言語間での転移学習を用いたロシア語文法誤り訂正

山下 郁海　勝又 智　金子 正弘　Imankulova Aizhan　小町 守
首都大学東京
{yamashita-ikumi, katsumata-satoru, kaneko-masahiro, imankulova-aizhan}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

1　はじめに

文法誤り訂正は入力文として文法的に誤った文が与えられ，その文を文法的に正しく訂正するタスクである．大規模データが存在する英語においてさまざまな文法誤り訂正手法が提案され高い精度を達成している．近年，英語以外のロシア語やチェコ語などの言語においても研究が行われ始めている [7, 11]．しかし，これらの言語では学習データが小規模でしか存在しない．

小規模な学習データしか存在しない問題に対処するために，さまざまなタスクで他言語データを活用する試みが進んでいる [6, 14]．一方で，文法誤り訂正では他言語を活用した研究がほとんど行われておらず，文法的な正誤を言語をまたいで転移させ精度向上することが可能であるか明らかにされていない．しかし，ある程度類似した言語，例えば，同じ語族に属するロシア語とチェコ語のような言語間では文法的な正誤が転移可能なのではないかと考えられる．表 1 に示すのは，「妹」を意味する単語の日本語，ロシア語，チェコ語の格変化の一部である．日本語では主格と属格の変化を単語の変化ではなく，別に助詞を用いて行っているのに対し，ロシア語とチェコ語は単語の活用で行っていることがわかる．このような事例において文法の正誤が転移可能なのではないかと考えられる．

本稿では，転移学習を用いて他言語の情報を考慮することで転移先言語の文法誤り訂正モデルの性能を向上可能であることを示す．そして，類縁関係にある他言語を用いることで特に類似した文法項目に関する文法誤りの訂正精度が向上することも明らかにする．他言語を転移学習するために他言語と転移先言語のコーパスで学習された多言語表現モデルを用いる．本研究では，転移元言語をチェコ語，英語と日本語とし，転移先言語をロシア語とした．我々の転移学習モデルは各言語に関する特別な文法知識などを必要としない．

主格	日本語	私 の 妹 が 住ん で いる 家
	ロシア語	Дом, где живет моя сестра
	チェコ語	Dům, kde bydlí moje sestra
属格	日本語	私 の 妹 の 家
	ロシア語	Дом моей сестры
	チェコ語	Dům mé sestry

表 1: チェコ語とロシア語で類似した格変化

2　関連研究

近年の文法誤り訂正の手法は，そのほとんどが Encoder-Decoder モデルを用いており，大規模なデータが学習に必要となる．文法的に正しい文から文法的に誤った文を作成することは容易であるため，大規模単言語コーパスから疑似データを作成し学習に活用する研究が盛んに行われている．小規模言語における文法誤り訂正でも同様に疑似データが用いられている [7, 11]． これらの研究は，我々と同様に小規模言語における文法誤り訂正の性能向上が目的であるが，単一言語の情報しか用いていない．

言語間の情報を用いた研究は機械翻訳で盛んに行われている．Zoph ら [14] は，ニューラル機械翻訳において， 大規模に学習データが存在する言語対で学習したモデルを，小規模な言語対でファインチューニングする手法を提案した．また，複数言語を 1 つのモデルで学習することで学習データが存在しない言語対の翻訳を行うことが可能であることが知られている [6]．これらの研究は，文法の正誤が重要となる文法誤り訂正とは違い，言語間の意味的な情報がより重要となる機械翻訳タスクが対象である．

3　言語間での転移学習を用いた文法誤り訂正

3.1　多言語表現モデルを用いた転移学習

本研究の目標は，文法誤り訂正において言語間で転移学習を行うことで共通した単語の活用や変形のような文法情報を転移させることである．そのためには，

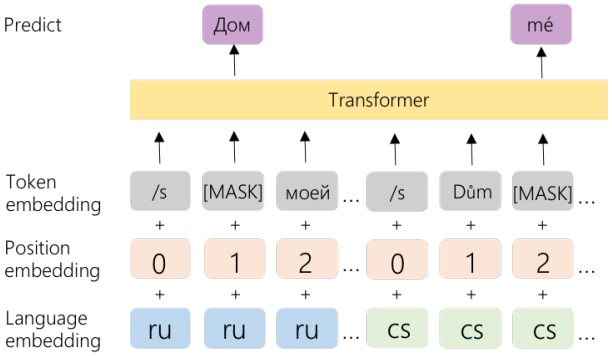

図 1: TLM のネットワーク構造

各言語の情報の学習や，言語間で対応する情報を学習する必要がある．そこで，我々は転移元言語と転移先言語の単言語または対訳コーパスを用いて学習した多言語表現モデルで文法誤り訂正モデルを初期化する．

