学習者の誤り傾向を考慮した擬似データを用いた文法誤り訂正

高橋 悠進　　勝又 智　　小町 守
首都大学東京
{takahashi-yujin, katsumata-satoru}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

近年，文法誤り訂正の研究では，英語だけでなく，様々な言語の学習者支援タスクとして盛んに研究されている．文法誤り訂正の訓練データが十分にない言語に対しても，事前学習時に擬似データを組み込む手法が盛んに用いられている．具体的には，擬似データで事前学習されたモデルを学習者データで fine-tuning することで，高い訂正性能を示している [16, 7, 5, 9, 4].

学習者データ不足の問題を解決するための擬似データを生成するいくつかの手法が提案されている．ランダムな誤りによる擬似データの生成は，無限にデータを扱える場合は全ての誤りをカバーできる．しかし，実際には擬似データとして扱えるデータの量は限られているため，ランダムな誤りで扱える誤りは限られてくる．つまり，限られたデータ量の擬似データ内で現実的な誤りを生成する必要がある．

そこで，本研究では学習者のエラータイプを考慮した擬似データを作成し，この擬似訓練データが訂正モデルに与える影響を調査する．我々は，英語とロシア語の評価データと同じドメインの開発データから学習者の誤り傾向を分析して，英語とロシア語でそれぞれ擬似データを構築した．

実験では，ランダムな誤りによる擬似データと，開発データから求めた学習者の誤り傾向上位のエラータイプを考慮した擬似データでそれぞれ事前学習し，学習者データで fine-tuning したモデルの訂正の性能を $F_{0.5}$ で比較した．また英語とロシア語で実験することで，英語，ロシア語でそれぞれ外部知識として導入した方が良いエラータイプが存在することを明らかにした．

2 関連研究

文法誤り訂正の擬似データを生成する手法はいくつか提案されている．Zhao ら [16] は誤りのないクリーンな文に対して，ランダムに単語の置換，削除，追加，およびシャッフルを行う．Kiyono ら [7] の擬似データ生成手法の一つは，与えられた文の各単語に対してランダムに単語のマスク，削除，追加および保持を行う．これらの手法では対象データの誤り方や言語ごとの違いを考慮していない．表 1 は英語とロシア語の学習者コーパスのエラータイプの統計を示している [13]．表 1 によると，英語の学習者は冠詞や語句選択の誤りが多いが，ロシア語の学習者は，スペリング誤り，挿入に次いで語形変化に由来した誤りが多いことを報告している．

English (CoNLL 2013)		Russian (RULEC-GEC dev)	
Error type	Rate (%)	Error type	Rate (%)
Article/Determiner	19.9	Spelling	22.8
Collocation/Idiom	12.5	Insert	13.2
Noun number	11.4	Noun case	10.2
Preposition	8.98	Replace	9.99
Word form	6.56	Delete	9.58

表 1: 英語とロシア語の開発データのエラータイプの統計

その中で，Grundkiewicz ら [5] は誤りのない文に対して，教師なしの spellchecker に基づいた confusion set を構築し，単語の置換候補を制限することで，より現実的な誤りを含む擬似データを生成した．実際に誤らせるには単語 *cor* に対する誤り *err* の確率 $P(err|cor)$ が得られれば正しい擬似データが生成できる．しかし，真の擬似誤り生成モデル $P(err|cor)$ を得るのは困難なため，Grundkiewicz らは手法では条件付き確率 $P(cor|err)$ を spellchecker に基づいた分布としている．一方で，我々の手法は，$P(err|cor)$ に対して，特に考慮したエラータイプを spellchecker ではなく，外部知識を用いて分布に対して制約を与える．これによって現実的な誤りを考慮するようにした．

Náplava と Straka [9] は Grundkiewicz らと同様の擬似データの作成方法で，英語，ロシア語，ドイツ語，チェコ語で実験を行った．さらに，Grundkiewicz と Junczys-Dowmunt [4] は spellchecker に基づいた confusion set 内の大文字小文字を一貫させ，OOV を含めた confusion set を構築した．この手法は，他にも文字

単位で，ランダムに削除，追加，置換，交換を行う．しかし，単語の表層情報を主に考慮しているため，現実的でない誤りも擬似データに含まれる問題がある．

Kasewa ら [6] は，文法誤り検出のタスクにおいて，対訳データから擬似誤り生成モデル $P(err|cor)$ の分布を決定している．本研究では，より困難な文法誤り訂正のタスクにおいて，$P(err|cor)$ の分布を外部知識に基づき決定した．

