Transfer Fine-tuning for Quality Estimation of Text Simplification

Yuki Hironaka, Tomoyuki Kajiwara, Takashi Ninomiya
Graduate School of Science and Engineering, Ehime University, Japan
{hironaka@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp

Abstract

To efficiently train quality estimation of text simplification on a small-scale labeled corpus, we train sentence difficulty estimation prior to fine-tuning the pre-trained language models. Our proposed method improves the quality estimation of text simplification in the framework of transfer fine-tuning, in which pre-trained language models can improve the performance of the target task by additional training on the relevant task prior to fine-tuning. Since the labeled corpus for quality estimation of text simplification is small (600 sentence pairs), an efficient training method is desired. Therefore, we propose a training method for pseudo quality estimation that does not require labels for quality estimation. As a relevant task for quality estimation of text simplification, we train the estimation of sentence difficulty. This is a binary classification task that identifies which sentence is simpler using an existing parallel corpus for text simplification. Experimental results on quality estimation of English text simplification showed that not only the quality estimation performance on simplicity that was trained, but also the quality estimation performance on fluency and meaning preservation could be improved in some cases.

Keywords: Transfer Fine-tuning, Quality Estimation, Text Simplification

1. Introduction

Text simplification (Alva-Manchego et al., 2020) is the task that paraphrases complex expressions into simpler ones while preserving their meaning. Automatic sentence simplification contributes to learning and reading support for children (De Belder and Moens, 2010) and language learners (Petersen and Ostendorf, 2007) as well as improves the performance of other natural language processing tasks such as relation extraction (Miwa et al., 2010) and machine translation (Štajner and Popovic, 2016).

The quality of text simplification models has been evaluated by human evaluation in terms of fluency, meaning preservation, and simplicity, and by automatic evaluation such as SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002) based on reference sentences and readability metrics such as FKGL (Kincaid et al., 1975). However, human evaluation has problems with cost and reproducibility, while automatic evaluation has a low correlation with human evaluation (Sulem et al., 2018; Tanprasert and Kauchak, 2021). In addition, when text simplification models are used in the real world, users often do not have reference sentences, so automatic evaluation based on reference sentences such as SARI cannot be used. Therefore, reference-less quality estimation (QE) for text simplification (Štajner et al., 2016; Kajiwara and Fujita, 2017; Martin et al., 2018; Alva-Manchego et al., 2021) has been studied.

Existing QE methods for text simplification (Kajiwara and Fujita, 2017; Martin et al., 2018) trained machine learning models with feature extraction using evaluation metrics based on word embeddings and word matching ratio. Although it is expected that the QE performance can be improved by employing context-aware deep learning models, this is difficult due to the small-scale of the labeled data for this task. Two existing datasets for QE of text simplification, QATS[1] (Štajner et al., 2016) targets models based on statistical machine translation and Simplicity-DA[2] (Alva-Manchego et al., 2021) targets models based on neural machine translation, both consisting of about 600 sentence pairs, which is small-scale to sufficiently train QE models based on deep learning.

To address this problem, we train the relevant task (pseudoQE) prior to QE training. This facilitates efficient training on a small-scale labeled corpus for QE of text simplification. As a pseudo-QE task, we propose the related task of identifying complex and simple sentences using an existing large-scale parallel corpus for text simplification (Jiang et al., 2020). Experimental results on QE of English text simplification using the Simplicity-DA dataset (Alva-Manchego et al., 2021) showed that QE performance on simplicity was improved. Moreover, beyond expectations, some deep learning models showed improvements in fluency and meaning preservation.

2. Related Work

2.1. Quality Estimation for Simplification

Text simplification as a sequence-to-sequence task has been studied based on monolingual parallel

[1] https://qats2016.github.io/
[2] https://github.com/feralvam/metaeval-simplification

LREC-COLING 2024, pages 16738–16744

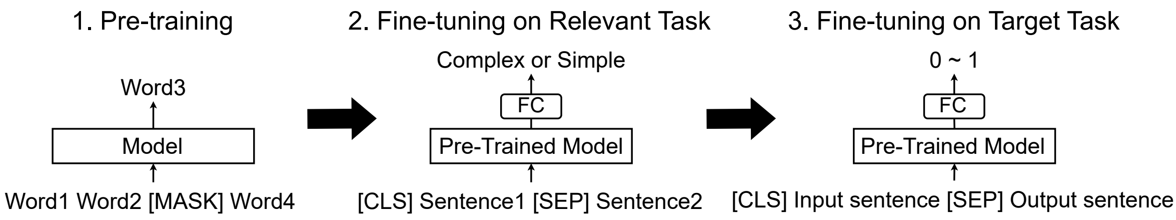

Figure 1: Overview of the proposed method. In both fine-tuning tasks, the latter sentence is evaluated.

