Personalizing Grammatical Error Correction: Adaptation to Proficiency Level and L1

Maria Nădejde
Grammarly
maria.nadejde@grammarly.com

Joel Tetreault*
Dataminr
jtetreault@dataminr.com

Abstract

Grammar error correction (GEC) systems have become ubiquitous in a variety of software applications, and have started to approach human-level performance for some datasets. However, very little is known about how to efficiently personalize these systems to the user's characteristics, such as their proficiency level and first language, or to emerging domains of text. We present the first results on adapting a general purpose neural GEC system to both the proficiency level and the first language of a writer, using only a few thousand annotated sentences. Our study is the broadest of its kind, covering five proficiency levels and twelve different languages, and comparing three different adaptation scenarios: adapting to the proficiency level only, to the first language only, or to both aspects simultaneously. We show that tailoring to both scenarios achieves the largest performance improvement (3.6 $F_{0.5}$) relative to a strong baseline.

1 Introduction

Guides for English teachers have extensively documented how grammatical errors made by learners are influenced by their native language (L1). Swan and Smith (2001) attribute some of the errors to "transfer" or "interference" between languages. For example, German native speakers are more likely to incorrectly use a definite article with general purpose nouns or omit the indefinite article when defining people's professions. Other errors are attributed to the absence of a certain linguistic feature in the native language. For example, Chinese and Russian speakers make more errors involving articles, since these languages do not have articles.

A few grammatical error correction (GEC) systems have incorporated knowledge about L1. Rozovskaya and Roth (2011) use a different prior for each of five L1s to adapt a Naive Bayes classifier for preposition correction. Rozovskaya et al. (2017) expand on this work to eleven L1s and three error types. Mizumoto et al. (2011) showed for the first time that a statistical machine translation (SMT) system applied to GEC performs better when the training and test data have the same L1. Chollampatt et al. (2016) extend this work by adapting a neural language model to three different L1s and use it as a feature in SMT-based GEC system. However, we are not aware of prior work addressing the impact of both proficiency level and native language on the performance of GEC systems. Furthermore, neural GEC systems, which have become state-of-the-art (Gehring et al., 2017; Junczys-Dowmunt et al., 2018; Grundkiewicz and Junczys-Dowmunt, 2018), are general purpose and domain agnostic.

We believe the future of GEC lies in providing users with feedback that is personalized to their proficiency level and native language (L1). In this work, we present the first results on adapting a general purpose neural GEC system for English to both of these characteristics by using fine-tuning, a transfer learning method for neural networks, which has been extensively explored for domain adaptation of machine translation systems (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Chu et al., 2017; Miceli Barone et al., 2017; Thompson et al., 2018). We show that a model adapted to both L1 and proficiency level outperforms models adapted to only one of these characteristics. Our contributions also include the first results on adapting GEC systems to proficiency levels and the broadest study of adapting GEC to L1 which includes twelve different languages.

[*] This research was conducted while the author was at Grammarly.

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 27–33
Hong Kong, Nov 4, 2019. ©2019 Association for Computational Linguistics

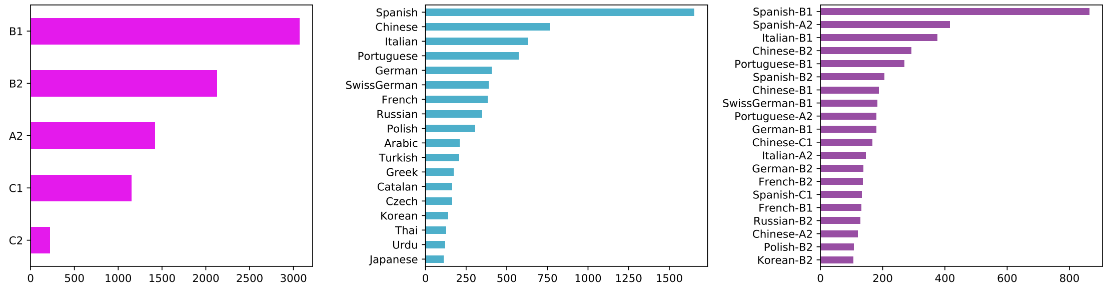

Figure 1: Corpus Distributions for CEFR Level, L1 and L1-Level.