単言語データに対してノイズを入れる際に外部知識を用いた研究として，Sun ら [14] は，入力文のいくつかの単語にマスクし，その単語を他の単語から予測する目的関数に対して，固有表現ごとにマスクをかけて事前学習を行うことで，中国語の複数の自然言語処理タスクで最高性能を達成している．本研究では，単言語データに対して擬似誤りを入れる際に，開発データの上位エラータイプに従い，外部知識を組み込む．

CoNLL-2014 Shared Task [10] の文法誤り訂正のタスクにおいては，複数の研究で訂正候補の集合として confusion set を用いた [12, 15]．本研究では，エラータイプを考慮した擬似誤りを作る際に，誤った語の候補の集合として confusion set を用いることでランダムより現実的な擬似誤りを作成する．

3 擬似誤り作成方法

学習者の誤りやすい傾向を求め，エラータイプの割合が上位の誤りを考慮した擬似データを作成する手法を述べる．図 1 にエラータイプに基づいた擬似誤り作成方法の例を示す．前置詞の誤りを例とすると，$P(err|cor$ = “at”) における誤り err は $err \in \{\text{about, by, for, from, in, of, with, on, to}\}$ として，外部知識を用いて制約を与える．このようにして，英語とロシア語のそれぞれで考慮したエラータイプの誤りに対して，誤り生成 $P(err|cor)$ の分布を一様分布にして擬似誤りを生成する．

表 1 より英語の場合は，冠詞や限定詞，コロケーションやイディオム，名詞の単数複数に関する誤り，前置詞の誤りなどがある．冠詞や限定詞に関する誤りは，ランダムの場合は置換される候補が語彙全体になるが，候補の集合を他の冠詞や限定詞に制限することで考慮する．名詞の単数複数の誤りは，単語の品詞が名詞だった場合[1] に，その単語の単数複数を入れ替えることで誤りを作成できる [1]．前置詞の誤りは，Bryant と Briscoe [2] で使用された上位 10 個の最頻出の前置詞の集合を置換先の候補集合として考慮する．コロケーションやイディオムの誤りについては，候補集合を作成することが困難なため今回は考慮しなった．

[1]名詞の可算，不可算は考慮していない．

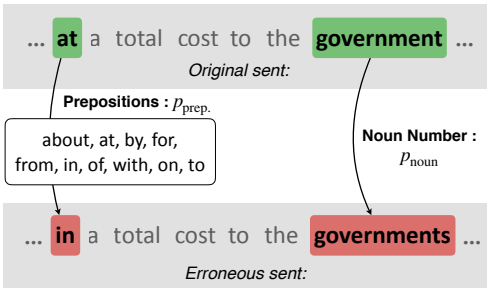

図 1: エラータイプに基づいた擬似誤りの例

Language	Dataset	Corpus	Sentences
English	One Billion Corpus	mono	10,000,000
	NUCLE	train	55,231
	CoNLL-2013	dev	1,381
	CoNLL-2014	test	1,312
Russian	Russian News Crawl	mono	10,000,000
	Lang-8 + RULEC-GEC	train	54,132
	RULEC-GEC dev	dev	2,500
	RULEC-GEC test	test	5,000

表 2: 本実験で使用したデータ

ロシア語の場合は，スペリング誤り，挿入，名詞の格変化に関する誤りが多い．スペリング誤り，挿入の誤りは先行研究の手法 [9, 4] で擬似データに含めることが可能である．名詞の格変化は，辞書を利用する．ある名詞の単語が辞書に含まれていた場合に，置換候補の集合をその単語の格変化の集合とする．

本研究では，英語とロシア語で共通の擬似誤りとして Grundkiewicz らの手法に従って spellchecker に基づいた confusion set を置換候補の集合として単語を置換し，文字単位の誤りも含めた．さらに我々は，この誤りにエラータイプに基づく誤りを追加した．

4 実験

データ 表 2 に本研究で使用したデータを示す．文法誤り訂正データとして英語の訓練データに NUCLE，開発データに CoNLL-2013 [11]，評価データに CoNLL-2014 [10] を用いた．今回は，少量データとして英語とロシア語でそれぞれ 55,000 文程度を想定したため，訓練データとして Lang-8 [8] を用いなかった．ロシア語の訓練データに Lang-8 と RULEC-GEC，開発データに RULEC-GEC dev，評価データに RULEC-GEC test [13] を用いた．擬似データを作成する際に，誤りを付与する単言語データとして，英語は One Billion Corpus

Method	CoNLL-2014 (En)			RULEC-GEC test (Ru)			That is the most essential elements to overcome the challenges. (Source)
	Prec.	Rec.	$F_{0.5}$	Prec.	Rec.	$F_{0.5}$	That is the most essential element to overcome the challenges. (Gold)
Baseline	51.3	1.99	8.60	14.4	2.74	7.79	That is the most essential elements to overcome the challenges.
Random	**66.4**	7.66	26.2	34.5	5.75	17.3	These are the most essential elements to overcome the challenges.
Random + Spell.	63.6	8.31	27.3	**44.8**	10.9	27.7	That is the most essential elements to overcome the challenges.
Ours	64.0	**8.72**	**28.2**	41.1	**12.4**	**28.1**	That is the most essential element to overcome the challenges.
w/o knowledge	–	–	–	44.1	9.75	25.9	

