Cross-lingual Transfer Learning for Grammatical Error Correction

Ikumi Yamashita Satoru Katsumata* Masahiro Kaneko
Aizhan Imankulova Mamoru Komachi
Tokyo Metropolitan University, Japan
yamashita-ikumi@ed.tmu.ac.jp
satoru.katsumata@retrieva.jp
{kaneko-masahiro, imankulova-aizhan}@ed.tmu.ac.jp
komachi@tmu.ac.jp

Abstract

In this study, we explore cross-lingual transfer learning in grammatical error correction (GEC) tasks. Many languages lack the resources required to train GEC models. Cross-lingual transfer learning from high-resource languages (the *source models*) is effective for training models of low-resource languages (the *target models*) for various tasks. However, in GEC tasks, the possibility of transferring grammatical knowledge (e.g., grammatical functions) across languages is not evident. Therefore, we investigate cross-lingual transfer learning methods for GEC. Our results demonstrate that transfer learning from other languages can improve the accuracy of GEC. We also demonstrate that proximity to source languages has a significant impact on the accuracy of correcting certain types of errors.

1 Introduction

Grammatical error correction (GEC) is the task of correcting grammatically incorrect sentences. The demand for GEC has grown significantly in recent decades because of the increasing opportunities for cross-cultural collaboration. Previous studies in the literature primarily focused on improving automated GEC for the English language. Thus, because of the large amount of data available for training, several machine-learning-based methods have achieved high scores in English GEC (Zhao et al., 2019; Grundkiewicz et al., 2019; Kiyono et al., 2019; Kaneko et al., 2020). In recent years, researchers have started working on other languages, including Russian and Czech (Rozovskaya and Roth, 2019; Náplava and Straka, 2019). However, for these languages, the language resources required to train the GEC models accurately are not sufficiently available.

It is known that using high-resource languages as the *source languages* can improve the accuracy of deep neural models for low-resource *target languages* in various settings (Johnson et al., 2017; Ruder et al., 2019; Dabre et al., 2020). One such setting involves cross-lingual transfer learning (Zoph et al., 2016), which aims to improve the accuracy of low-resource target models using knowledge from high-resource source models. The similarities between these languages is a key factor for successfully transferring grammatical knowledge (Cotterell and Heigold, 2017; Johnson et al., 2017). For example, languages within the same language family share several rules of grammar and nuances of vocabulary, which aid the learning process of the target models.

However, thus far, no study has investigated the use of cross-lingual transfer learning for GEC from other languages; therefore, it is unclear if useful grammatical knowledge (e.g., case inflection or conjugation) can be transferred. Table 1 shows example case inflections of words that mean "sister" in English, Russian, and Czech. In English, the difference between nominative and genitive is marked by the suffix "'s," whereas in Russian and Czech, it is marked by word conjugation. This example shows that Russian and Czech inflections are similar, suggesting that it may be possible to perform cross-lingual transfer learning by exploiting their grammatical similarities.

In this study, we investigate the following three research questions with respect to cross-lingual transfer learning for GEC:

*Currently at Retrieva, Inc.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4704–4715
Barcelona, Spain (Online), December 8-13, 2020

	Nominative	Genitive
English	The house that my sister lives	My sister's house
Russian	Дом, где живет моя сестра	Дом моей сестры
Czech	Dům, kde bydlí moje sestra	Dům mé sestry

Table 1: Case inflections in English, Russian and Czech.

(a) Does cross-lingual transfer learning improve GEC? To help answer this question, we compare the results of GEC models trained with and without transfer learning. In addition, we compare several cross-lingual transfer learning methods.

(b) Is information on grammatical errors transferable? To help answer this question, we analyze the correction results for each error type. Specifically, we transfer grammatical knowledge from languages that have similar grammatical structures and analyze the results for similar error types between the languages, including the noun case inflection.

(c) How does the size of data in the target language affect the results of transfer learning? Generally, in transfer learning, the transfer is performed from a high-resource language to a low-resource language. However, it is not clear whether it is effective to perform transfer learning from a low-resource language to a high-resource language. Therefore, we analyze the effectiveness of transfer learning, even for targets from a high-resource setting.

