Minimally-Augmented Grammatical Error Correction

Roman Grundkiewicz[†‡] and **Marcin Junczys-Dowmunt**[‡]
[†] University of Edinburgh, Edinburgh EH8 9AB, UK
[‡] Microsoft, Redmond, WA 98052, USA
rgrundki@inf.ed.ac.uk, marcinjd@microsoft.com

Abstract

There has been an increased interest in low-resource approaches to automatic grammatical error correction. We introduce Minimally-Augmented Grammatical Error Correction (MAGEC) that does not require any error-labelled data. Our unsupervised approach is based on a simple but effective synthetic error generation method based on confusion sets from inverted spell-checkers. In low-resource settings, we outperform the current state-of-the-art results for German and Russian GEC tasks by a large margin without using any real error-annotated training data. When combined with labelled data, our method can serve as an efficient pre-training technique.

1 Introduction

Most neural approaches to automatic grammatical error correction (GEC) require error-labelled training data to achieve their best performance. Unfortunately, such resources are not easily available, particularly for languages other than English. This has lead to an increased interest in unsupervised and low-resource GEC (Rozovskaya et al., 2017; Bryant and Briscoe, 2018; Boyd, 2018; Rozovskaya and Roth, 2019), which recently culminated in the low-resource track of the Building Educational Application (BEA) shared task (Bryant et al., 2019).[1]

We present Minimally-Augmented Grammatical Error Correction (MAGEC), a simple but effective approach to unsupervised and low-resource GEC which does not require any authentic error-labelled training data. A neural sequence-to-sequence model is trained on clean and synthetically noised sentences alone. The noise is automatically created from confusion sets. Additionally, if labelled data is available for fine-tuning (Hinton and Salakhutdinov, 2006), MAGEC can also serve as an efficient pre-training technique.

The proposed unsupervised synthetic error generation method does not require a seed corpus with example errors as most other methods based on statistical error injection (Felice and Yuan, 2014) or back-translation models (Rei et al., 2017; Kasewa et al., 2018; Htut and Tetreault, 2019). It also outperforms noising techniques that rely on random word replacements (Xie et al., 2018; Zhao et al., 2019). Contrary to Ge et al. (2018) or Lichtarge et al. (2018), our approach can be easily used for effective pre-training of full encoder-decoder models as it is model-independent and only requires clean monolingual data and potentially an available spell-checker dictionary.[2] In comparison to pre-training with BERT (Devlin et al., 2019), synthetic errors provide more task-specific training examples than masking. As an unsupervised approach, MAGEC is an alternative to recently proposed language model (LM) based approaches (Bryant and Briscoe, 2018; Stahlberg et al., 2019), but it does not require any amount of annotated sentences for tuning.

2 Minimally-augmented grammatical error correction

Our minimally-augmented GEC approach uses synthetic noise as its primary source of training data. We generate erroneous sentences from monolingual texts via random word perturbations selected from automatically created confusion sets. These are traditionally defined as sets of frequently confused words (Rozovskaya and Roth, 2010).

We experiment with three unsupervised methods for generating confusion sets:

[1] https://www.cl.cam.ac.uk/research/nl/bea2019st

[2] GNU Aspell supports more than 160 languages: http://aspell.net/man-html/Supported.html

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 357–363
Hong Kong, Nov 4, 2019.

Word	Confusion set
had	hard head hand gad has ha ad hat
night	knight naught nought nights bight might nightie
then	them the hen ten than thin thee thew
haben	habend halben gaben habe habet haken
Nacht	Nachts Nascht Macht Naht Acht Nach Jacht Pacht
dann	sann dank denn dünn kann wann bannen kannst
имел	им ел им-ел имела имели имело мел умел
ночь	ночью ночи дочь мочь ноль новь точь
затем	за тем за-тем затеем затеям зятем затеями

Table 1: Examples of spell-broken confusion sets for English, German and Russian.

Edit distance Confusion sets consist of words with the shortest Levenshtein distance (Levenshtein, 1966) to the selected confused word.

Word embeddings Confusion sets contain the most similar words to the confused word based on the cosine similarity of their word embedding vectors (Mikolov et al., 2013).