多言語表現モデルの学習には MLM (Masked Language Modeling) と TLM (Translation Language Modeling) を用いる [3]．TLM の構造を図 1 に示す．MLM と TLM の構造は基本的に同じであり，3 種類の embedding を足し合わせたものを Transformer [13] に入力し学習を行う．Language embedding には言語の区別の情報が入っており，Position embedding にはトークンの文中の位置に関する情報が，Token embedding にはトークン自体に関する情報が入っている．

MLM と TLM の学習時の違いは学習データに対訳データを用いるかどうかである．MLM の学習は単数または複数の単言語データに対して行う．入力として 1 文が与えられ，入力文中のいくつかのトークンをマスクし，そのマスクされたトークンを予測することで言語表現を獲得する．TLM は MLM を対訳コーパスも用いて学習するように拡張する．入力として対訳関係にある文対が与えられ，その文対を結合し 1 つの系列とし，MLM と同じように一部のトークンをマスクし予測することで多言語表現を獲得する．MLM や TLM を用いることで，言語間にまたがった言語表現を学習することが可能になる．

我々は MLM や TLM を用いて次のような手順で転移学習を行った．(1) 転移元言語と転移先言語のコーパスを用いて事前学習された MLM や TLM で文法誤り訂正モデルを初期化する．(2) 転移先と転移元の言語の学習者データで文法誤り訂正モデルを学習する．(3) 転移先言語の学習者データのみを使って文法誤り訂正モデルをファインチューニングする．

	Ru	En	Cs	Ja	学習	開発	評価
対訳							
TED	✓	✓			80K	1.3K	1.3K
TED	✓		✓		80K	1.3K	1.3K
TED	✓			✓	80K	1.3K	1.3K
単言語							
News（15-18）	✓				33M	-	-
News（18）	✓	✓	✓		1.2M	2.5K	2.5K
Wikipedia				✓	1.2M	2.5K	2.5K
学習者							
RULEC	✓				5K	2.5K	5K
Lang-8-Ru	✓				49K	-	-
NUCLE		✓			40K	-	-
CoNLL		✓			-	1.4K	-
AKCES			✓		40K	2.5K	-
Lang-8-Ja				✓	40K	-	-
NAIST				✓	-	3.3K	-

表 2: 対訳・単言語・学習者データの文数

3.2 転移学習に用いる言語

今回の実験では，文法誤り訂正の転移先言語としてロシア語 (Ru) を，転移元言語としてチェコ語 (Cs)，英語 (En)，日本語 (Ja) を用いた． 各転移元言語の文法をロシア語の文法と比較した際の特徴を以下に示す．

チェコ語：ロシア語と同じスラヴ語族に属する言語であり，使われている文字は異なるが，形容詞や名詞の格変化などロシア語との文法的共通点が多く，ロシア語に類似した言語と言える．

英語：ロシア語とは語族が違い，使われている文字も異なるが，基本的な文の構造に SVO が多いことや，主語の単数・複数により動詞が変化することなど，いくつかの共通点が挙げられる．

日本語：ロシア語とは語族も，文字も異なり，文の構造なども違うため，今回用いる言語の中で最もロシア語から遠い言語である．

4 実験

4.1 データセット

実験に用いたデータの概要を表 2 に示す．本研究では，MLM と TLM の学習のために，単言語データとして WMT-2019 の News Crawl と日本語の Wikipedia データを，対訳データとして TED talks [1] を，また，文法誤り訂正モデルの学習のために，誤文と正文が対になった学習者コーパス RULEC-GEC [11]，NUCLE [5]，AKCES-GEC [7]，および Lang-8 を，そして後述する文法誤り訂正モデルの Re-ranking に用いる言語モデルの学習のために Russian News Crawl（2015-2018）を用いた．また，MLM の開発データ，評価データは，学習データを除いたそれぞれの単言語データか

ら抜粋したものを，TLM の開発データ，評価データは，TED talks に付属しているものを用い，文法誤り訂正の開発データ，評価データは，ロシア語とチェコ語はそれぞれのコーパスに付随するものを，英語は CoNLL-13 [9] のデータを，日本語は NAIST 誤用コーパス [10] のものを，それぞれ用いた．なお，TED talks のデータは元々のデータである英語への翻訳データから各言語で対応する文対を抽出して再構成したものであり，News Crawl，Wikipedia，Lang-8，NUCLE のデータは元のデータからそれぞれ表 2 に示す文数を抽出したものである．