Our results indicate that using transfer learning from similar languages can improve the accuracy of GEC for certain grammatical errors. In particular, we show that the error correction performance of similar lexical items is improved and that the transfer of grammatical knowledge is possible. Additionally, we demonstrate that transfer learning is more effective for some types of errors than others, depending on the size of the target data.

2 Related Work

Most recent GEC methods use the encoder-decoder (EncDec) model, which requires large-scale training data (Zhao et al., 2019; Grundkiewicz et al., 2019). Therefore, several studies created additional pseudo-data in low-resource scenarios (Náplava and Straka, 2019; Rozovskaya and Roth, 2019). For example, because it is easy to generate a grammatically incorrect sentence from a grammatically correct sentence, extensive research has been conducted on generating pseudo-data from large-scale monolingual corpora (Xie et al., 2018; Kiyono et al., 2019). In addition, the use of EncDec models pretrained with large-scale unlabeled data is known to be effective for GEC (Kaneko et al., 2020). These studies aimed to improve the performance of GEC using large-scale training data.

Furthermore, research has also been conducted on the use of linguistic knowledge from other languages in neural machine translation (NMT). Zoph et al. (2016) proposed a method to fine-tune NMT models trained from high-resource language pairs on low-resource language pairs. Johnson et al. (2017) demonstrated that a language can be translated with no training data by jointly training one model by concatenating the training data from multiple languages. Schuster et al. (2018) presented a method that uses a bidirectional NMT encoder for cross-lingual contextual word representations, to generate dialog responses. For question answering, Lee and Lee (2019) proposed a cross-lingual transfer learning method that uses generative adversarial networks. These studies focused on the tasks for which semantic information is more important, as opposed to GEC, for which grammatical information is the key factor.

Various studies have analyzed the transfer of syntactic knowledge between languages. Kim et al. (2017) proposed a part-of-speech tagging method for learning language-independent and language-dependent expressions between languages by combining two models corresponding to the expressions. Ahmad et al. (2019) utilized adversarial training to train contextual encoders that produce invariant representations across languages, thereby facilitating cross-lingual transfers for dependency parsing. Wu

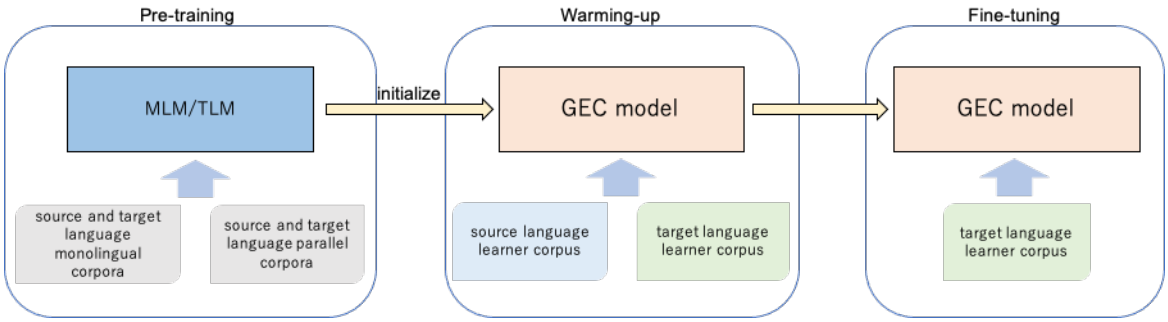

Figure 1: Overall training steps.

and Dredze (2019) used multilingual BERT for five tasks, such as POS tagging and dependency parsing, and demonstrated that its performance can be improved by using multilingual knowledge. As in our study, these studies perform transfer learning between languages in tasks for which syntactic information is important. However, it is not clear whether linguistic knowledge about grammatical errors can be transferred across languages.