Spell-breaking Confusion sets are composed of suggestions from a spell-checker; a suggestion list is extracted for the confused word regardless of its actual correctness.

These methods can be used to build confusion sets for any alphabetic language.[3] We find that confusion sets constructed via spell-breaking perform best (Section 4). Most context-free spell-checkers combine a weighted edit distance and phonetic algorithms to order suggestions, which produces reliable confusion sets (Table 1).

We synthesize erroneous sentences as follows: given a confusion set $C_i = \{c_1^i, c_2^i, c_3^i, ...\}$, and the vocabulary V, we sample word $w_j \in V$ from the input sentence with a probability approximated with a normal distribution $\mathcal{N}(p_{\text{WER}}, 0.2)$, and perform one of the following operations: (1) substitution of w_j with a random word c_k^j from its confusion set with probability p_{sub}, (2) deletion of w_j with p_{del}, (3) insertion of a random word $w_k \in V$ at $j+1$ with p_{ins}, and (4) swapping w_j and w_{j+1} with p_{swp}. When making a substitution, words within confusion sets are sampled uniformly.

To improve the model's capability of correcting spelling errors, inspired by Lichtarge et al. (2018); Xie et al. (2018), we randomly perturb 10% of characters using the same edit operations as above.

[3]For languages with logosylabic writing system like Chinese, the edit distance can be calculated on transcribed text, while word embeddings can be generated after word-segmentation.

Lang.	Corpus	Dev	Test	Train
EN	W&I+LOCNESS	4,384	4,477	34,308
DE	Falco+MERLIN	2,503	2,337	18,754[4]
RU	RULEC-GEC	2,500	5,000	4,980

Table 2: Sizes of labelled corpora in no. of sentences.

Character-level noise is introduced on top of the synthetic errors generated via confusion sets.

A MAGEC model is trained solely on the synthetically noised data and then ensembled with a language model. Being limited only by the amount of clean monolingual data, this large-scale unsupervised approach can perform better than training on small authentic error corpora. A large amount of training examples increases the chance that synthetic errors resemble real error patterns and results in better language modelling properties.

If any small amount of error-annotated learner data is available, it can be used to fine-tune the pre-trained model and further boost its performance. Pre-training of decoders of GEC models from language models has been introduced by Junczys-Dowmunt et al. (2018b), we pretrain the full encoder-decoder models instead, as proposed by Grundkiewicz et al. (2019).

3 Experiments

Data and evaluation Our approach requires a large amount of monolingual data that is used for generating synthetic training pairs. We use the publicly available News crawl data[5] released for the WMT shared tasks (Bojar et al., 2018). For English and German, we limit the size of the data to 100 million sentences; for Russian, we use all the available 80.5 million sentences.

As primary development and test data, we use the following learner corpora (Table 2):

- English: the new W&I+LOCNESS corpus (Bryant et al., 2019; Granger, 1998) released for the BEA 2019 shared task and representing a diverse cross-section of English language;
- German: the Falko-MERLIN GEC corpus (Boyd, 2018) that combines two German learner corpora of all proficiency levels;

[4]The original training part of Falco+MERLIN consists of 19,237 sentences, but is contaminated with some test sentences. We have removed training examples if their target sentences occur in the development or test set.

[5]http://data.statmt.org/news-crawl

System	P	R	$F_{0.5}$
Random	32.8	6.7	18.49
Edit distance	39.9	9.5	24.27
Word embeddings	39.7	9.0	23.56
Spell-breaking	43.1	10.6	26.66
+ OOV + Case	**44.9**	10.9	27.70
$\rightarrow$ WER $= 0.25$	43.3	11.8	27.50
$\rightarrow$ Edit-weighted $\Lambda = 2$	43.0	**12.6**	**28.99**

Table 3: Performance for different confusion sets and edit weighting techniques on W&I+LOCNESS Dev.

- Russian: the recently introduced RULEC-GEC dataset (Alsufieva et al., 2012; Rozovskaya and Roth, 2019) containing Russian texts from foreign and heritage speakers.