2 Personalizing GEC

Data In this work, we adapt a general purpose neural GEC system, initially trained on two million sentences written by both native and non-native speakers and covering a variety of topics and styles. All the sentences have been corrected for grammatical errors by professional editors.[1]

Adaptation of the model to proficiency level and L1 requires a corpus annotated with these features. We use the Cambridge Learner Corpus (CLC) (Nicholls, 2003) comprising examination essays written by English learners with six proficiency levels[2] and more than 100 different native languages. Each essay is corrected by one annotator, who also identifies the minimal error spans and labels them using about 80 error types. From this annotated corpus we extract a parallel corpus comprising of source sentences with grammatical errors and the corresponding corrected target sentences.

We do note the proprietary nature of the CLC which makes reproducibility difficult, though it has been used in prior research, such as Rei and Yannakoudakis (2016). It was necessary for this study as the other GEC corpora available are not annotated for both L1 and level. The Lang-8 Learner Corpora (Mizumoto et al., 2011) also provides information about L1, but it has no information about proficiency levels. The FCE dataset (Yannakoudakis et al., 2011) is a subset of the CLC, however, it only covers one proficiency level and there are not enough sentences for each L1 for our experiments. Previous work on adapting GEC classifiers to L1 (Rozovskaya et al., 2017) used the FCE corpus, and thus did not address adaptation to different proficiency levels. One of our future goals is to create a public corpus for this type of work.

Experimental Setup Our baseline neural GEC system is an RNN-based encoder-decoder neural network with attention and LSTM units (Bahdanau et al., 2015). The system takes as input an English sentence which may contain grammatical errors and decodes the corrected sentence. We train the system on the parallel corpus extracted from the CLC with the OpenNMT-py toolkit (Klein et al., 2018) using the hyper-parameters listed in the Appendix. To increase the coverage of the neural network's vocabulary, without hurting efficiency, we break source and target words into sub-word units. The segmentation into sub-word units is learned from unlabeled data using the Byte Pair Encoding (BPE) algorithm (Sennrich et al., 2016). The vocabulary, consisting of 20,000 BPE sub-units, is shared between the encoder and decoder.[3] We truncate sentences longer than 60 BPE sub-units and train the baseline system with early stopping on a development set sampled from the base dataset.[4]

To train and evaluate the adapted models, we extract subsets of sentences from the CLC that have been written by learners having a particular Level, L1, or L1-Level combination. We consider all subsets having at least 11,000 sentences, such that we can allocate 8,000 sentences for training, 1,000 for tuning and 2,000 for testing. We compare adapted models trained and evaluated on the same subset of the data. For example, we adapt a model using the Chinese training data and then evaluate it on the Chinese test set.

Since our base dataset and CLC are different domains, we wanted to make sure that improve-

[1] To maintain anonymity, we do not include more details.

[2] The CLC uses levels defined by the Common European Framework of Reference for Languages: A1 - Beginner, A2 - Elementary, B1 - Intermediate, B2 - Upper intermediate, C1 - Advanced, C2 - Proficiency.

[3] Although the source and target vocabularies are the same, the embeddings are not tied.

[4] Performance did not improve after 15 epochs.

ments by fine-tuning by Level or L1 were not due to simply being in-domain with the test data, which is also from the CLC. To control for this, we construct another baseline system ("Random") by adapting the general purpose GEC system to a random sample of learner data drawn from the CLC. In Figure 1 we show the distribution of Level, L1 and L1-Level sentences in a random CLC sample, for the subsets having at least 100 sentences. B1 is the most frequent level, while A2, the lowest proficiency level included in this study, is half as frequent in the random sample. The L1 distribution is dominated by Spanish, with Chinese second with half as many sentences. Among the L1-Level subsets, Spanish-B2 is the most frequent with Spanish-A2 covering half as many sentences.