Several GEC studies using L1 information have been conducted. Rozovskaya and Roth (2011) adopted information from five L1s with different priorities to preposition correction using the naïve Bayes classifier. Rozovskaya et al. (2017) extended this method to eleven L1s and three error types. Mizumoto et al. (2011) demonstrated that using the same L1 for training and test data in an SMT-based GEC system improved the system performance. Chollampatt et al. (2016) extended this method by incorporating three different L1 neural language models into an SMT-based GEC model as features to adapt to each L1. In these studies, GEC was performed considering the L1 information; however, unlike our study, the objective of these studies did not include the transfer of grammatical knowledge between languages.

3 GEC using Cross-lingual Transfer Learning

3.1 Overall Training Steps

We train a GEC model that employs the Masked Language Modeling (MLM) / Translation Language Modeling (TLM) (Conneau and Lample, 2019) shown in Subsection 3.2, and the transfer learning method shown in Subsection 3.3.

Figure 1 illustrates the overall training steps. First, we pre-train the MLM/TLM with the monolingual/parallel corpora of the source and target languages. Second, we initialize the GEC model using the MLM/TLM and train it with the learner corpora of the source and target languages. Finally, we fine-tune the GEC model with the learner corpus of the target language.

3.2 Using Pre-trained Language Representations

We use pre-trained language representations in our cross-lingual transfer learning to transfer cross-lingual linguistic knowledge that cannot be obtained solely from a learner corpus. We use a method based on MLM and TLM to learn the language representations [1].

MLM training uses monolingual source and target language data. Input is given as one sentence with some tokens being masked, and training is performed by predicting the masked tokens. When training the MLM, a batch includes sentences coming from the same language at each iteration. TLM extends the capability of MLM to use parallel data for training. If the input data are parallel, then the sentence pair is combined into one sequence while masking some tokens. Training and prediction are performed in a manner similar to those in MLM. Because MLM and TLM are used in combination, they are trained alternately. Henceforth, in this paper, the model trained by combining MLM and TLM will be referred to

[1] In this study, we do not use the published pre-trained model; instead, we pre-train the MLM/TLM by ourselves. This is because the purpose of this study is to analyze whether grammatical knowledge can be transferred between two languages; therefore, we want to eliminate the influence of other languages.

target	source similarity		
	high	moderate	low
Russian	Czech	English	Japanese
Czech	Russian	English	Japanese
English	German	Russian	Japanese

Table 2: Languages used in the experiment.

as TLM. It is assumed that using a parallel corpus with TLM provides better knowledge transfer between languages than using only MLM.

We compare the GEC model initialized using MLM with that initialized using TLM, and investigate whether it is better to use a parallel corpus for cross-lingual transfer for GEC. After initializing the GEC model using MLM or TLM, we initialize the learning rate of the model and train the GEC model. While training the GEC models, we do not use language embedding, which was proposed by (Conneau and Lample, 2019).

3.3 Transfer-learning Method for GEC

In this study, we investigate whether grammatical knowledge can be transferred in GEC using cross-lingual transfer learning. Various methods are utilized for cross-lingual transfer learning, as discussed in Section 2. In this study, we focused on sharing both lexical and grammatical knowledge between languages. Thus, we use the EncDec model to facilitate transfer learning.

Several studies on NMT have demonstrated that training a source model on a combination of different languages is effective when performing fine-tuning on low-resource language pairs (Imankulova et al., 2019; Dabre et al., 2019). Therefore, we train the GEC models by concatenating the learner data from both the source and the target languages, wherein each batch may consist of tokens from the two languages; subsequently, we fine-tune the models using the learner data of the target language. Finally, the outputs of all models are re-ranked, as proposed by Chollampatt et al. (2018).

4 Experiments

4.1 Languages

In this study, we perform experiments with GEC on three target languages: Russian, Czech, and English. For each target language, we use three source languages: one with high similarity, one with moderate similarity, and one with low similarity (i.e., Japanese, which is not used as a target language). For the Russian GEC, we use Czech, English, and Japanese as the source languages. Russian and Czech are languages that belong to the Slavic family; hence, they have considerable commonalities, such as in the inflections of adjectives and nouns. Although English is different from Russian in terms of the language family and the characters used, it has some similarities; for example, the verb forms change depending on the singular or plural status of the subject. Japanese is the language that is least related to Russian among the languages used in this study. It differs not only in terms of the language family but also in terms of characters and sentence structure.

