	−0.462	−0.610	0.502	0.016
BERT w/ our data	文法性	**0.976**	**0.973**	**0.745**	**0.502**
	流暢性	**0.979**	**0.978**	**0.741**	**0.494**
	意味保存性	−0.517	−0.621	**0.504**	0.022

なっており，訂正システムの出力に対する人手評価に最適化をすることの有効性が確認できる．

表 6.2 に文法性・流暢性・意味保存性の各項目の自動評価モデルの，Grundkiewicz ら [15] の評価データを用いたメタ評価の結果を示す．浅野ら [4] のシステムレベルにおける結果は浅野ら [4] からの引用，文レベルにおける結果は再実装したものを用いている．表 6.1 では，各自動評価モデルは対応する項目の人手評価との相関を計算しているが，表 6.2 では，各自動評価モデルは総合的な人手評価との相関を計算している．意味保存性はシステムレベルでは負の相関，文レベルでは無相関となったが，文法性と流暢性に関しては全ての項目で我々の作成したデータで BERT を再学習した手法がベースラインより高い相関となっている．意味保存性のシステムレベルで負の相関になっている理由は，訂正を行わないシステムに対する意味保存性単体の評価は高くなるのに対して，総合的な人手の評価では訂正しないシステムに対して低い評価となるため負の相関になっていると考えられる．実際に，評価データ中の各システムの訂正を行わない文数と提案手法の意味保存評価モデルの予測値のスピアマンの順位相関係数は 0.923 となり，各システムの訂正を行わない文

表 6.3: システムレベル（左）と文レベル（右）のメタ評価

	システムレベル			文レベル		
	ピアソン	スピアマン	重み (α:β:γ)	正解率	ケンドール	重み (α:β:γ)
M^2	0.674	0.720	-	0.464	0.294	-
GLEU	0.846	0.816	-	0.670	0.354	-
浅野ら	0.898	0.912	-	0.690	0.390	0.02:0.82:0.16
BERT w/ existing data	0.937	0.912	0.85:0.00:0.15	0.744	0.502	0.88:0.12:0.00
BERT w/ our data	**0.979**	**0.978**	0.00:1.00:0.00	**0.749**	**0.510**	0.55:0.43:0.02

数とシステムの人手の順位とのスピアマンの順位相関係数は -0.549 となった．

表 6.3 に，3 章で説明した各評価項目のモデルを組み合わせた手法の，システムレベルおよび文レベルの人手評価との相関によるメタ評価の結果を示す．浅野ら [4] の結果は，システムレベルにおいては浅野ら [4] における 3 項目を組み合わせた手法の中で最大の相関値を引用，文レベルにおいては再実装したものを使用している．システムレベルと文レベルの両方で，BERT の再学習に基づく自動評価手法が他の自動評価手法よりも大幅に高い性能を示したことから，文法誤り訂正の自動評価において事前学習された文符号化器 BERT を用いることの有効性が確認できる．BERT の再学習に使用するデータセットの違いからは，システムレベルと文レベルの両方で我々のデータセットを用いる再学習が人手評価とのより高い相関を達成することがわかる．このことから，実際のシステムの訂正文に対する各項目の人手評価に自動評価モデルをそれぞれ最適化することの有効性が確認できる．文法性・流暢性・意味保存性の 3 項目を組み合わせる手法では，全体的に意味保存性の重み γ が小さくなっている．これは，文法誤り訂正では入力文と訂正文の多くの単語が共通するため，訂正の良し悪しによらず多くの場合に意味が変わらないことが原因だと考える．文レベルでは各項目の重み (α:β:γ) が比較的均等に重みづけされているのに対し，システムレベルでは 1 つの重みに偏っている．これは，重みのチューニングに，4 つのシステムに対する人手のランキングがついた JFLEG データセットを用いており，4 つのシステムの評価は各項目を組み合わせずとも，文法性・流暢性の単体の評価器で高い精度で予測ができてしまうからであると考えられる．実際に，作成したデータにおける文法性単体の評価器は，JFLEG データセットに対し