全てのデータは fastBPE [12] を用いて BPE (Byte Pair Encoding) 処理を施し subword 化した後に使用した．また，ロシア語のデータは BPE の前に pyspellchecker[1] を用いてスペルチェックを行った．

4.2 実験設定

TLM と MLM，文法誤り訂正モデルには全て Guillaume ら [3] と同じ Transformer を用い，モデルの層数は 6 層，バッチサイズは 32 として，それ以外のパラメータはデフォルトのものを用いた．提案手法との比較に用いるベースラインには，ロシア語の学習者データのみを用いて学習したモデル，MLM や TLM を用いずに 2 言語で学習したモデルを用いた．また，開発データに対しては GLEU [8] を用いてモデル選択を行い，RULEC-GEC のテストデータに対しては，適合率，再現率，$F_{0.5}$ [4] と，GLEU で評価した．なお，評価を行う前に全てのモデルの出力に Chollampatt ら [2] と同じ Re-ranking を行なっている．

実験には Ru の学習者データのみを用いて学習したモデルと JaRu，EnRu，CsRu の各言語の学習者データを結合し学習したモデル，ロシア語の単言語データのみで事前学習された MLM で初期化したモデル，各言語の単言語データを結合し事前学習した MLM で初期化したモデル，各言語の単言語と対訳データで事前学習された TLM により初期化されたモデルを用いる．

4.3 実験結果

表 3 はロシア語文法誤り訂正の結果である．まず，MLM または TLM で初期化を行っていない設定では，全ての言語対でロシア語のみで学習したベースラインと比べて GLEU は高いが $F_{0.5}$ は低いことがわかる．また，適合率と再現率を見ると適合率が上がっているものの再現率は下がっていることがわかる．これによ

[1] https://github.com/barrust/pyspellchecker

モデル	P	R	$F_{0.5}$	GLEU
Ru のみ	22.82	14.27	20.38	68.34
Ja→Ru	23.02	12.81	19.86	69.61
En→Ru	23.74	12.17	19.95	69.86
Cs→Ru	22.96	13.21	20.01	69.50
MLM (Ru のみ)	22.58	**22.32**	22.53	68.25
MLM (Ja→Ru)	24.15	16.70	22.15	70.87
MLM (En→Ru)	25.05	19.91	23.82	70.69
MLM (Cs→Ru)	26.62	18.30	24.41	71.27
TLM (Ja→Ru)	22.87	17.22	21.62	69.80
TLM (En→Ru)	25.09	19.48	23.72	70.89
TLM (Cs→Ru)	**26.84**	20.31	**25.22**	**71.75**

表 3: ロシア語文法誤り訂正の結果

誤りタイプ	Ru のみ	Ja→Ru	En→Ru	Cs→Ru
Adj:Case	5.30	10.61	14.39	**15.15**
Noun:Case	8.09	15.90	15.90	**20.23**
Verb:Num/Pers	9.03	15.48	17.42	**22.58**

表 4: 誤りタイプごとの各モデルの再現率．Ja→Ru，En→Ru，Cs→Ru はそれぞれ TLM を用いたモデルの結果である．

り，単に他言語の学習データを追加するだけでは誤り訂正の性能は改善しないと考えられる．

次に，ロシア語のみの MLM を用いたモデルはベースラインのモデルよりも $F_{0.5}$ は高いが，適合率が下がって再現率が上がっていることが読み取れる．また，GLEU のスコアも MLM を用いない場合と比べて下がっていることがわかる．このことから，単言語の MLM だけでは，MLM を用いない場合と同じく，性能が十分には改善しないと考えられる．

続いて，2 言語で事前学習した MLM を用いた実験結果は MLM を用いない場合と比べてどの設定でもスコアが高くなっていることが読み取れる．また，どのモデルもベースラインのモデルと比べて適合率，再現率が共に上がっており，他言語を考慮した MLM を用いた転移学習を行うことで文法誤り訂正の性能が改善可能であることがわかる．

最後に，TLM を用いた実験結果はチェコ語とロシア語の TLM を用いて転移学習を行ったモデルが $F_{0.5}$ と GLEU のスコアのどちらも全てのモデルの中で最も高いことがわかる．一方で，英語から転移学習したモデルと日本語から転移学習をしたモデルは MLM を用いたモデルと比較してほとんど変わらないことがわかる．これはチェコ語がロシア語に近い言語であるため TLM を用いて言語間の対応を学習できたのに対して，日本語や英語はロシア語から遠い言語であるため言語間の対応を学習できなかったのではないかと考えられる．また，MLM を用いたモデルと TLM を用い