For the Czech GEC, we use Russian, English, and Japanese as the source languages. The similarities between Russian and Czech have already been mentioned, and the relation of Czech to English and Japanese is similar to that of Russian.

For the English GEC, we use German, Russian, and Japanese as the source languages. English and German are considerably similar languages and belong to the same Germanic family. The relation between English and Russian has been described above; Japanese is again the least similar language to English among the languages used.

Data type	Corpus	train	dev	test
Parallel	TED Talks	80K	1.3K	-
Monolingual	News Crawl (to train a language model for re-ranking)	33M	-	-
	News Crawl (18)	1.2M	2.5K	-
	Wikipedia	1.2M	2.5K	-
Learner corpora	RULEC-GEC (Ru)	5K	2.5K	5K
	Lang-8-Ru	49K	-	-
	NUCLE (En)	57K	-	-
	CoNLL 2013 (En)	-	1.4K	-
	CoNLL 2014 (En)	-	-	1.3K
	Lang-8-En	1.3M	-	-
	AKCES-GEC (Cs)	40K	2.5K	2.4K
	Falko-MERLIN-GEC (De)	15K	2.5K	-
	Lang-8-De	39K	-	-
	Lang-8-Ja	54K	-	-
	NAIST Goyo Corpus (Ja)	-	3.3K	-

Table 3: GEC data overview.

4.2 Data

The data used in the experiment are presented in Table 3. In this study, we use WMT-2019's News Crawl [2] and Japanese Wikipedia data [3] as monolingual data for training the MLM, and we used TED talks (Cettolo et al., 2012) [4] as parallel data for training the TLM. The development and test data for MLM are extracted from each monolingual dataset, excluding the training data. The development and test data for TLM include data from TED Talks, in addition to the data for MLM.

We use RULEC-GEC (Ru) (Rozovskaya and Roth, 2019) [5], Lang-8 [6], NUCLE (En) (Dahlmeier et al., 2013) [7], AKCES-GEC (Cs) (Náplava and Straka, 2019) [8], and Falko-MERLIN-GEC (De) (Boyd, 2018) [9] as the learner corpora for training the GEC model [10]. For the development and test data for GEC, we use the Russian, Czech, and German data attached to each corpus. We use English data from CoNLL 2013 (Ng et al., 2013) and CoNLL 2014 and Japanese data from the NAIST Goyo Corpus (Oyama et al., 2013) for the development and test data [11]. We also use Russian News Crawl (2015–2018), Czech News Crawl (2014–2018), and English News Crawl (2015–2018) to train the language model for re-ranking the GEC model.

The TED Talks data are reconstructed from the original English translation data by extracting the corresponding sentence pairs in each language. News Crawl, Wikipedia, Lang-8, and NUCLE data are obtained by extracting the number of sentences depicted in Table 3 from the original data. To maintain a consistent experimental setting, the size of each source language data is adjusted to be the same as that of the language that has the smallest data size. The source languages data size is 40K in the Russian GEC and 54K in the Czech and English GEC.

To investigate the effect of the size of the target language data on transfer learning, the experiment on GEC in English is conducted under two data settings, unlike the experiments on GEC in Russian

[2] http://www.statmt.org/wmt19/translation-task.html
[3] https://dumps.wikimedia.org/jawiki/
[4] https://wit3.fbk.eu/
[5] https://github.com/arozovskaya/RULEC-GEC
[6] https://sites.google.com/site/naistlang8corpora/
[7] https://www.comp.nus.edu.sg/~nlp/corpora.html
[8] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3057
[9] https://github.com/adrianeboyd/boyd-wnut2018
[10] When English is used as the source language, we use only NUCLE as the training data.
[11] The English test data exclude sentences that require more than 30 minutes for evaluation using m2scorer.

