

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[†]・金子 正弘[†]・梶原 智之^{†,††}・小町 守[†]

信頼できる文法誤り訂正の自動評価手法の構築は、文法誤り訂正の研究および開発の発展に有用である。可能な参照文を網羅することが難しいため、先行研究では参照文を用いない自動評価手法が提案されてきた。そのうちの一つは、文法性・流暢性・意味保存性を評価する3つの評価モデルを用いることで、参照文を用いる手法よりも人手評価との高い相関を達成した。しかし、各項目の評価モデルは人手評価には最適化されておらず、改善の余地が残されていた。本研究では、より適切な評価を行える自動評価手法の構築を目的として、各項目の評価モデルを事前学習された文符号化器を用いて人手評価に対して最適化する手法を提案する。また、最適化に理想的である、訂正システムの出力文に対して人手評価が付与されたデータセットの作成を行う。実験の結果、項目ごとの評価モデルおよびそれらを組み合わせた手法の両方で、従来手法と比べて人手評価との相関が向上し、事前学習された文符号化器を用いることおよび訂正文の人手評価に最適化することの両方が貢献していることがわかった。分析の結果、提案手法は従来手法に比べて多くのエラータイプの訂正を正しく評価できていることがわかった。

キーワード：文法誤り訂正、自動評価

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

RYOMA YOSHIMURA[†], MASAHIRO KANEKO[†], TOMOYUKI KAJIWARA^{†,††} and MAMORU KOMACHI[†]

The development of a reliable automatic evaluation metric of grammatical error correction (GEC) is useful for the research and development of GEC. Since it is difficult to cover all possible reference sentences, previous studies have proposed reference-less metrics. One of them achieved a higher correlation with manual evaluation than reference-based metrics by integrating metrics from the three perspectives of grammaticality, fluency, and meaning preservation. However, the correlation with the manual evaluation can be further improved because they are not considered for optimizing each metric for each manual evaluation. Therefore, in this study, we propose a method of optimizing each metric. Furthermore, we create a dataset with manual evaluation of system output that is ideal for optimization. Experimental results show

[†] 東京都立大学, Tokyo Metropolitan University

^{††} 愛媛大学, Ehime University

that the proposed method improves correlation with the manual evaluation in both the metric of each perspective and combining the metrics. We also demonstrate that both using pre-trained language models for optimization and optimizing to manual evaluation on system output of GEC contribute to improvement. As a result of the analysis, it was found that our proposed metric appropriately rewarded more edits of error types than the conventional methods.

Key Words: *Grammatical Error Correction, Automatic Evaluation Metric*

1 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。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）などを用いて評価する。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 (Dahlmeier and Ng 2012; Napoles et al. 2015) と用いない手法 (Napoles et al. 2016; 浅野 他 2018) 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 (Bryant and Ng 2015) 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 他 (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 (Sakaguchi et al. 2016)。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意味が異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価手法を構築するために、浅野 他 (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) (Devlin et al. 2019) を用い、各項目の人手評価値付きデータセットで再学習することで各評価モデルを最適化する。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、文法性や流暢性の評価値が付与され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ために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、データセットを作成する。

実験では、人手評価との相関および MAEGE (Choshen and Abend 2018) 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を正しく評価できていないのに対して、提案手法は全てのエラータイプの訂正を正しく評価でき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¹。
- 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を正しく評価できていることを示した。

2 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 (Dale and Kilgariff 2011) 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。Dahlmeier and Ng (2012) 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。Felice and Briscoe (2015) 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。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 et al. (2015) は機械翻訳の自動評価で使われる BLEU (Papineni et al. 2002) を文法誤り訂正のために改善した GLEU を提案した。BLEU は訂正文と参照文を用

¹ <https://github.com/tmu-nlp/TMU-GFM-Dataset>

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 et al. (2016) 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. 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 (Bryant and Ng 2015) 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 et al. (2016) 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. 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 et al. (2014) 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 他 (2018) 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. 文法性は GUG データセット (Heilman et al. 2014) 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 (Denkowski and Lavie 2014) 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 (Napoles et al. 2017) を用いてチューニングし, システムレベル・文レベルで Grundkiewicz et al. (2015) の評価データを用いてメタ評価している. 浅野 他 (2018) 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 他 (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最適化する. 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 他 (2018) と同様に実験する.

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 他 (2018) 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² (Heilman et al. 2014) がある. GUG データセットは, 学習者の書いた文に対して文法性の人手評価が付与されている. 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疑似誤り文に対して流暢性 (Acceptability) の人手評価が付与されたデータセット³ (Lau et al. 2015) 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.