corpora consisting of complex and simple sentences (Coster and Kauchak, 2011; Xu et al., 2015). In the early 2010s (Specia, 2010; Wubben et al., 2012; Narayan and Gardent, 2014; Štajner et al., 2015; Kajiwara and Komachi, 2016), text simplification based on phrase-based statistical machine translation (Koehn et al., 2003) has been studied. Since the late 2010s (Nisioi et al., 2017; Zhang and Lapata, 2017; Dong et al., 2019; Kriz et al., 2019; Nishihara et al., 2019), following the success of neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), text simplification based on recurrent neural networks has been studied. In recent years (Zhao et al., 2018; Kajiwara, 2019; Martin et al., 2020; Maddela et al., 2021; Yanamoto et al., 2022), text simplification based on the Transformer model (Vaswani et al., 2017) has become mainstream, as have other sequence-to-sequence tasks such as machine translation.

QE is a task of estimating the quality of the output sentences from the input and output sentence pairs. Previous QE studies for text simplification have trained machine learning models, such as support vector machine and ridge regression, on QATS dataset (Štajner et al., 2016) for evaluating text simplification models based on statistical machine translation. Kajiwara and Fujita (2017) performed QE as a classification model of Good, OK, and Bad based on word embeddings-based feature extraction (Mikolov et al., 2013). Martin et al. (2018) performed feature extraction based on machine translation evaluation metrics such as BLEU (Papineni et al., 2002) and readability metrics such as FKGL (Kincaid et al., 1975) for both regression and classification QE. Alva-Manchego et al. (2021) constructed the Simplicity-DA dataset for evaluating text simplification models based on deep learning. Unlike QATS, which targets text simplification models based on statistical machine translation, Simplicity-DA targets recent text simplification models, but like QATS, it is small-scale, at about 600 sentence pairs. Efficient training methods are desired for high-quality QE from small-scale labeled corpora.

2.2. Transfer Fine-Tuning

In recent natural language processing, transfer learning approaches, in which pre-trained models such as masked language models (Devlin et al., 2019; Liu et al., 2019; He et al., 2021) are fine-tuned on the target task, have achieved high performance in a variety of applications (Wang et al., 2019). Its performance can be further improved by training on a task with similar characteristics to the target task before fine-tuning, which is called transfer fine-tuning (Arase and Tsujii, 2019). Masked language modeling at the sentence level for the summarization task (Zhang et al., 2020) and reconstruction of round-trip translations for the paraphrase generation task (Kajiwara et al., 2020) have been reported to be effective as pre-training with similar characteristics to the target task, respectively. Transfer fine-tuning is also effective for classification and regression tasks. For example, additional training to classify paraphrases between pre-training and fine-tuning can improve the performance of sentence similarity estimation (Arase and Tsujii, 2019). However, effective additional training methods have not been identified in transfer fine-tuning for the QE of text simplification task.

3. Proposed Method

In this study, we train QE models for text simplification by fine-tuning a pre-trained model in two steps as shown in Figure 1. While labeled corpora for QE of text simplification are available only on a small-scale, our additional task does not require QE labels and uses only existing parallel corpora for text simplification, allowing it to be trained on a large-scale. Following previous studies (Štajner et al., 2016; Kajiwara and Fujita, 2017; Martin et al., 2018), we train each QE model on the aspects of fluency, meaning preservation, and simplicity.

3.1. Pre-training

We employ the Transfomer encoder (Vaswani et al., 2017) for our QE model. To train efficiently from a small-scale labeled corpus, we first pre-train our QE model on a large-scale raw corpus. Although QE models can be pre-trained on any task, this study

Task	Dataset	Type	Sentences
pseudoQE	Wiki-Auto	Train	488,332
	Turk Corpus	Dev	2,000
	Newsela-Auto	Train	394,300
		Dev	43,317
QE	Simplicity-DA	Train	400
		Dev	100
		Test	100

Table 1: Corpus size

employs masked language modeling and uses pre-trained models such as BERT (Devlin et al., 2019).

3.2. Fine-tuning on Relevant Task

To address the low-resource problem in QE for text simplification, we train the pre-trained model on the pseudo-QE task before fine-tuning it on the QE-labeled corpus. As shown in the center of Figure 1, sentence pairs of complex and simple sentences are concatenated and input into the QE model to train a binary classification of whether the latter sentence is more complex or simpler. Since such a task of sentence difficulty estimation is similar to the task of QE for simplicity, this additional training of pseudo-QE can be expected to improve the performance of QE for simplicity of text simplification. Note that our pseudo-QE training does not require manually labeled QE labels, and only an existing parallel corpus for text simplification (e.g., Wiki-Auto (Jiang et al., 2020) or Newsela (Xu et al., 2015)) is required, which allows for large-scale training at low cost.