表 3: 文法誤り訂正の結果 (Ours : Random + Spell. + Error-type) と出力例

[2]，ロシア語は Russian News Crawl [3] を用いた．

実験設定　本研究では文法誤り訂正モデルとして，Transformer をベースにコピー機構を追加した Zhao ら [16] の実装を用い，ハイパーパラメータとして，事前学習時の max-epoch を 3，学習時の max-epoch を 15 にし，それ以外の数値を全て彼らと同様の値設定した．

擬似データは，各言語の単言語コーパスに 3 節で示したように擬似誤りを組み込むことで作成した．本研究では，単言語コーパスに組み込む擬似誤りの影響を調査するために，以下の 3 つのベースラインと比較した; 擬似データの必要性を調査するためのベースラインとして，擬似データ無し (Baseline)，ランダムに選択された単語に対して単語の削除，追加，置換，文自体のシャッフルを行う誤りを加えた擬似データ (Random)，ランダムの追加と削除の手法に加え，置換先の候補の集合として spellchecker に基づいた confusion set を用いた擬似データ (Random + Spell.)．これらの手法と，Random + Spell. の手法に加えエラータイプに基づいた擬似誤りを付与した擬似データ (Ours) を比較した．

本研究では，以下に示す各パラメータは，全体の擬似誤りに含まれる各擬似誤りの割合を表し，開発データのスコアをもとにパラメータを設定した．

英語の場合には，ランダムで単語の削除，挿入の擬似誤りのパラメータは $(p_{\text{delete}}, p_{\text{insert}}) = (0.05, 0.05)$，spellchecker による confusion set と文字単位の擬似誤りのパラメータは $(p_{\text{spell}}, p_{\text{char}}) = (0.2, 0.2)$，考慮したエラータイプである冠詞や限定詞，名詞の単数複数，前置詞の擬似誤りのパラメータを $(p_{\text{art./det}}, p_{\text{noun}}, p_{\text{prep.}}) = (0.1, 0.3, 0.1)$ とした．

ロシア語の実験で，公開されている語の活用に関して展開済みの辞書 を使用した．[4] 英語と共通する擬似誤りのパラメータとして $(p_{\text{delete}}, p_{\text{insert}}, p_{\text{spell}}, p_{\text{char}}) = (0.1, 0.1, 0.32, 0.16)$，格変化を考慮した辞書を用いた擬似誤りのパラメータは $p_{\text{dict}} = 0.32$ とした．擬似誤りの種類による影響を確認するために，手法ごとに擬似誤りの数の合計に大きく差が生じないように統一した．評価手法としては，英語は CoNLL-2014，ロシア語は RULEC-GEC test に対して，適合率，再現率，$F_{0.5}$ スコア [3] で評価した．

実験結果　文法誤り訂正の実験結果を表 3 に示す．1 つ目のベースライン Baseline は，その他の手法と比較すると，英語，ロシア語共通して全ての値が上昇していることがわかる．これは Náplava と Straka [9] と同様の結果であった．

2 つ目のベースライン Random と Baseline を比較すると，英語の場合，ロシア語ほど大きな性能向上はなく，他の手法と比較しても Precision が最も高い値である．ロシア語の場合，英語よりも $F_{0.5}$ の向上が小さい．ロシア語において，ランダムで擬似誤りを作成することは，英語と比べて効果が得られないことがわかる．また，Random と提案手法と比較すると，ロシア語において，大きく $F_{0.5}$ が向上している．ただし，英語とロシア語で共通して Recall が小さい．

3 つ目のベースライン Random + Spell. と提案手法を比較すると，ロシア語では Precision が下がってしまうが，提案手法の方が僅かに性能が向上していることがわかる．提案手法は，英語とロシア語で共通して Recall が上昇していることがわかる．

w/o knowledge は，ロシア語の場合に，語の活用に関する辞書や spellchecker の外部知識を用いず，Random の手法と文字単位での誤りのみで，擬似誤りデータを作成した場合の結果である．これにより，外部知識を利用した擬似誤り作成方法が効果的であることを示している．

5　分析

エラータイプごとの Recall　表 4 に提案手法と Random + Spell. におけるエラータイプごとの Recall の比

[2]https://www.statmt.org/lm-benchmark/
[3]http://www.statmt.org/wmt18/translation-task.html
[4]http://opencorpora.org/?page=downloads