² <https://github.com/EducationalTestingService/gug-data>

³ <https://clasp.gu.se/about/people/shalom-lappin/smog/experiments-and-datasets>

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。

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 (Banerjee and Lavie 2005)、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、機械翻訳の自動評価手法の Shared task である WMT 2019 Metrics Shared Task (Ma et al. 2019) 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 (Grundkiewicz et al. 2015; Sakaguchi et al. 2016; Napoles et al. 2016; 浅野 他 2018)。浅野 他 (2018) 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 and Abend (2018) 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。人手の訂正を用いることで、人手の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、人手の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のため、人手評価との相関だけでなく、人手の訂正に基づく MAEGE によるメタ評価も行う。人手評価との相関に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。

3 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を最適化する。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 (Devlin et al. 2019) を用いることを提案する。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、入力文の一部をマスクしマスクされた単語を当てるように学習を行うことで、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。

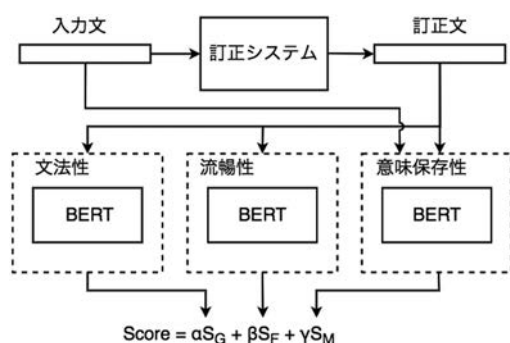


図 1 評価手法の全体像 (α, β, γ は各評価値の重み)

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 1 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。再学習時には回帰モデルの学習と同時に文符号化器の学習も行う。

訂正文の最終的な評価は、浅野 他 (2018) と同様に各項目の評価値の線形和で求める。

$$\text{Score} = \alpha \cdot \text{Score}_G + \beta \cdot \text{Score}_F + \gamma \cdot \text{Score}_M. \quad (1)$$

ここで、 Score_G 、 Score_F 、 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。

4 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 2013⁴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⁵。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 <https://www.comp.nus.edu.sg/~nlp/conll13st.html>

⁵ 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是不適切なため使用は避けた。

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⁶に公開されている中で最高性能のシステム (Zhao et al. 2019) 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 and Junczys-Dowmunt (2018) の実装⁷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⁸のうち part00 から part02 までの 30 億文を用いた。KenLM⁹ (Heafield 2011) を用いて言語モデルの学習を行った。単語クラスの学習には Word2Vec (Mikolov et al. 2013) を用いた。

RNN: Recurrent Neural Network に基づく系列変換モデル。実装には fairseq¹⁰ (Ott et al. 2019) 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 et al. (2015) に従った。

CNN: Convolutional Neural Network に基づく系列変換モデル。実装には fairseq¹⁰ (Ott et al. 2019) 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 and Ng (2018) に従った。

SAN: Self-Attention Network に基づく系列変換モデル。実装には fairseq¹⁰ (Ott et al. 2019) を用いた。設定は Vaswani et al. (2017) 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 et al. (2019) の実装¹¹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¹¹ (Lang-8, NUCLE, FCE) を、その他のモデルは Bryant et al. (2019) の訓練用データ¹² (Lang-8, NUCLE, FCE, W&I) を用いた。ただし、100 トークン以上の文を削除した後に、Byte Pair Encoding¹³ (Sennrich et al. 2016) を用い、結合回数を 30,000 として単語のサブワード化を行った。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 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 の評価を行っている。先行研究の報告

⁶ 2019 年 7 月時点。

⁷ <https://github.com/grammatical/baselines-emnlp2016/tree/master/train-2018>

⁸ <http://data.statmt.org/romang/gec-emnlp16/sim/>

⁹ <https://kheafeld.com/code/kenlm/>

¹⁰ <https://github.com/pytorch/fairseq>

¹¹ <https://github.com/zhawe01/fairseq-gec>

¹² <https://www.cl.cam.ac.uk/research/nl/bea2019st/#data>

¹³ <https://github.com/rsennrich/subword-nmt>

システム	本研究	先行研究	
SMT	34.20	32.30	(Mita et al. 2019)
RNN	35.50	33.76	(Mita et al. 2019)
CNN	33.04	31.96	(Chollampatt and Ng 2018)
SAN	35.57	36.20	(Mita et al. 2019)
SAN+Copy	40.89	—	—

表 1 CoNLL 2013 における各システムの $F_{0.5}$ 値

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 et al. (2014) 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 et al. (2015) 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 et al. (2016) 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 Turk¹⁴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, 全ての項目に 4 を回答させるダミー設問を設置した. 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」¹⁵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. 時給 7.25 ドルと見積もって作業を複数人に分割して依頼し, データ作成時点 (2019 年) のレート換算で合計約 10 万円を使用してデータセットを作成した.