	Model	P	R	$F_{0.5}$
PLAIN	Ru-only	19.29	14.08	17.96
	Cs→Ru	19.05	12.78	17.35
	En→Ru	23.76	13.65	20.70
	Ja→Ru	20.70	13.57	18.73
MLM	Ru-only	19.95	**23.15**	20.52
	Cs→Ru	26.36	19.02	24.47
	En→Ru	26.02	19.74	24.47
	Ja→Ru	27.23	16.13	23.93
TLM	Cs→Ru	**28.51**	22.47	**27.06**
	En→Ru	27.60	22.18	26.31
	Ja→Ru	26.11	19.99	24.61

Table 4: Russian GEC results.

	Model	P	R	$F_{0.5}$
PLAIN	Cs-only	52.05	39.29	48.88
	Ru→Cs	59.93	38.73	54.01
	En→Cs	61.35	39.01	55.05
	Ja→Cs	56.22	38.53	51.49
MLM	Cs-only	57.46	47.40	55.12
	Ru→Cs	63.58	47.15	59.43
	En→Cs	63.54	48.63	59.87
	Ja→Cs	62.15	47.35	58.50
TLM	Ru→Cs	**65.09**	**50.82**	**61.63**
	En→Cs	63.20	48.47	59.58
	Ja→Cs	63.84	45.70	59.15

Table 5: Czech GEC results.

and Czech. In the first setting, only NUCLE is used as the English training data. Thus, in total, 57K sentences are included in the English training data. We call this setting, "NUCLE only." In the second setting, NUCLE and Lang-8-En are used as the training data for English. Thus, in total, 1.3M + 57K sentences are included in the English training data. We call this setting, "NUCLE + Lang-8-En."

To tokenize the Japanese sentences, we use `MeCab` [12] with the UniDic (v.2.1.1) dictionary. Other languages are tokenized using NLTK [13]. For the target language for GEC, we use pyspellchecker [14] to preprocess all data. Then we convert them into subwords via byte pair encoding (Sennrich et al., 2016) using fastBPE [15].

4.3 Settings

We use the same architecture as Conneau and Lample (2019) for the MLM/TLM and transformer encoder and decoder for GEC models. Both the encoder and the decoder of the GEC model are initialized with the parameters of MLM/TLM. The number of layers in the model is six, the dimension of the hidden and embedding layers is 1,024, the batch size is 32, and a dropout with a probability of 0.1 is applied. The best model is selected using perplexity on the development data. We report the precision, recall, and $F_{0.5}$ scores using m2scorer (Dahlmeier and Ng, 2012) for the test data.

4.4 Baseline

In this study, we use two baselines to compare the effects of transfer learning and the MLM/TLM.

PLAIN In this setting, we do not use the MLM/TLM. Therefore, the GEC model learns grammatical knowledge from the learner corpus only.

MLM {Ru,Cs,En}-only In this setting, we pre-train the MLM with the target language monolingual corpus only and train the GEC model with the target language learner corpus only. This model learns grammatical knowledge from the large-scale target language monolingual corpus and learner corpus; it does not use knowledge of other languages.

4.5 Results

Table 4 shows the results for the Russian GEC. The models that use transfer learning with MLM or TLM obtain a higher $F_{0.5}$ score than the PLAIN models and the MLM Ru-only model; moreover, the model that uses transfer learning from Czech with TLM obtains the highest precision and $F_{0.5}$ score.

[12] `http://taku910.github.io/mecab`, v.0.996.
[13] `https://www.nltk.org/`
[14] `https://github.com/barrust/pyspellchecker`
[15] `https://github.com/glample/fastBPE`

	Model	NUCLE only P	NUCLE only R	NUCLE only $F_{0.5}$	NUCLE + Lang-8-En P	NUCLE + Lang-8-En R	NUCLE + Lang-8-En $F_{0.5}$
PLAIN	En-only	36.67	17.65	30.17	48.11	27.08	41.65
	De→En	35.78	16.88	29.23	47.39	26.18	40.78
	Ru→En	36.32	18.29	30.33	48.07	27.78	41.94
	Ja→En	33.96	17.70	28.69	50.90	29.38	44.39
MLM	En-only	42.66	24.70	37.24	49.74	**35.17**	45.94
	De→En	44.08	24.96	38.23	50.33	33.94	45.90
	Ru→En	41.54	24.91	36.65	50.88	33.03	45.92
	Ja→En	43.44	20.23	35.34	45.94	32.04	42.28
TLM	De→En	46.26	**28.71**	**41.22**	50.15	34.26	45.89
	Ru→En	**46.65**	24.79	39.65	51.84	34.00	46.92
	Ja→En	46.06	21.27	37.35	**52.21**	33.48	**46.96**