て 0.976 のピアソンの積率相関係数，1.0 のスピアマンの相関係数であり，流暢性単体の評価器ではそれぞれ 0.978，1.0 であった．JFLEG データセットは流暢性を重要視して作られているため，BERT w/ our data では流暢性の重みが高くなっている．BERT w/ existing data では，流暢性単体よりも文法性単体の評価器の方が JFLEG データセットに対する相関が高いため文法性の重みが高くなっていると考えられる．実際に，文法性単体の評価器は JFLEG データセットに対して 0.963 のピアソンの積率相関係数，流暢性単体の評価器は 0.957 のピアソンの積率相関係数であった．

6.2 MAEGE によるメタ評価

表 6.4 に MAEGE によるコーパスレベルと文レベルのメタ評価の結果を示す．上段の参照文を用いる手法と下段の参照文を用いない手法を比較すると，コーパスレベルおよび文レベルの両方，特に文レベルにおいて参照文を用いない手法が高い相関となっている．参照文を用いる手法の M^2 は文レベルにおいてピアソンの積率相関係数は無相関，ケンドールの順位相関ではある程度の相関となっており，GLEU は M^2 と逆の振る舞いを示している．M^2 は同じ文の異なる訂正に対する順序付けはある程度予測できるが，全体的に一貫した評価をすることはできず，GLEU はその逆の性質を持つことを意味している．参照文を用いない手法を比較

表 6.4: MAEGE によるメタ評価

	コーパスレベル		文レベル	
	ピアソン	スピアマン	ピアソン	ケンドール
M^2	0.286	0.236	0.036	0.361
GLEU	0.798	0.755	0.250	0.045
浅野ら	0.874	0.918	0.271	0.480
BERT w/ existing data	0.882	0.973	**0.376**	0.642
BERT w/ our data	**0.883**	**0.976**	0.372	**0.680**

すると，BERT を用いた手法が 浅野ら [4] の手法より高い相関となっている．特に BERT を用いる手法はベースライン手法に比べて文レベルの両方の評価尺度で高い相関を持つことがわかる．このことから，BERT を用いた手法は同じ文の訂正ペアの順序付けと，全体的に一貫した評価の両方を行うことができることがわかる．再学習に用いるデータセットの違いでは，総合的に見ると，我々のデータセットを用いる方が相関が高くなった．

6.1 節の人手評価との相関を用いる実験および 6.2 節の MAEGE による実験の結果，作成したデータで BERT を再学習する手法が最もよい結果となった．各項目の評価モデルを作成したデータセットに最適化することで，各項目の評価モデルの性能が向上し（表 6.1，表 6.2），その結果，各評価モデルを組み合わせた手法で性能が向上した（表 6.3，表 6.4）．評価データ（CoNLL 2014）の性質とは異なる既存のデータで再学習した場合と評価データに近い性質のデータ（CoNLL 2013）で作成したデータで再学習した場合の両方で改善されているため，再学習に用いるデータの性質は大きな問題にならないと考えられる．表 2 や表 3 のように各項目の評価を見ると大きく改善しているため，データ作成にかかるコスト約 10 万円を考慮しても，各項目をより適切に評価するためにシステム出力にラベル付けして新たにデータを作る意味は十分あると考えられる．