人手評価値のヒストグラムおよび各項目間のピアソンの積率相関係数を図 2 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. また, 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

¹⁴ <https://www.mturk.com/>

¹⁵ 不完全な文や意味不明な文.

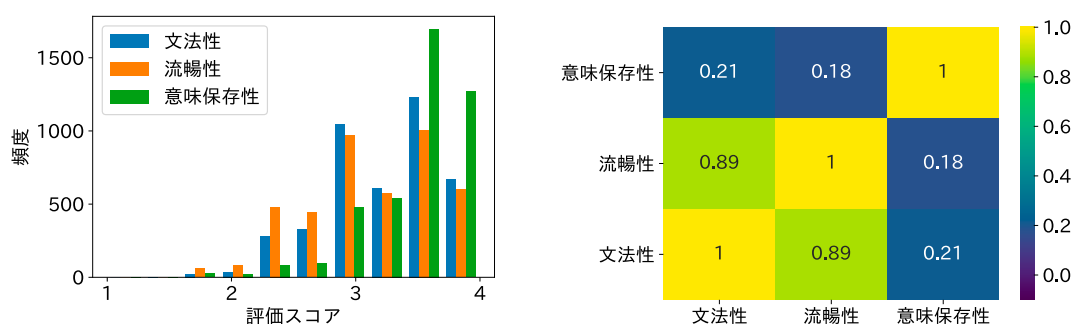


図 2 各項目のヒストグラム（左図）と各項目間の相関行列（右図）

入力文	This will <i>inversely</i> improve the <i>sale</i> of the shop.		
システム出力	This will <i>definitely</i> improve the <i>sales</i> of the shop.		
	文法性: 3.8	流暢性: 3.8	意味保存性: 1.6
入力文	The <i>increasing</i> longevity is due to fast development of the society so as the living pressure.		
システム出力	The <i>increase</i> in longevity is due to <i>the</i> fast development of society so as the living pressure.		
	文法性: 2.6	流暢性: 2.4	意味保存性: 3.8

図 3 実際の評価例

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 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、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。

5 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¹ および流暢性² 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 (Cer et al. 2017) のデータセット¹⁶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ついているのに対し, このデータセットは誤りを含まない文に対して評価がついている. 事前学習済みの BERT (bert-base-cased)¹⁷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. 検証データでスパイアマンの順位相関係数が最大となるモデルを選択した.

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 4 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 他 (2018) と同様に, Grundkiewicz et al. (2015) 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 et al. (2015) はそのランクか

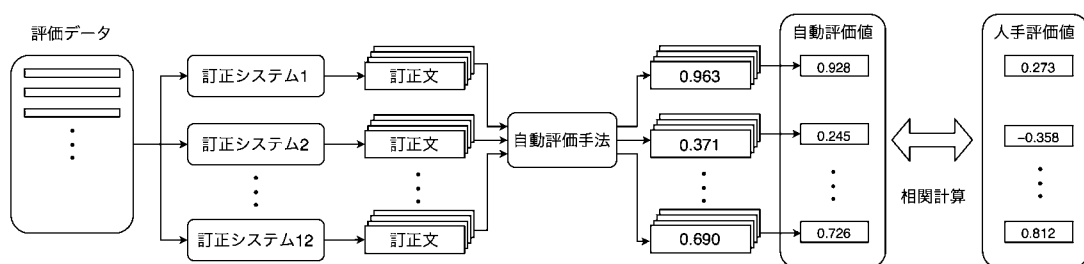


図 4 システムレベルでのメタ評価

¹⁶ <http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark>

¹⁷ <https://github.com/huggingface/transformers>

らレーティングアルゴリズムである True Skill (Herbrich et al. 2007) を用いてシステムレベルでの人手評価を計算した。浅野 他 (2018) と比較を行うために, Grundkiewicz et al. (2015) の Table 3c¹⁸ 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 (1) の重み α, β, γ は 浅野 他 (2018) と同様に JFLEG データセット (Napoles et al. 2017) 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 他 (2018) 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 他 (2018) と同様に, Grundkiewicz et al. (2015) 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 (2)) と WMT2017 (Bojar et al. 2017) で使われているケンドールの順位相関係数 τ (Kendall's τ) (式 (3)) 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。

$$Accuracy = \frac{\text{適切に評価したペア数}}{\text{人手評価の大小が異なるペア数}}. \quad (2)$$

$$Kendall's \tau = \frac{\text{適切に評価したペア数} - \text{逆に評価したペア数}}{\text{人手評価の大小が異なるペア数}}. \quad (3)$$