Unless explicitly stated, we do not use the training parts of those datasets. For each language we follow the originally proposed preprocessing and evaluation settings. English and German data are tokenized with Spacy[6], while Russian is preprocessed with Mystem (Segalovich, 2003). We additionally normalise punctuation in monolingual data using Moses scripts (Koehn et al., 2007). During training, we limit the vocabulary size to 32,000 subwords computed with SentencePiece using the unigram method (Kudo and Richardson, 2018).

English models are evaluated with ERRANT (Bryant et al., 2017) using $F_{0.5}$; for German and Russian, the M2Scorer with the MaxMatch metric (Dahlmeier and Ng, 2012) is used.

Synthetic data Confusion sets are created for each language for $V = 96,000$ most frequent lexical word forms from monolingual data. We use the Levenshtein distance to generate edit-distance based confusion sets. The maximum considered distance is 2. Word embeddings are computed with *word2vec*[7] from monolingual data. To generate spell-broken confusion sets we use Enchant[8] with Aspell dictionaries.[9] The size of confusion sets is limited to top 20 words.

Synthetic errors are introduced into monolingual texts to mimic word error rate (WER) of about 15%, i.e. $p_{\text{WER}} = 0.15$, which resembles error frequency in common ESL error corpora. When confusing a word, the probability p_{sub} is set to 0.7, other probabilities are set to 0.1.

[6] https://spacy.io
[7] https://github.com/tmikolov/word2vec
[8] https://abiword.github.io/enchant
[9] ftp://ftp.gnu.org/gnu/aspell/dict

System	Dev	P	R	$F_{0.5}$
Top BEA19 (Low-res.)	44.95	70.2	48.0	64.24
Top BEA19 (Restricted)	53.00	72.3	60.1	69.47
Spell-checker	10.04	23.7	7.4	16.45
Spell-checker w/ LM	12.00	41.5	6.8	20.52
MAGEC w/o LM	28.99	53.4	26.2	44.22
MAGEC	31.87	49.1	37.5	46.22
MAGEC Ens.	33.32	53.0	34.5	47.89
Fine-tuned (Real)	44.29	61.2	54.1	59.62
Fine-tuned (Real+Synth.)	49.49	66.0	58.8	64.45

(a) English (W&I+LOCNESS)

System	Dev	P	R	$M^2_{0.5}$
Boyd (2018) (Unsup.)	—	30.0	14.0	24.37
Boyd (2018)	—	52.0	29.8	45.22
Spell-checker	20.97	33.0	9.5	22.06
Spell-checker w/ LM	24.14	43.6	8.6	24.27
MAGEC w/o LM	49.25	58.1	27.2	47.30
MAGEC	52.06	57.9	34.7	51.10
MAGEC Ens.	53.61	58.3	36.9	**52.22**
Fine-tuned (Real)	68.13	72.2	54.0	67.67
Fine-tuned (Real+Synth.)	70.51	73.0	61.0	70.24

(b) German (Falko-MERLIN)

System	Dev	P	R	$M^2_{0.5}$
Rozovskaya and Roth (2019)	—	38.0	7.5	21.0
Spell-checker	18.32	19.2	7.2	14.39
Spell-checker w/ LM	22.01	30.7	7.5	18.99
MAGEC w/o LM	24.82	30.1	20.4	27.47
MAGEC	27.13	32.3	29.5	31.71
MAGEC Ens.	27.87	33.3	29.4	**32.41**
Fine-tuned (Real)	30.28	35.4	31.1	34.45
Fine-tuned (Real+Synth.)	30.64	36.3	28.7	34.46

(c) Russian (RULEC-GEC)

Table 4: Unsupervised and fine-tuned MAGEC systems for English, German and Russian, contrasted with systems from related work and spell-checking baselines.

Training settings We adapt the recent state-of-the-art GEC system by Junczys-Dowmunt et al. (2018b), an ensemble of sequence-to-sequence Transformer models (Vaswani et al., 2017) and a neural language model.[10]

We use the training setting proposed by the authors[11], but introduce stronger regularization: we increase dropout probabilities of source words to 0.3, add dropout on transformer self-attention and filter layers of 0.1, and use larger mini-batches with

[10] Models and outputs are available from https://github.com/grammatical/magec-wnut2019
[11] https://github.com/grammatical/neural-naacl2018

~2,500 sentences. We do not pre-train the decoder parameters with a language model and train directly on the synthetic data. We increase the size of language model used for ensembling to match the Transformer-big configuration (Vaswani et al., 2017) with 16-head self-attention, embeddings size of 1024 and feed-forward filter size size of 4096. In experiments with fine-tuning, the training hyperparameters remain unchanged.