3.3. Fine-tuning on Target Task

For our QE model fine-tuned on the relevant task in the previous section, we finetune it on the actual QE task using sentence pairs of input sentences and output sentences of the text simplification system, as shown in the right side of Figure 1. We expect that QE models that can be evaluated at a coarse level by training in the pseudo-QE task can be evaluated at a finer level by fine-tuning on the actual QE task.

4. Experiment

This experiment evaluates sentence-level QE of English text simplification on the Simplicity-DA dataset (Alva-Manchego et al., 2021). We trained each regression model on the aspects of fluency, meaning preservation, and simplicity. The performance of QE models was automatically evaluated using the Pearson correlation between predicted scores and human labels.

	F	M	S
Kajiwara-17	0.405	0.670	0.373
Martin-18	0.462	0.680	0.320
BERT	0.766	0.638	0.482
+ pseudoQE (Wiki)	0.739	**0.710**	**0.503**
+ pseudoQE (News)	0.679	**0.734**	0.470
RoBERTa	0.790	0.779	0.543
+ pseudoQE (Wiki)	0.741	0.738	0.517
+ pseudoQE (News)	0.746	0.764	**0.568**
DeBERTa	0.716	0.734	0.473
+ pseudoQE (Wiki)	**0.754**	0.728	**0.522**
+ pseudoQE (News)	0.682	**0.766**	**0.519**

Table 2: QE performance by Pearson correlation coefficient. F: Fluency, M: Meaning preservation, and S: Simplicity. Values that are improved over the baseline model are in bold, and the highest performance is highlighted by underlining.

4.1. Setting

Data Table 1 shows the number of sentence pairs for the datasets used in this experiment. For the pseudo-QE task, we used two parallel corpora, Wikipedia and Newsela, which are commonly used for training English text simplification models. For Wikipedia, we used Wiki-Auto[3] (Jiang et al., 2020) for training and Turk Corpus[4] (Xu et al., 2016) for validation. For Newsela, we used Newsela-Auto[3] (Jiang et al., 2020) for both training and validation. For fine-tuning on the QE task, we used Simplicity-DA dataset[2] (Alva-Manchego et al., 2021). This dataset consisted of 600 sentence pairs, randomly divided into 400 for training and 100 each for validation and evaluation.

Model We began training QE models from three pre-trained models: BERT[5] (Devlin et al., 2019), RoBERTa[6] (Liu et al., 2019), and DeBERTa[7] (He et al., 2021). We implemented each model using HuggingFace Transformers (Wolf et al., 2020). For each pre-trained model, we trained three QE models: baseline, which fine-tunes only QE task, pseudoQE (Wiki), which applies our proposed method with Wikipedia, and pseudoQE (News), which applies our proposed methods with Newsela.

For the pseudo-QE task, we trained three epochs

[3] https://github.com/chaojiang06/wiki-auto
[4] https://github.com/cocoxu/simplification
[5] https://huggingface.co/bert-base-uncased
[6] https://huggingface.co/roberta-base
[7] https://huggingface.co/microsoft/deberta-base

of cross-entropy loss minimization with a batch size of 1,024, a learning rate of 5e-5, and the optimization method AdamW (Loshchilov and Hutter, 2019). The model in the epoch with the highest accuracy on the validation data was then used for the fine-tuning of the QE task. For the QE task, we trained for mean squared error minimization with a batch size of 32 and the optimization method AdamW. Training stopped when the Pearson correlation in the validation data stopped improving for 10 epochs. Four learning rates were tried: 5e-5, 4e-5, 3e-5, and 2e-5, and we used the model with the highest Pearson correlation on the validation data. For all models, we train five times each with changing random seeds and report the average score of the three models excluding the maximum and minimum values.

Comparative Model We compare the performance of two existing methods based on machine learning with our method based on deep learning. The Kajiwara-17 model (Kajiwara and Fujita, 2017) was implemented using scikit-learn.[8] The Martin-18 model (Martin et al., 2018) was implemented using their implementation.[9]

4.2. Result

Experimental results are shown in Table 2. For all pre-trained models, the proposed method (+pseudoQE) was able to improve QE performance on simplicity in at least one of the domains. Since the proposed method added training related to the QE of simplicity, we expected to improve the QE performance on simplicity. However, beyond our expectations, the QE performance on fluency of DeBERTa and meaning preservation of BERT and DeBERTa were also improved.