Table 6: English GEC results. The data size of "NUCLE only" is 57K and the data size of "NUCLE + Lang-8-En" is 1.3M + 57K.

error type	PLAIN Ru-only	MLM Ru-only	TLM Cs→Ru	TLM En→Ru	TLM Ja→Ru
Spelling	40.64	46.79	**48.96**	48.11	45.65
Insert	20.29	26.43	21.64	**26.95**	23.93
Lexical choice	1.54	4.63	**4.99**	3.68	2.85
Noun:Case	5.20	23.84	**27.02**	22.69	18.93
Delete	14.83	**37.08**	29.41	20.20	20.20
Word form	5.20	11.60	12.00	**13.20**	8.40
Preposition	13.64	20.45	25.00	**27.27**	19.89
Verb:Number/Person	5.16	**34.84**	30.32	27.10	24.52
Adj:Case	5.30	18.18	**20.45**	16.67	14.39
Noun number	7.38	**21.31**	19.67	15.57	13.11

Table 7: Recall of the Russian GEC models on the top-10 most frequent error types in RULEC-GEC.

Table 5 shows the results for the Czech GEC. In this table, a trend similar to the Russian results can be observed. The models that use transfer learning with MLM or TLM obtain a higher $F_{0.5}$ score than the PLAIN models and the MLM Cs-only model; furthermore, the model that uses transfer learning from Russian with TLM obtains the highest precision and $F_{0.5}$ score.

Table 6 shows the results for the English GEC. In the "NUCLE only" setting, the results in Table 6 show a similar trend to that in the Russian and Czech GEC results thus far. The $F_{0.5}$ scores of models that use transfer learning with MLM or TLM are higher than the $F_{0.5}$ scores of the PLAIN models and the MLM En-only model; the model that use transfer learning from German with TLM obtains the highest $F_{0.5}$ scores. In the "NUCLE + Lang-8-En" setting, the $F_{0.5}$ scores of the models that use transfer learning with MLM or TLM are higher than those of the PLAIN models; however, unlike the other GEC results, they are almost the same as the score of the MLM En-only model. No difference is observed between the scores of the models that use transfer learning with MLM or TLM.

original	Безусловно , в России существует более политический транспарант в новостях когда речь идет о международных отношениях , но не о внутренний{A:Nom} политики{N:Gen} .
gold	Безусловно , в России существует более политический транспарант в новостях , когда речь идет о международных отношениях , но не о внутренней{A:Pre} политике{N:Pre} .
English	Of course, in Russia there is a more political transparency in the news when it comes to international relations, but not about domestic politics.
Ru-only	Безусловно , в России существует более политический транспарант в новостях , когда речь идет о международных отношениях , но не о внутренний{A:Nom} политики{N:Gen} .
TLM (Ja→Ru)	Безусловно , в России существует более политический транспарант в новостях , когда речь идет о международных отношениях , но не о внутренних{A:Pre} отношениях{N:Pre} .
TLM (En→Ru)	Безусловно , в России существует более политический транспарант в новостях , когда речь идет о международных отношениях , но не о внутренним{A:Inst} политики{N:Gen} .
TLM (Cs→Ru)	Безусловно , в России существует более политический транспарант в новостях , когда речь идет о международных отношениях , но не о внутренней{A:Pre} политике{N:Pre} .

Table 8: Output example of the Russian GEC model for Adj:Case and Noun:Case. The words in red are incorrect, and those in blue are correct. Brackets show the types of error. The first term indicates a part of speech, where "A" is an adjective, and "N" is a noun. The second term indicates the word case, where "Nom" is nominative, "Gen" is genitive, "Pre" is prepositional, and "Inst" is instrumental.