	En (test)			Ru (test)	
Error type	Ours	Spell.	Error type	Ours	Spell.
*Art/Det	**7.67**	5.29	Spelling	25.7	**28.2**
Collocation/Idiom	0.99	**1.36**	Insert	**11.6**	10.8
*Noun number	**24.6**	17.0	*Noun case	**17.1**	9.39
*Preposition	**4.07**	3.23	Replace	**5.00**	0.00
Word form	9.14	**10.3**	Delete	**11.0**	9.97

表 4: エラータイプごとの Recall (* は考慮したエラータイプを示す)

較を示した．英語の結果から，提案手法として考慮した，冠詞，名詞の単数複数，前置詞の誤りの Recall が向上している．その中でも，名詞の単数複数が最も高く向上している．ロシア語の場合，スペリング誤り以外のエラータイプにおいて Recall が向上していることがわかる．特に，提案手法で考慮した名詞の格変化の誤りは，他のエラータイプと比べて大きく向上している．$P(err|cor)$ について，分布に制約を加えており，Spellchecker と比較して現実的な誤りを生成していると考えられる．そのため，実際の誤りに対するカバー率も提案手法の方が多く，Recall が向上していると考えられる．このことより，エラータイプを考慮して擬似データを構築することが有効であることがわかる．

しかし，今回考慮したエラータイプである冠詞，前置詞の誤りは名詞の単数複数の誤りよりも Recall の向上が小さい．これは名詞の単数複数の誤りのパラメータが，他 2 つの誤りよりも多いことや，名詞の単数複数の誤りが 2 値分類であるのに対して，冠詞，前置詞の誤りは多値分類であることから当てることが難しいため，Recall の向上が小さいと考えられる．

出力例　英語の名詞の単数複数の誤りのみを含む文に対する各実験設定の出力例を表 3 に示す．赤字になっている単語は，誤っている箇所，青字になっている単語は，原文から正しく訂正されている箇所を表している．

Baseline と Random + Spell. の出力は，Source の文をそのまま出力してしまっている．Random の出力は，訂正すべきでない箇所を訂正してしまい，文の意味が異なった出力になっている．それらに対して，提案手法の出力は，訂正すべき名詞の単数複数の誤りの箇所を正しく訂正することができている．

6　おわりに

本研究は，文法誤り訂正タスクにおいて事前学習に用いる擬似データを構築する際に，学習者の誤りやすい傾向を考慮して，擬似誤りを作成した手法の影響を調査した．提案手法は，既存手法より Recall を向上させることで，$F_{0.5}$ を向上させた．分析から，開発データのエラータイプを考慮して作成した擬似誤りに対して，Recall が向上することを明らかにした．本研究は，外部知識を用いてエラータイプに制約を設けることで，より現実的な誤りを生成できたが，誤りを生成する際に重みは定式化しておらず，現実的な誤り分布にはなっていない．そのため，今後の研究ではエラータイプの統計を考慮することで現実的な誤り分布を再現できるようにしたい．

参考文献

[1] Chris Brockett, William B. Dolan, and Michael Gamon. Correcting ESL errors using phrasal SMT techniques. In *ACL*, 2006.

[2] Christopher Bryant and Ted Briscoe. Language model based grammatical error correction without annotated training data. In *BEA*, 2018.

[3] Daniel Dahlmeier and Hwee Tou Ng. Better evaluation for grammatical error correction. In *NAACL-HLT*, 2012.

[4] Roman Grundkiewicz and Marcin Junczys-Dowmunt. Minimally-augmented grammatical error correction. In *W-NUT*, 2019.

[5] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *BEA*, 2019.

[6] Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. Wronging a right: Generating better errors to improve grammatical error detection. In *EMNLP*, 2018.

[7] Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP*, 2019.

[8] Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. The effect of learner corpus size in grammatical error correction of ESL writings. In *COLING*, 2012.

[9] Jakub Náplava and Milan Straka. Grammatical error correction in low-resource scenarios. In *W-NUT*, 2019.

[10] 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 *CoNLL*, 2014.

[11] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 shared task on grammatical error correction. In *CoNLL*, 2013.

[12] Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. The Illinois-Columbia system in the CoNLL-2014 shared task. In *CoNLL*, 2014.

[13] Alla Rozovskaya and Dan Roth. Grammar error correction in morphologically rich languages: The case of Russian. *TACL*, 2019.

[14] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv preprint arXiv:1904.09223*, 2019.

[15] Jian-Cheng Wu, Tzu-Hsi Yen, Jim Chang, Guan-Cheng Huang, Jimmy Chang, Hsiang-Ling Hsu, Yu-Wei Chang, and Jason S. Chang. NTHU at the CoNLL-2014 shared task. In *CoNLL*, 2014.

[16] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *ACL*, 2019.