第 7 章　分析

7.1　評価例

M^2，GLEU，浅野ら [4]，作成したデータで BERT の再学習を行なった場合の提案手法による評価例を分析する．Grundkiewicz ら [15] の評価データに対する各評価手法を比較する．表 7.1 に，提案手法のみが正しく評価できた例を示す．M^2 は訂正文 1 と訂正文 2 に対して同じ評価値を与えており正しく評価できていない．表層的な語句の一致率に基づく GLEU は，参照文に “problems” が含まれないために訂正文 1 を低く評価し，表層的には似ているが余計な “the” が入っている訂正文 2 を高く評価してしまっている．一方で提案手法は，“disadvantages” が “problems” になっていても余計な “the” が入っている訂正文 2 よりも訂正文 1 のほうを高く評価できており，表層的な語句の一致率に依存せず評価できている．浅野ら [4] が正しく評価できていないのは，浅野ら [4] が主に流暢性，次に意味保存性を見ている（表 6.3）が，流暢性の評価モデルは人手評価との相関が低い（表 6.1）ことおよび，意味保存性の評価モデルは METEOR であり入力文との表層一致で評価値を計算するため入力文に表層の近い訂正文 2 を高く評価するからであると考えられる．

表 7.2 に，参照文を用いる手法では正しく評価できたが，参照文を用いない手法では正しく評価できなかった例を示す．参照文を用いる手法では，参照文には “child” が含まれているため，“child” が含まれている訂正文 2 を高く評価してい

表 7.1: 提案手法のみが正しく評価できた例

入力文	There are a lot of disadvantages that people may not realize of .				
参照文	There are a lot of disadvantages that people may not realize .				
	There are a lot of *problems* that people may not realize .				
訂正文 1	人手評価	M^2	GLEU	浅野ら [4]	提案手法
	✓	0.556	0.586	0.949	**0.913**
	There are a lot of *the* disadvantages that people may not realize .				
訂正文 2	人手評価	M^2	GLEU	浅野ら [4]	提案手法
	×	0.556	0.630	0.977	0.826

表 7.2: 参照文を用いる手法のみが正しく評価できている例

入力文	Therefore I believe the parents have their right to know the healthiness of their child .
参照文	Therefore , I believe the parents have the right to know about the healthiness of their child .
訂正文 1	Therefore , I believe parents have their right to know the healthiness of their child . 人手評価 M^2 GLEU 浅野ら 提案手法 ✓ **0.456** **0.320** 0.850 0.873
訂正文 2	Therefore , I believe parents have their right to know the healthiness of their *children* . 人手評価 M^2 GLEU 浅野ら 提案手法 × 0.333 0.245 0.883 0.881

る．一方で，参照文を用いない手法では，“child” の部分が “children” となっている訂正文 2 を高く評価している．これは，参照文を用いないため，参照文の情報を考慮した評価ができないためである．

7.2 エラータイプ別の評価分析

MAEGE では，特定のエラータイプの訂正に対する評価の分析を行うことができる．特定のエラータイプの訂正のみが異なる文のペア (c, c') の集合を用意し，各ペア (c, c') の各文を評価して，その差分 $m(c) - m(c')$ の平均を計算する．c は特定のエラータイプの訂正を適用した後の文，c' はそのエラータイプの訂正を適用する前の文，m は評価手法を示す．負の値の平均差は評価手法がエラータイプの訂正に対して減点し，正の値の平均差は加点することを意味している．例えば，エラータイプが Vt（Verb tense）で，c が “The medical treatment technology during that time ***is*** not advanced enough to completely cure him .”，c' が “The medical treatment technology during that time ***was*** not advanced enough to completely cure him .” のペアに対して各文の評価を行い $m(c) - m(c')$ が正ならば訂正した文の方が高く評価できているため，適切に評価できており，負ならば，