5 Analysis

5.1 Cross-lingual Transfer Learning for GEC

The GEC results for the three languages provide some interesting insights that are common to all languages for GEC using cross-lingual transfer learning. In all languages, the models that use transfer learning with MLM or TLM score higher than those that use transfer learning without MLM or TLM. This shows that transfer learning with MLM and TLM is effective for GEC, irrespective of the language pairs. The models that use transfer learning with MLM or TLM in the most similar languages obtain a higher score than those that use MLM pretrained in the target language instead of transfer learning. This suggests that a better GEC model can be trained by considering knowledge in multiple languages instead of only the target language. In any language, the model that uses transfer learning with TLM from the language closest to the target language obtains the highest score. This indicates that it is important to perform transfers from a closely related language.

5.2 Similar Lexical Items between Languages

In this subsection, we present our investigation of whether transfer learning between languages improves the error correction accuracy for similar grammatical items[16]. Table 7 shows the recall of different models for the error types calculated using manually annotated evaluation data for Russian GEC. Using Czech as a source language improves the accuracy of most error types compared to the baseline models. Moreover, with similar grammatical items between Czech and Russian, (e.g., Spelling, Lexical choice, Adj:Case (errors of adjective case inflection), and Noun:Case (errors of noun case inflection)), the TLM model transferred from Czech obtains the highest scores, thereby demonstrating the positive effect of transfer learning in similar grammatical items.

Table 8 shows the output examples of the Russian GEC model for the error types of Adj:Case and Noun:Case. The underlined words represent the errors in Adj:Case, and Noun:Case. When the original sentence is compared with the gold sentence, the case in the gold sentence is observed to have changed. The case of a word is denoted by suffixes in Russian and Czech. There are seven cases in Czech, including six of the same cases in Russian. Ru-only fails to correct errors in this case. We speculate that this is caused by a lack of training data. The use of Japanese as a source language leads to erroneous changes in nouns. However, in English, the prepositional relations are marked using standalone words, which is not helpful in correcting the prepositional errors in Russian, wherein the prepositional relations

[16]The error types assigned to Czech data are unique and considerably different from those assigned to Russian and English data. Thus, we do not analyze the Czech GEC model by error type.

NUCLE only					
	PLAIN	MLM	TLM		
error type	En-only	En-only	De→En	Ru→En	Ja→En
Determiner	15.91	23.58	**25.85**	22.44	22.44
Preposition	13.22	18.06	**22.91**	16.74	17.18
Punctuation	1.03	5.15	**9.28**	4.12	8.25
Verb	5.23	**13.95**	11.63	11.05	9.88
Verb Tense	9.47	**15.26**	14.74	8.95	12.11
Spelling	59.72	62.50	61.11	**68.06**	13.89
Pronoun	5.41	**14.86**	**14.86**	9.46	5.41
Verb Form	7.55	**29.25**	**29.25**	27.36	**29.25**
Morphology	9.59	15.07	**20.55**	13.70	13.70
SVA	17.71	32.29	**33.33**	25.00	28.12

NUCLE + Lang-8-En					
	PLAIN	MLM	TLM		
error type	En-only	En-only	De→En	Ru→En	Ja→En
Determiner	27.56	**38.92**	35.23	38.07	34.66
Preposition	21.15	**36.12**	32.60	31.72	32.60
Punctuation	4.12	7.22	6.19	**9.28**	2.06
Verb	12.21	**18.60**	14.53	14.53	15.12
Verb Tense	10.00	**18.95**	14.21	17.89	13.68
Spelling	70.83	69.44	70.83	**72.22**	**72.22**
Pronoun	6.76	17.57	**18.92**	17.57	13.51
Verb Form	24.53	**33.96**	28.30	33.02	32.08
Morphology	17.81	21.92	**28.77**	20.55	23.29
SVA	25.00	43.75	44.79	**45.83**	40.62

Table 9: Recall of the English GEC models on the top-five and bottom-five in terms of the numbers of errors in the Lang-8-En corpus, excluding error types whose number of errors in the test data is less than or equal to 50.

are marked by case systems. Only the TLM model transferred from Czech generates the correct output. We hypothesize that the reason that Adj:case and Noun:case errors are corrected into prepositional cases is because our model captures the grammatical information of Czech, which is also useful in Russian.