All models are trained with Marian (Junczys-Dowmunt et al., 2018a). The training is continued for at most 5 epochs or until early-stopping is triggered after 5 stalled validation steps. We found that using 10,000 synthetic sentences as validation sets, i.e. a fully unsupervised approach, is as effective as using the development parts of error corpora and does not decrease the final performance.

4 Results and analysis

Confusion sets On English data, all proposed confusion set generation methods perform better than random word substitution (Table 3).Confusion sets based on word embeddings are the least effective, while spell-broken sets perform best at 26.66 $F_{0.5}$. We observe further gains of +1.04 from keeping out-of-vocabulary spell-checker suggestions (OOV) and preserving consistent letter casing within confusion sets (Case).

The word error rate of error corpora is an useful statistic that can be used to balance precision/recall ratios (Rozovskaya and Roth, 2010; Junczys-Dowmunt et al., 2018b; Hotate et al., 2019). Increasing WER in the synthetic data from 15% to 25% increases recall at the expense of precision, but no overall improvement is observed. A noticeable recall gain that transfers to a higher F-score of 28.99 is achieved by increasing the importance of edited fragments with the edit-weighted MLE objective from Junczys-Dowmunt et al. (2018b) with $\Lambda = 2$. We use this setting for the rest of our experiments.

Main results We first compare the GEC systems with simple baselines using a greedy and context spell-checking (Table 4); the latter selects the best correction suggestion based on the sentence perplexity from a Transformer language model. All systems outperform the spell-checker baselines.

On German and Russian test sets, single MAGEC models without ensembling with a language model already achieve better performance than reported by Boyd (2018) and Rozovskaya and Roth (2019) for their systems that use authentic error-annotated data for training (Table 4b and 4c). Our best unsupervised ensemble systems that combine three Transformer models and a LM[12] outperform the state-of-the-art results for these languages by +7.0 and +11.4 $F_{0.5}$.

Our English models do not compete with the top systems (Grundkiewicz et al., 2019) from the BEA shared task trained on publicly available error-annotated corpora (Table 4a). It is difficult to compare with the top low-resource system from the shared task, because it uses additional parallel data from Wikipedia (Grundkiewicz and Junczys-Dowmunt, 2014), larger ensemble, and n-best list re-ranking with right-to-left models, which can be also implemented in this work.

MAGEC systems are generally on par with the results achieved by a recent unsupervised contribution based on finite state transducers by Stahlberg et al. (2019) on the CoNLL-2014 (Dahlmeier et al., 2013) and JFLEG test sets (Napoles et al., 2017) (Table 5).

System	CoNLL	JFLEG
Bryant and Briscoe (2018) ⋆	34.09	48.75
Stahlberg et al. (2019) ⋆	44.43	52.61
Stahlberg et al. (2019) (659K real data)	58.40	58.63
MAGEC Ens. ⋆	44.23	56.18
MAGEC Fine-tuned (34K real data)	56.54	60.01

Table 5: Comparison with LM-based GEC on the CoNLL (M^2) and JFLEG (GLEU) test sets for unsupervised (⋆) and supervised systems trained or fine-tuned on different amounts of labelled data.

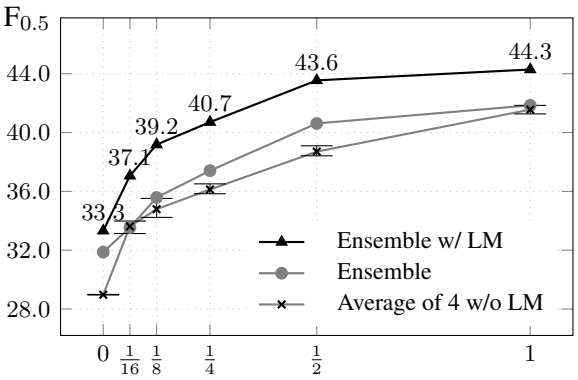

Figure 1: Improvements from fine-tuning on subsets of W&I+LOCNESS Train. The smallest $\frac{1}{16}$ part of the dataset contains 2,145 sentences. Averaged F-scores over 4 runs trained on different subsets of the data.