表 7.3: エラータイプ分析

エラータイプ	M^2	GLEU	浅野ら	BERT_exist	BERT_our
Wci: (Wrong collocation/idiom)	0.000	-0.047	0.062	0.036	0.064
ArtOrDet: (Article or determiner)	0.000	0.003	0.055	0.040	0.068
Mec: (Spelling, punctuation, capitalication, etc.)	0.000	-0.010	0.052	0.038	0.072
Prep: (Preposition)	0.000	0.005	0.063	0.038	0.063
Nn: (Noun number)	0.000	0.025	0.066	0.045	0.078
Vt: (Verb tense)	0.000	0.038	0.048	0.035	0.063
Rloc-: (Redundancy)	0.000	-0.025	0.052	0.037	0.074
SVA: (Subject-verb agreement)	0.000	0.024	0.047	0.029	0.058
Pref: (Pronoun reference)	0.000	-0.042	0.042	0.028	0.042
Vform: (Verb form)	0.000	0.025	0.048	0.039	0.070
Trans: (Linking words/phrases)	0.000	-0.041	0.052	0.038	0.069
Wform: (Word form)	0.000	0.053	0.076	0.044	0.079
Others: (Other errors)	0.000	-0.057	0.047	0.035	0.052
Vm: (Verb modal)	0.000	-0.072	0.024	0.032	0.057
Ssub: (Subordinate clause)	0.000	-0.036	0.034	0.044	0.087
WOinc: (Incorrect word order)	0.000	-0.032	0.065	0.036	0.076
V0: (Missing verb)	0.000	-0.012	0.048	0.060	0.101
Pform: (Pronoun form)	0.000	-0.029	0.047	0.032	0.052
Um: (Unclear meaning)	0.000	0.009	-0.006	-0.005	0.012
WOadv: (Incorrect adjective/adverb order)	0.000	0.021	0.076	0.044	0.076
Npos: (Noun possessive)	0.000	0.010	0.067	0.020	0.054
Spar: (Parallelism)	0.000	-0.106	0.045	0.028	0.068
Srun: (Run-on sentences, comma splices)	0.000	-0.090	-0.046	0.039	0.084
Wtone: (Tone formal/informal)	0.000	0.017	0.052	0.038	0.060
Sfrag: (Sentence fragment)	0.000	-0.142	0.032	0.027	0.039
Smod: (Dangling modefiers)	0.000	-0.162	0.021	0.025	0.033
Wa: (Acronyms)	0.000	0.285	0.176	0.113	0.237

訂正した文の方を低く評価しているため，適切に評価できていない．エラータイプには CoNLL2014 [34] で指定されている 27 のエラータイプを使用する．

表 7.3 に結果を示す．M^2 では全てのエラータイプで平均差がほぼ 0（0.0001 など）になっており*，特定のエラータイプに対して一貫した評価を行えていないこと

*Choshen ら [8] の結果でも同様に M^2 は全てのエラータイプが 0 または 0 に近い値となっている．

がわかる．GLEU では多くのエラータイプの訂正に対して減点している．一方，参照文を用いない評価手法ではほとんど全てのエラータイプで正の値となっており，多くのエラータイプの訂正に対して加点できていることがわかる．特に，我々が作成したデータで BERT の再学習を行なった手法においては全てのエラータイプの訂正に対して加点できている．また，既存のデータで BERT の再学習を行った場合に比べて，各エラータイプに対する平均差の値が大きくなっており，各エラータイプの訂正に対してより大きく加点できていることがわかる．特に，他の手法に比べて連続文やコンマ区切りの誤り（Srun）や並列性の誤り（Spar）に対する平均差が大きく，得意な誤りであることがわかる．逆に，明確にしないと修正ができないような誤り（Um）は他の手法と同様に平均差が小さく，苦手な誤りであることがわかる．

第 8 章　おわりに

本研究では，文法誤り訂正の自動評価で重要である，文法性・流暢性・意味保存性の 3 項目の評価モデルを人手評価に最適化する手法を提案した．最適化するための理想的なデータセットが存在しないため，5 つの訂正システムの出力に対して，文法性，流暢性，意味保存性の観点で人手評価を行いデータセットの作成を行った．作成した各項目のデータセットを用いて，事前学習された文符号化器の再学習を行い，各項目の評価モデルを作成してそれらを組み合わせて文法誤り訂正の自動評価を行なった．