5.3 Size of Target Language Data

In this subsection, we analyze the effect of the size of the target language data on transfer learning. We use ERRANT [17] (Bryant et al., 2017) to annotate the training and evaluation data with error types and analyze the knowledge that is effective (i.e., transferable) in transfer learning to a high-resource target. Table 9 shows the recall results of the English GEC models on the top-five error types (determiner, preposition, punctuation, verb, and verb tense) and the bottom-five error types (spelling, pronoun, verb form, morphology, and subject-verb agreement) in terms of the numbers of errors in the Lang-8-En corpus.

The results of "NUCLE only," wherein the target is at a low-resource setting, demonstrate that the model that uses transfer learning from German obtains the highest recall for most error types. This

[17] https://github.com/chrisjbryant/errant

indicates that transfer learning is effective for most error types for any low-resource languages.

For the "NUCLE + Lang-8-En" setting, there is a tendency wherein the recall of the top-five error types does not increase much through transfer learning from any language. However, the tendency of recall of the bottom-five error types is slightly different. The recall of the model that uses transfer learning from Japanese barely increases from that of the baseline model in terms of pronoun, verb form, and subject-verb agreement, which are dissimilar error types between Japanese and English. In contrast, the recall results of the model transferred from German are higher than those of the MLM En-only model for similar errors, such as pronoun, morphology, and subject-verb agreement in English and German. Accordingly, it can be considered that transfer learning from other languages is effective for the error types that are infrequent but similar between the source and target languages, even if the target is a high-resource language.

6 Conclusion

In this study, we show that certain grammatical knowledge can be transferred across languages for GEC. In particular, the correction performance of a model transferred from a similar language is greatly improved. Additionally, we show that the performance improves more when correcting errors for similar grammatical items are corrected than when those for dissimilar grammatical items are corrected. We also show that cross-lingual transfer learning is effective for GEC in both low-resource and high-resource languages to some extent.

In the future, we plan to compare and visualize word embedding in models with and without transfer learning and investigate why MLM and TLM are effective for transfer learning in GEC. We also plan to investigate the results of training and transferring multiple source languages instead of only one.

Acknowledgements

We gratefully thank Yangyang Xi and Lang-8 contributors for sharing their data. This work has been partly supported by the programs of the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS KAKENHI) Grant Numbers 19K12099 and 19KK0286.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *CoNLL*, pages 372–382.

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *W-NUT*, pages 79–84.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *ACL*, pages 793–805.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *EAMT*, pages 261–268.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *AAAI*, pages 5755–5762.

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *EMNLP*, pages 1901–1911.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NeurIPS*, pages 7059–7069.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *EMNLP*, pages 748–759.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *EMNLP-IJCNLP*, pages 1410–1416.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *ArXiv*, abs/2001.01115.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *NAACL-HLT*, pages 568–572.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *BEA*, pages 22–31.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *BEA*, pages 252–263.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *MT Summit*, pages 128–139.

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. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, pages 339–351.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *ACL*, pages 4248–4254.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *EMNLP*, pages 2832–2838.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP-IJCNLP*, pages 1236–1242.

Chia-Hsuan Lee and Hung-yi Lee. 2019. Cross-lingual transfer learning for question answering. *ArXiv*, abs/1907.06042.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *IJCNLP*, pages 147–155.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *W-NUT*, pages 346–356.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *CoNLL*, pages 1–12.

Hiromi Oyama, Mamoru Komachi, and Yuji Matsumoto. 2013. Towards automatic error type classification of Japanese language learners' writings. In *PACLIC*, pages 163–172.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *ACL*, pages 924–933.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *TACL*, pages 1–17.

Alla Rozovskaya, Dan Roth, and Mark Sammons. 2017. Adapting to learner errors with minimal supervision. *CL*, pages 723–760.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *JAIR*, page 569630.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL-HLT*, pages 3795–3805.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *EMNLP-IJCNLP*, pages 833–844.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *NAACL-HLT*, pages 619–628.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *NAACL*, pages 156–165.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *EMNLP*, pages 1568–1575.