Fine-tuning We build adapted GEC models using fine-tuning, a transfer learning method for neural networks. We continue training the parameters of the general purpose model on the "in-domain" subset of the data covering a particular Level, L1, or L1-Level. Thompson et al. (2018) showed that adapting only a single component of the encoder-decoder network is almost as effective as adapting the entire set of parameters. In this work, we fine-tune the parameters of the source embeddings and encoder, while keeping the other parameters fixed.

To avoid quickly over-fitting to the smaller "in-domain" training data, we reduce the batch size (Thompson et al., 2018) and continue using the dropout regularization (Miceli Barone et al., 2017). We apply dropout to all the layers and to the source words, as well as variational dropout (Gal and Ghahramani, 2016) on each step, all with probability 0.1. We also reduce the learning rate by four times and use the `start_decay_at` option which halves the learning rate after each epoch. Consequently, the updates become small after a few epochs. To enable the comparison between different adaptation scenarios, all fine-tuned models are trained for 10 epochs on 8,000 sentences of "in-domain" data.

3 Results

We report the results for the three adaptation scenarios: adapting to Level only, adapting to L1 only, and adapting to both L1 and Level. We summarize the results by showing the average M^2 F$_{0.5}$ score (Dahlmeier and Ng, 2012) across all the test sets included in the respective scenario.

We first note that the strong baseline ("Random"), which is a model adapted to a random sample of CLC , achieves improvements between 11 to 13 F$_{0.5}$ points on average on all scenarios. While not the focus of the paper, this large improvement shows the performance gains by simply adapting to a new domain (in this case CLC data). Second, we note that the models adapted only by Level or by L1 are on average better than the "Random" model by 2.1 and 2.3 F$_{0.5}$ points respectively. Finally, the models adapted to both Level and L1 outperform all others, beating the "Random" baseline on average by 3.6 F$_{0.5}$ points.

On all adaptation scenarios we report the performance of the single best model released by Junczys-Dowmunt et al. (2018). Their model, which we call *JD single*, was trained on English learner data of comparable size to our base dataset and optimized using the CoNLL14 training and test data.

Adaptation by Proficiency Level We adapt GEC models to five of the CEFR proficiency levels: A2, B1, B2, C1, C2. The results in Table 1 show that performance improves for all levels compared to the "Random" baseline. The largest improvement, 5.2 F$_{0.5}$ points, is achieved for A2, the lowest proficiency level. We attribute the large improvement to this level having a higher error rate, a lower lexical diversity and being less represented in the random sample on which the baseline is trained on. In contrast, for the B1 and B2 levels, the most frequent in the random sample, improvements are more modest: 0.7 and 0.2 F$_{0.5}$ points respectively. Our adapted models are better than the *JD single* model on all levels, and with a large margin on the A2 and C1 levels.

Adapt	A2	B1	B2	C1	C2	Avg.
No	30.4	34.9	33.1	32.5	33.0	32.8
Rand.	48.4	47.9	42.5	41.4	39.2	43.8
Level	**53.6**	**48.6**	**42.7**	**43.3**	**41.1**	**45.9**
JD single	44.1	47.1	41.7	37.8	35.0	44.1

Table 1: Adaptation to Proficiency Level in F$_{0.5}$

Adaptation by L1 We adapt GEC models to twelve L1s: Arabic, Chinese, French, German, Greek, Italian, Polish, Portuguese, Russian, Spanish, Swiss-German and Turkish. The results in Table 2 (top) show that all L1-adapted models are better than the baseline, with improvements ranging from 1.2 F$_{0.5}$ for Chinese and French, up

Adapt	AR	CN	FR	DE	GR	IT	PL	PT	RU	ES	CH	TR	Avg
No	37.5	36.2	32.7	31.4	32.7	29.3	36.0	31.7	35.8	32.1	31.1	35.4	33.5
Random	46.3	45.0	44.9	44.7	46.4	44.9	46.2	45.2	45.3	47.6	44.2	47.0	45.6
L1	**48.3**	**46.2**	**46.1**	**47.1**	**49.0**	**46.8**	**48.4**	**47.6**	**47.8**	**49.8**	**47.1**	**50.6**	**47.9**
JD single	47.0	44.7	44.2	41.4	44.1	40.7	46.0	44.6	43.7	44.8	40.7	47.5	44.1