実験では，まず各項目の評価モデルの個別のメタ評価を行い，各項目をどれだけ適切に予測できるか実験した．そして，人手評価との相関によるメタ評価および，人手の訂正を用いたメタ評価で，各項目の評価モデルを組み合わせた手法を評価した．実験の結果，提案手法は従来手法に比べて，各項目をより適切に評価できること，各項目の評価モデルを組み合わせた評価でもより適切に評価できることを示した．そして，事前学習された文の符号化器である BERT を回帰モデルとして用いることおよび，我々が作成した訂正文に対する人手評価に最適化することが相関の改善に有効であることがわかった．分析の結果，従来手法ではいくつかのエラータイプの訂正に対して減点しているのに対して，提案手法は全てのエラータイプの訂正に対して適切に加点できていることがわかった．

今後の発展としては，文法誤り訂正の精度改善に提案した自動評価手法を活用することである．例えば，提案した自動評価手法の評価値を報酬として強化学習による学習を行うことや，複数モデルのアンサンブルを行う時に使用して，最も良い訂正を選択することなどが考えられる．

発表リスト

国際会議

1. Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara and Mamoru Komachi. **SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction.** In Proceedings of the 28th International Conference on Computational Linguistics, pp.6516–6522, 2020.
2. Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi and Mamoru Komachi. **Filtering Pseudo-References by Paraphrasing for Automatic Evaluation of Machine Translation.** In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp.521–525, 2017.

国内会議

1. 吉村綾馬，杉山雅和，友松祐太，小町守．**音声認識誤り検出における文単位のラベルからの単語単位の誤り予測．** NLP 若手の会第 15 回シンポジウム，2020.
2. 喜友名朝視顕，吉村綾馬，金子正弘，小町守．**項目別マルチタスク学習による系列変換タスクの品質推定．** NLP 若手の会第 15 回シンポジウム，2020.
3. 吉村綾馬，松村雪桜，山岸駿秀，小町守．**機械翻訳の自動評価のための擬似参照訳作成．** 言語処理学会第 25 回年次大会，2019.
4. 吉村綾馬，松村雪桜，山岸駿秀，小町守．**機械翻訳の自動評価のための N-best を用いたマルチリファレンス作成手法の提案．** NLP 若手の会第 13 回

シンポジウム，2018.

5. 中澤 真人，池田 可奈子，山田 美知花，吉村 綾馬，鈴木 由衣，小町 守．**リビュー文書を対象とした句単位の日本語評価極性タグ付きコーパス．** 言語処理学会第 24 回年次大会，2018.

謝辞

本論文の執筆にあたり，多くの方々にご協力，ご助言をいただきましたことに，心より感謝申し上げます．指導教員である小町守准教授には，毎週の進捗報告での議論や論文執筆の指導だけでなく，研究しやすい環境を提供してくださり心より感謝申し上げます．研究室配属時にメンターとして指導して下さった松村雪桜さん，山岸駿秀さんには基礎から細かく教えて頂き，本当に感謝しています．同じ領域の研究をされており，様々なアドバイスを頂きました嶋中宏希さん，本研究の方針や手法，論文執筆に関して直接の指導を頂いた金子正弘さん，梶原智之さんに心より感謝申し上げます．また，研究会や日々の議論などで様々なアドバイスを頂きました研究室のみなさん，ありがとうございます．最後に，副査を引き受けて下さり多くのアドバイスを頂きました山口亨教授と高間康史教授に感謝します．

参考文献

[1] D. Dahlmeier and H.T. Ng, “Better evaluation for grammatical error correction,” Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.568–572, 2012.

[2] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, “Ground truth for grammatical error correction metrics,” 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, 2015.

[3] C. Napoles, K. Sakaguchi, and J. Tetreault, “There’s no comparison: Reference-less evaluation metrics in grammatical error correction,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.2109–2115, 2016.

[4] 浅野広樹，水本智也，乾健太郎，“文法性・流暢性・意味保存性に基づく文法誤り訂正の参照無し評価，” 自然言語処理，vol.25，no.5，pp.555–576，2018.