4.3. Analysis

Figure 2 shows the change in QE performance when the amount of training data for the pseudo-QE task is reduced to 250,000, 125,000, 50,000, and 25,000 sentence pairs. We found that for all pre-trained models, the impact of the proposed method peaks at 50,000 to 100,000 sentence pairs of training. Therefore, there is no need to prepare a large-scale parallel corpus for text simplification with more than 100,000 sentence pairs. Since parallel corpora for text simplification on the scale of tens of thousands of sentence pairs are available for languages other than English, such as Italian (Brunato et al., 2016) and Japanese (Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018), our method may be applicable to QE of text simplification in other languages.

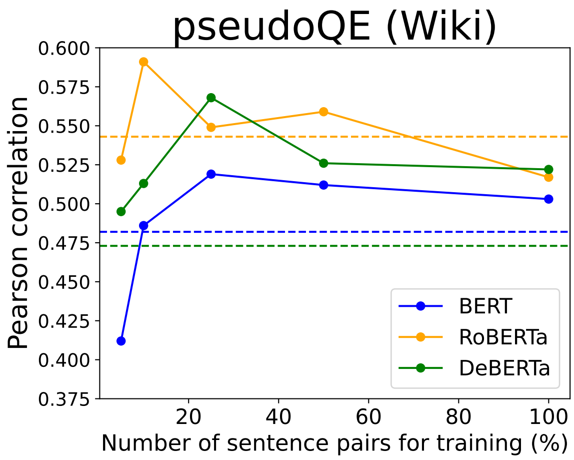

Figure 2: Analysis of training data size and QE performance.

We observe the number of sentence pairs in the parallel corpus for text simplification used for additional training for each model. For BERT, the proposed method can outperform the QE performance of the baseline when using a parallel corpus of 50,000 sentence pairs or more. For DeBERTa, the proposed method can outperform the QE performance of the baseline even with a parallel corpus of 25,000 sentence pairs. Although RoBERTa achieves the highest performance, there is a large variation in performance for each experiment.

5. Conclusion

To efficiently train QE models for text simplification with small-scale labeled corpora, we proposed transfer fine-tuning, in which pre-trained models are additionally trained with a pseudo-QE task prior to fine-tuning. As a pseudo-QE task, the proposed method trains a binary classification that identifies which sentence is simpler using a general parallel corpus for text simplification without QE labels.

Experimental results on English text simplification showed that the proposed method not only improves QE performance on simplicity, but also improves fluency and meaning preservation, depending on the pre-trained model. Our detailed analysis reveals that a parallel corpus of text simplification for additional training is enough on the scale of tens of thousands of sentence pairs. This is the size of the corpus also accessible in languages other than English.

Our future work includes designing additional training methods that focus on fluency and meaning preservation, as well as working on quality estimation of text simplification in non-English languages.

[8] https://scikit-learn.org

[9] https://github.com/facebookresearch/text-simplification-evaluation

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP22H03651.

Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer Fine-Tuning: A BERT Case Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5393–5404.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. PaCCSS-IT: A Parallel Corpus of Complex-Simple Sentences for Automatic Text Simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*, pages 19–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *Proceedings of the Ninth International Conference on Learning Representations*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.

Tomoyuki Kajiwara. 2019. Negative Lexically Constrained Decoding for Paraphrase Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052.

Tomoyuki Kajiwara and Atsushi Fujita. 2017. Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 109–115.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158.

Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. 2020. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8042–8049.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 461–466.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical report, Defence Technical Information Center (DTIC) Document*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-Weighted Loss and Diverse Reranking for Sentence Simplification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3137–3147.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the Seventh International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable Text Simplification with Explicit Paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable Sentence Simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less Quality Estimation of Text Simplification Systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation*, pages 29–38.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified Corpus with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1153–1160.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations*.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-Focused Sentence Simplification for Relation Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 788–796.

Shashi Narayan and Claire Gardent. 2014. Hybrid Simplification using Deep Semantics and Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 435–445.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 85–91.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Sarah E Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of the Workshop on Speech and Language Technology in Education*, pages 69–72.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, pages 30–39.

Sanja Štajner, Hannah Béchara, and Horacio Saggion. 2015. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 823–828.

Sanja Štajner and Maja Popovic. 2016. Can Text Simplification Help Machine Translation? In *Proceedings of the 19th Annual Conference of the*

European Association for Machine Translation, pages 230–242.

Sanja Štajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared Task on Quality Assessment for Text Simplification. In *Proceedings of Shared Task on Quality Assessment for Text Simplification*, pages 22–31.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is Not Suitable for the Evaluation of Text Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 28th Conference on Neural Information Processing Systems*, pages 3104–3112.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 1–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the Seventh International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable Text Simplification with Deep Reinforcement Learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 398–404.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the Thirty-seventh International Conference on Machine Learning*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating Transformer and Paraphrase Rules for Sentence Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173.