Adapt	CN-B2	CN-C1	FR-B1	DE-B1	IT-B1	PT-B1	ES-A2	ES-B1	ES-B2	Avg.
No	36.1	32.5	31.8	31.2	28.1	31.4	28.9	31.9	33.7	31.8
Random	42.7	39.1	45.3	46.1	43.5	45.2	50.2	46.4	44.1	44.7
Level	43.4	41.0	46.5	46.9	45.3	46.1	56.6	47.5	43.7	46.3
L1	44.1	40.9	46.5	48.1	46.5	46.2	53.8	47.6	44.4	46.5
L1 & Level	**45.5**	**43.1**	**48.1**	**50.2**	**47.3**	**47.9**	**58.2**	**48.8**	**45.6**	**48.3**
JD single	43.0	35.8	46.9	43.8	41.6	46.7	43.4	45.0	41.0	43.0

Table 2: Top: Adaptation to L1 Only. Bottom: Adaptation to Level and L1. Eval metric: $F_{0.5}$

to 3.6 $F_{0.5}$ for Turkish. For the languages that are less frequent in the random sample of CLC (Greek, Turkish, Arabic, Polish and Russian) we see consistent improvements of over 2 $F_{0.5}$ points. Our adapted models are better than the *JD single* model on all L1s, and with a margin larger than 5 $F_{0.5}$ points on German, Swiss-German, Italian, Greek and Spanish.

Adaptation by L1 and Proficiency Level Finally, we adapt GEC models to the following nine L1 – Level subsets: Chinese-B2, Chinese-C1, French-B1, German-B1, Italian-B1, Portuguese-B1, Spanish-A2, Spanish-B1 and Spanish-B2. We include these subsets in our study because they meet the requirement of having at least 8,000 sentences for training. All the models adapted to both Level and L1 outperform the models adapted to only one of these features, as shown in Table 2 (bottom). Focusing on the two levels for Chinese native speakers, we see the model adapted to C1 achieves a larger improvement over the baseline, 4.1 $F_{0.5}$ points, compared to 2.7 $F_{0.5}$ points for the B2 level. Again, this is explained by the lower frequency of the C1 level in the random sample of CLC, which is also reflected by the lowest $F_{0.5}$ score for the baseline model. Similarly, among the models adapted to different levels of Spanish native speakers, the one adapted to Spanish-A2 achieves the largest gains of 8 $F_{0.5}$ points. The Spanish-A2 testset has the highest number of errors per 100 words among all the L1-Level testsets, as shown in Table 1 in the Appendix. Furthermore, the A2 level is only half as frequent as the B1 level in the random sample of CLC. Finally, our adapted models are better than the *JD single* model on all L1–Level subsets, with a margin of 5 $F_{0.5}$ points on average.

Adapted	P	R	F0.5
Random	61.9	35.6	54.0
CN-C1	61.1	37.0	54.1
CN-B2	62.4	37.5	55.1
+ spellcheck	63.6	40.3	57.0
JD single	59.1	40.4	54.1
JD ensemble	63.1	42.6	57.5

Table 3: Results on the CoNLL14 testsets for Chinese models.

CoNLL14 Evaluation We compare our adapted models on the CoNLL14 testset (Ng et al., 2014) in Table 3. The model adapted to Chinese-B2 improves the most over the baseline, achieving 55.1 $F_{0.5}$. This result aligns with how the test set was constructed: it consists of essays written by university students, mostly Chinese native speakers. When we pre-process the evaluation set before decoding with a commercial spellchecker[5], our adapted model scores 57.0 which places it near other leading models, trained on a similar amount of data, such as Chollampatt and Ng (2018) (56.52) and Junczys-Dowmunt et al. (2018)[6] (57.53) even though we do not use the CoNLL14 in-domain training data. We note that the most recent state-of-the-art models (Zhao et al., 2019; Grundkiewicz et al., 2019), are trained on up to one hundred million additional synthetic parallel sentences, while we adapt models with only eight thousand parallel sentences.

[5] Details removed for anonymity.