[12]The weight of the language model is grid-searched on the development set.

Lang.	Spell.+punc. P	R	$F_{0.5}$	Other errors P	R	$F_{0.5}$
EN	28.8	24.1	27.68	33.5	16.8	27.93
DE	54.8	71.4	57.43	63.6	55.8	61.83
RU	26.7	75.0	**30.70**	14.6	19.7	**15.37**

Table 6: Performance of single MAGEC w/ LM models on two groups of errors on respective development sets.

All unsupervised systems benefit from domain-adaptation via fine-tuning on authentic labelled data (Miceli Barone et al., 2017). The more authentic high-quality and in-domain training data is used, the greater the improvement, but even as few as ~2,000 sentences are helpful (Fig. 1). We found that fine-tuning on a 2:1 mixture of synthetic and oversampled authentic data prevents the model from over-fitting. This is particularly visible for English which has the largest fine-tuning set (34K sentences), and the difference of 5.2 $F_{0.5}$ between fine-tuning with and without synthetic data is largest.

Spelling and punctuation errors The GEC task involves detection and correction of all types of error in written texts, including grammatical, lexical and orthographical errors. Spelling and punctuation errors are among the most frequent error types and also the easiest to synthesize.

To counter the argument that – mostly due to the introduced character-level noise and strong language modelling – MAGEC can only correct these "simple" errors, we evaluate it against test sets that contain either spelling and punctuation errors or all other error types; with the complement errors corrected (Table 6). Our systems indeed perform best on misspellings and punctuation errors, but are capable of correcting various error types. The disparity for Russian can be explained by the fact that it is a morphologically-rich language and we suffer from generally lower performance.

5 Conclusions and future work

We have presented Minimally-Augmented Grammatical Error Correction (MAGEC), which can be effectively used in both unsupervised and low-resource scenarios. The method is model independent, requires easily available resources, and can be used for creating reliable baselines for supervised techniques or as an efficient pre-training method for neural GEC models with labelled data. We have demonstrated the effectiveness of our method and outperformed state-of-the-art results for German and Russian benchmarks, trained with labelled data, by a large margin.

For future work, we plan to evaluate MAGEC on more languages and experiment with more diversified confusion sets created with additional unsupervised generation methods.

References

Anna Alsufieva, Olesya Kisselev, and Sandra Freels. 2012. Results 2012: Using flagship data to develop a Russian learner corpus of academic writing. *Russian Language Journal*, 62:79–105.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

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.

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.

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 *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126, Gothenburg, Sweden. Association for Computational Linguistics.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Fluency boost learning and inference for neural grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1055–1065, Melbourne, Australia. Association for Computational Linguistics.

Sylviane Granger. 1998. The computer learner corpus: A versatile new source of data for SLA research. In Sylviane Granger, editor, *Learner English on Computer*, pages 3–18. Addison Wesley Longman, London and New York.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing — Lecture Notes in Computer Science*, volume 8686, pages 478–490. Springer.

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.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Kengo Hotate, Masahiro Kaneko, Satoru Katsumata, and Mamoru Komachi. 2019. Controlling grammatical error correction using word edit rate. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 149–154. Association for Computational Linguistics.

Phu Mon Htut and Joel Tetreault. 2019. The unbearable weight of generating artificial errors for grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 478–483, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018a. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018b. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Jared Lichtarge, Christopher Alberti, Shankar Kumar, Noam Shazeer, and Niki Parmar. 2018. Weakly supervised grammatical error correction using iterative decoding. *CoRR*, abs/1811.01710.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 2017 Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain. Association for Computational Linguistics.

Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. 2017. Artificial error generation with machine translation and syntactic patterns. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 287–292, Copenhagen, Denmark. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970, Cambridge, MA. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically-rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

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

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. CSREA Press.

Felix Stahlberg, Christopher Bryant, and Bill Byrne. 2019. Neural grammatical error correction with finite state transducers. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. 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 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

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. *CoRR*, abs/1903.00138.