式 (1) の重み α, β, γ は Grundkiewicz et al. (2015) 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 and Abend (2018) らは, 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 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

¹⁸ 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。

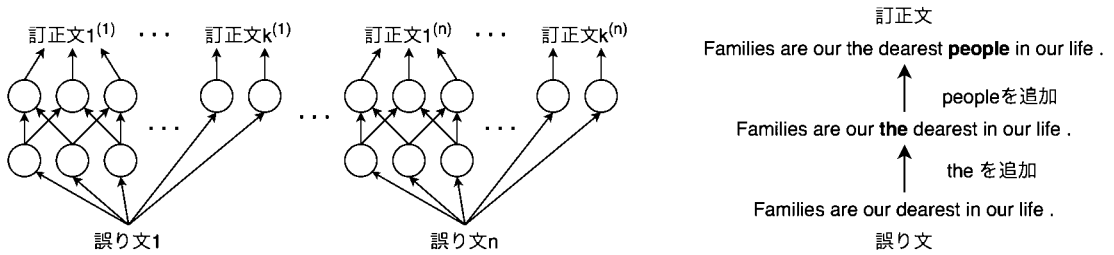


図 5 各誤り文に対する人手の訂正を元とした束（左図）と実際の例（右図）．各有向エッジは人手の訂正の適用を表し，訂正文 k は誤り文に対して k 人目の訂正者の全ての訂正を適用した文である．

価し，定義された順序関係を正解順位として比較を行いメタ評価を行う．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 and Abend (2018) を参照されたい．

本研究の実験では，CoNLL 2014 (Ng et al. 2014) のテストデータを使用する．実験は公開されている実装¹⁹を使用した，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 and Abend (2018) 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 and Abend (2018) と同様の設定で，ケンドールの順位相関係数 τ ²⁰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 and Abend (2018) では，2 人のアノテーションのうち 1 人が訂正を行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．

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 (Dahlmeier and Ng 2012) および GLEU (Napoles et al. 2015) 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 and Ng (2015) の追加の 8 つの参照文と Sakaguchi et al. (2016) 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 他 (2018) を用いる．

¹⁹ <https://github.com/borggr/EoE>

²⁰ 通常用いられる，全順序集合に対して定義された指標ではなく，半順序集合に合わせて Choshen and Abend (2018) が再定義した指標を用いている．

6 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 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, 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 他 (2018) のシステムレベルにおける結果は再実装したものをを用いた. 浅野 他 (2018) 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 3 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 et al. (2015) の評価データを用いたメタ評価の結果を示す. 浅野 他 (2018) のシステムレベルにおける結果は浅野 他 (2018) からの引用, 文レベルにおける結果は再実装したものをを用いている. 表 2 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 3 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

	項目	ピアソン	スピアマン
浅野 他 (2018)	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 2 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.

保存性単体の評価は高くなるのに対して, 総合的な評価では訂正しないシステムに対しては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 4 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 他 (2018) の結果は, システムレベルにおいては浅野 他 (2018) 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

	項目	システムレベル		文レベル	
		ピアソン	スピアマン	正解率	ケンドール
浅野 他 (2018)	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

表 3 Grundkiewicz et al. (2015) 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		
	ピアソン	スピアマン	重み ($\alpha:\beta:\gamma$)	正解率	ケンドール	重み ($\alpha:\beta:\gamma$)
M ²	0.674	0.720	—	0.464	0.294	—
GLEU	0.846	0.816	—	0.670	0.354	—
浅野 他 (2018)	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

表 4 システムレベル (左) と文レベル (右) のメタ評価

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。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 5 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すると、BERT を用いた手法が 浅野 他 (2018) 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

	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 他 (2018)	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

表 5 MAEGE によるメタ評価

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 2, 表 3), 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 4, 表 5). 各評価モデルを組み合わせた場合の, 再学習に用いるデータセットの違いによる性能の改善は大きくないが, 表 2, 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.