原文	… когда речь идет о международных отношениях , но не о внутренний$_{1\{\text{A:Nom}\}}$ политики$_{2\{\text{N:Gen}\}}$ ·
正解文	… , когда речь идет о международных отношениях , но не о внутренней$_{1\{\text{A:Pre}\}}$ политике$_{2\{\text{N:Pre}\}}$ ·
日本語訳	… ， 国際 関係 に関する とき で ， 国内$_1$ 政治$_2$ について で は あり ませ ん ．
Ru のみ	… , когда речь идет о международных отношениях , но не о внутренний$_{1\{\text{A:Nom}\}}$ политики$_{2\{\text{N:Gen}\}}$ ·
TLM (Ja→Ru)	… , когда речь идет о международных отношениях , но не о внутренних$_{1\{\text{A:Pre}\}}$ отношениях$_{2\{\text{N:Pre}\}}$ ·
TLM (En→Ru)	… , когда речь идет о международных отношениях , но не о внутренним$_{1\{\text{A:Inst}\}}$ политики$_{2\{\text{N:Gen}\}}$ ·
TLM (Cs→Ru)	… , когда речь идет о международных отношениях , но не о внутренней$_{1\{\text{A:Pre}\}}$ политике$_{2\{\text{N:Pre}\}}$ ·

表 5: Adj:Case と Noun:Case の誤りタイプに対するモデルの出力例．色がついた単語の添え字のうち，数字は誤りの識別番号を示す．中括弧の中の第 1 項は品詞であり，A は Adj，N は Noun を示す．第 2 項はその単語の格であり，Nom は主格，Gen は属格，Pre は前置格，Inst は具格を示す．

たモデルの実験結果の両方でロシア語に近い言語であるチェコ語から転移したモデルのスコアが高く，遠い言語である日本語から転移したモデルのスコアが低くなっていることがわかる．これらのことから，MLM や TLM を用いて事前学習した訂正モデルの転移学習を行うことで，類似した言語間で文法誤り訂正の性能を改善可能であることがわかる．

5 転移された文法項目に関する分析

言語間の転移学習により類似した文法項目に関する誤りの訂正精度が向上していることを明らかにする．表 4 は評価データに人手で付与された誤りタイプに対する，各モデルの再現率を示している．結果をわかりやすくするためにベースラインのモデルと TLM を用いたモデルの結果を載せており，全誤りタイプのうち分析のためにチェコ語と英語それぞれに類似した誤りタイプの結果を載せている．まず，チェコ語とロシア語で類似した誤りである，Adj:Case（形容詞の格変化に関する誤り）と Noun:Case（名詞の格変化に関する誤り）において，チェコ語から転移したモデルのスコアが最も高いことがわかる．また，英語とロシア語に類似した誤りである Verb:Num/Per（主語に対する動詞の誤り）において，英語から転移したモデルのスコアはロシア語のみのモデルや日本語から転移したモデルよりも高いことがわかる．このことから，類似した文法項目において性能が改善していることがわかる．

Adj:Case と Noun:Case の誤りタイプに関する各モデルの出力例を表 5 に示す．原文の下線 1 が Adj:Case の誤りであり，下線 2 が Noun:Case の誤りである．原文と正解文を比べると，訂正後の文ではそれぞれ格が変わっていることがわかる．各モデルの出力を見ると，Ru のみのモデル，日本語からの転移学習を行ったモデル，英語からの転移学習を行ったモデルは誤りを訂正できていないのに対して，チェコ語から転移学習を行ったモデルのみが正しく訂正できている．これは，訂正後の単語の格である前置格が，ロシア語とチェコ語では単語の語尾の変化で表現されるのに対して，英語や日本語では他の単語（英語なら「about something」，日本語なら「〜について」など）を用いて表現されるためではないかと考えられる．

6 おわりに

本研究では，文法誤り訂正において言語をまたぎ類似した文法誤りの情報を転移可能であることを示した．さらに，言語間で類似する文法項目の誤りの訂正精度が向上していることを明らかにした．今後は，英語のような大規模な学習者データが存在する言語において，学習データと転移の効果の相関などを調査したい．

参考文献

[1] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit3:web inventory of transcribed and translated talks. In *EAMT*, 2012.

[2] Shamil Chollampatt and Hwee Tou Ng. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *AAAI*, 2018.

[3] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *NIPS*, 2019.

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

[5] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *BEA*, 2013.

[6] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 2017.

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

[8] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Ground truth for grammatical error correction metrics. In *ACL-IJCNLP*, 2015.

[9] 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.

[10] Hiromi Oyama, Mamoru Komachi, and Yuji Matsumoto. Towards automatic error type classification of Japanese language learners' writings. In *PACLIC*, 2013.

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

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[14] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, 2016.