[5] C. Bryant and H.T. Ng, “How far are we from fully automatic high quality grammatical error correction?,” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp.697–707, 2015.

[6] K. Sakaguchi, C. Napoles, M. Post, and J. Tetreault, “Reassessing the goals of grammatical error correction: Fluency instead of grammaticality,” Transactions of the Association for Computational Linguistics, vol.4, pp.169–182, 2016.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp.4171–4186, 2019.

[8] L. Choshen and O. Abend, “Automatic metric validation for grammatical error correction,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1372–1382, 2018.

[9] R. Dale and A. Kilgarriff, “Helping our own: The HOO 2011 pilot shared task,” Proceedings of the 13th European Workshop on Natural Language Generation, pp.242–249, 2011.

[10] M. Felice and T. Briscoe, “Towards a standard evaluation method for grammatical error detection and correction,” Proceedings of the 2015 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.578–587, 2015.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.311–318, 2002.

[12] M. Heilman, A. Cahill, N. Madnani, M. Lopez, M. Mulholland, and J. Tetreault, "Predicting grammaticality on an ordinal scale," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.174–180, 2014.

[13] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.376–380, 2014.

[14] C. Napoles, K. Sakaguchi, and J. Tetreault, "JFLEG: A fluency corpus and benchmark for grammatical error correction," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp.229–234, 2017.

[15] R. Grundkiewicz, M. Junczys-Dowmunt, and E. Gillian, "Human evaluation of grammatical error correction systems," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.461–470, 2015.

[16] J.H. Lau, A. Clark, and S. Lappin, "Unsupervised prediction of acceptability judgements," Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp.1618–1628, 2015.

[17] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp.65–72, 2005.

[18] Q. Ma, J. Wei, O. Bojar, and Y. Graham, "Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges," Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp.62–90, 2019.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems, pp.5998–6008, 2017.

[20] W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu, "Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data," Proceedings of the 2019 Conference of the North American Chapter of the Asso-

ciation for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp.156–165, 2019.

[21] R. Grundkiewicz and M. Junczys-Dowmunt, "Near human-level performance in grammatical error correction with hybrid machine translation," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp.284–290, 2018.

[22] K. Heafield, "KenLM: Faster and smaller language model queries," Proceedings of the Sixth Workshop on Statistical Machine Translation, pp.187–197, 2011.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Workshop Track Proceedings of 1st International Conference on Learning Representations, 2013.

[24] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp.48–53, 2019.

[25] T. Luong, H. Pham, and C.D. Manning, "Effective approaches to attention-based neural machine translation," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.1412–1421, 2015.

[26] S. Chollampatt and H.T. Ng, "A multilayer convolutional encoder-decoder neural network for grammatical error correction," Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp.5755–5762, 2018.

[27] M. Mita, T. Mizumoto, M. Kaneko, R. Nagata, and K. Inui, "Cross-corpora evaluation and analysis of grammatical error correction models — is single-corpus evaluation enough?," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp.1309–1314, 2019.

[28] C. Bryant, M. Felice, Ø.E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction," Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp.52–75, 2019.

[29] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1715–1725, 2016.

[30] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," Transactions of the Association for Computational Linguistics, vol.4, pp.401–415, 2016.

[31] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017

task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation," Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp.1–14, 2017.

[32] R. Herbrich, T. Minka, and T. Graepel, "TrueSkill™: A Bayesian skill rating system," Advances in Neural Information Processing Systems 19, eds. by B. Schölkopf, J.C. Platt, and T. Hoffman, pp.569–576, MIT Press, 2007.

[33] O. Bojar, Y. Graham, and A. Kamran, "Results of the WMT17 metrics shared task," Proceedings of the Second Conference on Machine Translation, pp.489–513, 2017.

[34] H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant, "The CoNLL-2014 shared task on grammatical error correction," Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pp.1–14, 2014.