7 分析

7.1 評価例

M², GLEU, 浅野 他 (2018), 作成したデータで BERT の再学習を行った場合の提案手法による評価例を分析する. Grundkiewicz et al. (2015) の評価データに対する各評価手法を比較する. 表 6 に, 提案手法のみが正しく評価できた例を示す. M² 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 他 (2018) が正しく評価できていないのは, 浅野 他 (2018) が主に流暢性, 次に意味保存性を見ている (表 4) が,

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 .				
訂正文 1	There are a lot of <i>problems</i> that people may not realize .				
	人手評価	M ²	GLEU	浅野 他 (2018)	提案手法
	✓	0.556	0.586	0.949	0.913
訂正文 2	There are a lot of <i>the</i> disadvantages that people may not realize .				
	人手評価	M ²	GLEU	浅野 他 (2018)	提案手法
	×	0.556	0.630	0.977	0.826

表 6 提案手法のみが正しく評価できた例

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 他 (2018)	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 <i>children</i> .				
	人手評価	M ²	GLEU	浅野 他 (2018)	提案手法
	×	0.333	0.245	0.883	0.881

表 7 参照文を用いる手法のみが正しく評価できている例

流暢性の評価モデルは人手評価との相関が低い（表 2）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 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、適切に評価できしており、負ならば、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 (Ng et al. 2014) で指定されている 27 のエラータイプを使用する。

表 8 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²¹, 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, 連続文やコンマ区切りの誤り (Srun) や並列性の誤り (Spar) は他の手法に比べて平均差が大きく, 他の手法と比較してこ

エラータイプ	M^2	GLEU	浅野 他 (2018)	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 modifiers)	0.000	-0.162	0.021	0.025	0.033
Wa: (Acronyms)	0.000	0.285	0.176	0.113	0.237

表 8 エラータイプ分析

²¹ Choshen and Abend (2018) 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.

これらの誤りに対する感度が高いことがわかる。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、他手法と同様に感度が低いことがわかる。

8 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・意味保存性・流暢性の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。

謝 辞

本論文の実験をするにあたり、実装コードの提供をして頂いた浅野広樹氏、JFLEG の評価データで使用されているシステム出力を提供して頂いた坂口慶祐氏、Amazon Mechanical Turk で公開するタスクの確認を行って頂いた Siti Oryza Khairunnisa 氏に感謝します。本研究の一部は JSPS 科研費（若手研究，課題番号：JP20K19861）の助成を受けたものです。本研究の内容の一部は、The 28th International Conference on Computational Linguistics (Yoshimura et al. 2020) で発表したものです。

参考文献

- 浅野広樹, 水本智也, 乾健太郎 (2018). 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, **25** (5), pp. 555–576. [H. Asano et al. (2018). A Reference-less Evaluation Metric Based on Grammaticality, Fluency, and Meaning Preservation in Grammatical Error Correction. *Journal of Natural Language Processing*, 25 (5), pp. 555–576.].
- Banerjee, S. and Lavie, A. (2005). “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72.
- Bojar, O., Graham, Y., and Kamran, A. (2017). “Results of the WMT17 Metrics Shared Task.” In *Proceedings of the Second Conference on Machine Translation*, pp. 489–513.
- Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). “The BEA-2019 Shared Task on Grammatical Error Correction.” In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–75.
- Bryant, C. and Ng, H. T. (2015). “How Far are We from Fully Automatic High Quality Grammatical Error Correction?” In *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.
- Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). “SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation.” In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14.
- Chollampatt, S. and Ng, H. T. (2018). “A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction.” In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 5755–5762.
- Choshen, L. and Abend, O. (2018). “Automatic Metric Validation for Grammatical Error Correction.” In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1372–1382.
- Dahlmeier, D. and Ng, H. T. (2012). “Better Evaluation for Grammatical Error Correction.” In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 568–572.
- Dale, R. and Kilgariff, A. (2011). “Helping Our Own: The HOO 2011 Pilot Shared Task.” In *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 242–249.