[6] We call their ensemble of four models with language model re-scoring *JD ensemble* and their single best model without language model re-scoring *JD single*

Adapt	Det	Prep	Verb	Tense	NNum	Noun	Pron
CN-C1	3.53	5.90	2.99	1.77	8.28	8.02	22.78
FR-B1	2.34	1.99	12.54	5.16	9.16	3.48	1.13
DE-B1	8.85	1.77	2.04	2.37	3.86	7.18	22.75
IT-B1	2.37	5.32	12.48	6.74	4.40	3.29	8.99
ES-A2	6.06	12.52	7.51	8.54	8.73	12.39	10.57

Table 4: L1-Level breakdown by error type in relative improvements in $F_{0.5}$ over the "Random" baseline.

Error-type Analysis We conclude our study by reporting improvements on the most frequent error types, excluding punctuation, spelling and orthography errors. We identify the error types in each evaluation set with Errant, a rule-based classifier (Bryant et al., 2017). Table 4 shows the results for the systems adapted to both L1 and Level that improved the most in overall $F_{0.5}$. The adapted systems consistently outperform the "Random" baseline on most error types. For Chinese-C1, the adapted model achieves the largest gains on pronoun (Pron) and noun number agreement errors (NNum). The Spanish-A2 adapted model achieves notable gains on preposition (Prep), noun and pronoun errors. Both the French-B1 and Italian-B1 adapted models gain the most on verb errors. For German-B1, the adapted model improves the most on pronoun (Pron) and determiner (Det) errors. The large improvement of 22.75 $F_{0.5}$ points for the pronoun category is in part an artefact of the small error counts. The adapted model corrects 35 pronouns (P=67.3) while the baseline corrects only 15 pronouns (P=46.9). We leave an in depth analysis by error type to future work.

Below, we give an example of a confused auxiliary verb that the French-B1 adapted model corrects. The verb phrase corresponding to "go shopping" in French is "faire des achats", where the verb "faire" would translate to "make/do".

Orig	He told me that celebrity can be bad because he can't *do* shopping normally.
Rand	He told me that *the* celebrity can be bad because he can't *do* shopping normally.
FR-B1	He told me that celebrity can be bad because he can't *go* shopping normally.
Ref	He told me that celebrity can be bad because he can't *go* shopping normally.

4 Conclusions

We present the first results on adapting a neural GEC system to proficiency level and L1 of language learners. This is the broadest study of its kind, covering five proficiency levels and twelve different languages. While models adapted to either proficiency level or L1 are on average better than the baseline by over 2 $F_{0.5}$ points and the largest improvement (3.6 $F_{0.5}$) is achieved when adapting to both characteristics simultaneously.

We envision building a single model that combines knowledge across L1s and proficiency levels using a mixture-of-experts approach. Adapted models could also be improved by using the *mixed fine tuning* approach which uses a mix of in-domain and out-of-domain data (Chu et al., 2017).

Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback. We are also grateful to our colleagues for their assistance and insights: Dimitrios Alikaniotis, Claudia Leacock, Dmitry Lider, Courtney Napoles, Jimmy Nguyen, Vipul Raheja and Igor Tytyk.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

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 *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing, pages 1901–1911. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 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*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 1027–1035, USA. Curran Associates Inc.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 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)*, pages 284–290. Association for Computational Linguistics.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. Opennmt: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 177–184. Association for Machine Translation in the Americas.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494. Association for Computational Linguistics.

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 *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155. Asian Federation of Natural Language Processing.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1181–1191, Berlin, Germany. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for esl correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 924–933, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alla Rozovskaya, Dan Roth, and Mark Sammons. 2017. Adapting to learner errors with minimal supervision. *Computational Linguistics*, 43(4):723–760.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, Berlin, Germany. Association for Computational Linguistics.

Michael Swan and Bernard Smith. 2001. *Learner English: A Teacher's Guide to Interference and other Problems. Second Edition.* Cambridge University Press.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 124–132, Belgium, Brussels. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Wei Zhao, Liang Wang, Kewei Shen, Jia Ruoyu, and Jingming Liu. 2019. Better evaluation for grammatical error correctionimproving 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*, Minneapolis, MN, USA. Association for Computational Linguistics.