- Denkowski, M. and Lavie, A. (2014). “Meteor Universal: Language Specific Translation Evaluation for Any Target Language.” In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pp. 376–380.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *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.
- Felice, M. and Briscoe, T. (2015). “Towards a Standard Evaluation Method for Grammatical Error Detection and Correction.” In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 578–587.
- Grundkiewicz, R. and Junczys-Dowmunt, M. (2018). “Near Human-Level Performance in Grammatical Error Correction with Hybrid Machine Translation.” In *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.
- Grundkiewicz, R., Junczys-Dowmunt, M., and Gillian, E. (2015). “Human Evaluation of Grammatical Error Correction Systems.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 461–470.
- Heafield, K. (2011). “KenLM: Faster and Smaller Language Model Queries.” In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 187–197.
- Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., and Tetreault, J. (2014). “Predicting Grammaticality on an Ordinal Scale.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 174–180.
- Herbrich, R., Minka, T., and Graepel, T. (2007). “TrueSkillTM: A Bayesian Skill Rating System.” In Schölkopf, B., Platt, J. C., and Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 569–576. MIT Press.
- Lau, J. H., Clark, A., and Lappin, S. (2015). “Unsupervised Prediction of Acceptability Judgments.” In *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.
- Luong, T., Pham, H., and Manning, C. D. (2015). “Effective Approaches to Attention-based Neural Machine Translation.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421.
- Ma, Q., Wei, J., Bojar, O., and Graham, Y. (2019). “Results of the WMT19 Metrics Shared

- Task: Segment-Level and Strong MT Systems Pose Big Challenges.” In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 62–90.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). “Efficient Estimation of Word Representations in Vector Space.” In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Mita, M., Mizumoto, T., Kaneko, M., Nagata, R., and Inui, K. (2019). “Cross-Corpora Evaluation and Analysis of Grammatical Error Correction Models—Is Single-Corpus Evaluation Enough?” In *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.
- Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). “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.
- Napoles, C., Sakaguchi, K., and Tetreault, J. (2016). “There’s No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction.” In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2109–2115.
- Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). “JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction.” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 229–234.
- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). “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.
- Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). “fairseq: A Fast, Extensible Toolkit for Sequence Modeling.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “BLEU: a Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.
- Sakaguchi, K., Napoles, C., Post, M., and Tetreault, J. (2016). “Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality.” *Transactions of the Association for Computational Linguistics*, 4, pp. 169–182.

- Sennrich, R., Haddow, B., and Birch, A. (2016). “Neural Machine Translation of Rare Words with Subword Units.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). “Attention is All you Need.” In *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). “Optimizing Statistical Machine Translation for Text Simplification.” *Transactions of the Association for Computational Linguistics*, 4, pp. 401–415.
- Yoshimura, R., Kaneko, M., Kajiwara, T., and Komachi, M. (2020). “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.
- Zhao, W., Wang, L., Shen, K., Jia, R., and Liu, J. (2019). “Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data.” In *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. 156–165.

略歴

吉村 綾馬：2019 年首都大学東京（現東京都立大学）システムデザイン学部システムデザイン学科情報通信システムコース卒業。同年、首都大学東京システムデザイン研究科情報科学域博士前期課程に進学。2021 年博士前期課程修了。同年、株式会社ユーザーローカルに入社。現在に至る。

金子 正弘：2016 年北見工業大学工学部情報システム工学科卒業。同年、首都大学東京システムデザイン研究科博士前期課程に進学。2018 年博士前期課程修了。同年、首都大学東京システムデザイン研究科博士後期課程に進学。2019 年より日本学術振興会特別研究員 (DC2) を経て、2021 年博士後期課程修了。博士（情報科学）。同年より東京工業大学情報理工学院研究員。現在に至る。

梶原 智之：愛媛大学大学院理工学研究科助教。2013 年長岡技術科学大学工学部電気電子情報工学課程卒業。2015 年同大学大学院理工学研究科修士課程修了。2018 年首都大学東京大学院システムデザイン研究科博士後期課程修了。博士（工学）。大阪大学データビリティフロンティア機構の特任助教を経て、2021 年より現職。情報処理学会，人工知能学会，言語処理学会，ACL 各会員。

小町 守：2005 年東京大学教養学部基礎科学科科学史・科学哲学分科卒業。

2007 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了. 2008 年より日本学術振興会特別研究員 (DC2) を経て, 2010 年博士後期課程修了. 博士 (工学). 同年より同研究科助教を経て, 2013 年より首都大学東京システムデザイン学部准教授. 大規模なコーパスを用いた意味解析および統計的自然言語処理に関心がある. 情報処理学会, 人工知能学会, 言語処理学会, ACL 各会員.

(2020 年 9 月 1 日 受付)

(2020 年 12 月 1 日 再受付)

(2021 年 1 月 19 日 採録)