Experimenting Sentence Split-and-Rephrase Using Part-of-Speech Labels

P. Berlanga Neto and E. Y. Okano and E. E. S. Ruiz

Departamento de Computação e Matemática, FFCLRP
Universidade de São Paulo (USP).
Av. Bandeirantes, 3900, Monte Alegre. 14040-901, Ribeirão Preto, SP – Brazil
[pauloberlanga, okano700, evandro]@usp.br

Abstract. Text simplification (TS) is a natural language transformation process that reduces linguistic complexity while preserving semantics and retaining its original meaning. This work aims to present a research proposal for automatic simplification of texts, precisely a split-and-rephrase approach based on an encoder-decoder neural network model. The proposed method was trained against the WikiSplit English corpus with the help of a part-of-speech tagger and obtained a BLEU score validation of 74.72%. We also experimented with this trained model to split-and-rephrase sentences written in Portuguese with relative success, showing the method's potential.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Keywords: natural language processing, neural networks, sentence simplification

1. INTRODUCTION

Text Simplification (TS) is the process of modifying natural language to reduce complexity and improve both readability and understandability [Shardlow 2014]. A simplified vocabulary or a simplified text structure can benefit different publics, such as people with low education levels, children, non-native individuals, people who have learning disorders (such as autism, dyslexia, and aphasia), among others [Štajner et al. 2015]. Although the simplicity of a text seems intuitively obvious, it does not have a precise definition in technical terms. Traditional assessment metrics consider factors such as sentence length, syllable count, and other linguistic features of the text to identify it as elaborate or not [Shardlow 2014].

In the last decades, several models and systems have been developed to improve the task of automatic text simplification. They are primarily based on two main approaches: lexical simplification (LS) and/or syntactic simplification (SS) [Shardlow 2014]. Lexical simplification has the goal to identify and replace words or expressions for synonyms that can be understood by a larger audience [Hartmann et al. 2018; Shardlow 2014]. On the other hand, syntactic simplification must identify grammatical complexities presented in the sentences and rewrite them in simpler structures [Tajner and Glava 2017; Scarton et al. 2019].

Since complex texts often contain a portion of simple sentences in their structure, some late approaches have focused on the analysis of specific aspects at the sentence level, expanding a study branch known as sentence simplification (SentS). Split-and-rephrase is a SentS method proposed by Narayan and colleagues [Narayan et al. 2017] that aims to split a complex sentence into shorter sentences while preserving the meaning of the original sentence. This method conceptualizes the notion that shorter sentences are generally better understood by the majority of the individuals. Nevertheless, it may also be easier to be processed by NLP (Natural Language Processing) systems, facilitating tasks such as parsers, semantic role labelers, and machine translation systems.

In this paper, we make the main contribution in exploring split-and-rephrase from word labels gained from a part-of-speech tagger. We focus on coordinate clauses, a relative dependence clause to the main sentence clause. Also, we add a minor improvement to this experiment by performing a simple transfer learning. We train split-and-rephrase systems in English sentences and apply the learned knowledge to Portuguese sentences. Based on the literature surveyed, this is the first reference to an automatic split-and-rephrase method successfully used to the Portuguese language. In the next section, we describe some previous work on text simplification. In Section 3 we present the WikiSplit corpus [Botha et al. 2018] and our suggestion of simplification process. In Section 4 we present the results obtained by our proposed model. In Section 5, we briefly discuss these results, while in Section 6 we have the conclusion and the expected challenges to the future.

2. RELATED WORK

Advanced neural computational methods have transformed the task of text simplification. For a survey of the academic work before 2014, see the excellent survey from Advaith Siddharthan [Siddharthan 2014]. A quick search on Google Scholar for the string query *"Text Simplification" OR "Lexical Simplification" OR "Syntactic Simplification"* for documents published up to 2014 resulted in 3,200 hits.

Wang and colleagues [Wang et al. 2016] affirm that, up to 2016, some of the TS techniques were limited to either lexical-level applications or manually defining a large number of rules. In this same paper, the authors propose to use a Long Short-Term Memory (LSTM) encoder-decoder neural model for sentence-level TS. By applying this model, they examined operation rules such as reversing, sorting and replacing constituents from sequence pairs, which has the potential of sentence simplification.

Later, Nisioi and collaborators [Nisioi et al. 2017] present another attempt at using encoder-decoder neural networks to model TS, named Neural Text Simplification. They apply this model to simultaneously perform lexical simplification and content reduction, inspired by the success observed in the Neural Machine Translation approach [Bahdanau et al. 2014].

Vu and colleagues [Vu et al. 2018] also worked to simplify the content and structure of complex sentences. For this reason, they adopt an architecture with augmented memory capacities called Neural Semantic Encoders [Munkhdalai and Yu 2017] for sentence simplification. They have demonstrated the effectiveness of their approach by automatic evaluation measures and human judgments.

As for the simplification of texts written in Portuguese, we highlight the work of Hartman et al. [Hartmann et al. 2018] that targets lexical simplification compiling the SIMPLEX-PB, the first available corpus of lexical simplification for Brazilian Portuguese. In the article by Scarton et al. [Scarton et al. 2010], the authors present the 'Simplifica' tool, also as an integral part of the PorSimples project [Aluísio et al. 2008]. This technology encourages writers to write simplified texts in Brazilian Portuguese. The tool has two modules, one for simplifying lexical terms and the other for assessing the complexity of the input texts.

Concerning datasets for research in the split-and-rephrase task, until very recently, the WebSplit corpus introduced by Narayan et al. [Narayan et al. 2017] was the main corpus used as a benchmark for the split-and-rephrase job, nowadays WikiSplit [Botha et al. 2018] is the latest reference corpus, and it is the dataset used in this paper.

3. DATA AND METHODS

WikiSplit, by Botha and his AI team at Google [Botha et al. 2018], is a corpus for the split-and-rephrase task. It is composed of one million naturally occurring sentence rewrites obtained from mining English Wikipedia's edit history. WikiSplit is a public corpus, freely available[1], and licensed under CC BY-SA 4.0 [2].

The dataset is released as text files formatted as tab-separated values. It contains 1,004,994 English sentences, each split into two sentences that together preserve the original meaning. This corpus is divided into four datasets, the training dataset with 989,944 sentences, and the other three datasets, tune, validation, and test, containing 5,000 sentences each. The sentences below are an example of one may see in this corpus:

> Street Rod is the first in a series of two games released for the PC and Commodore 64 in 1989 .
>
> Street Rod is the first in a series of two games . <::::> It was released for the PC and Commodore 64 in 1989 .

The string <::::> marks the start of the split-up sentences.

We define the split-and-rephrase task as follows. Given a complex sentence C, the goal is to produce a simplified text T consisting of a sequence of sentences $T_1, T_2, \ldots, T_n$, $n \geq 2$, in such a way that T preserves the meaning of C.

3.1 Sentence selection

Given the vast amount of aligned sentence pairs in the WikiSplit corpus, two specific cuts were made in the original training dataset to train the proposed model. The first cut was selecting the alignments with a length of, at least 15, and a maximum of 30 time steps per sentence. We considered only sentences that had equivalent counts between NLTK[3] and Spacy[4] tokenizers with no special characters. This first cut selected 171,133 alignments.

The second cut aimed to individually select compound sentences formed by a main clause and a coordinate clause. See the highlighted example below extracted from the WikiSplit corpus. This second cut extracted 63,623 alignments, thus consolidating the final selection.

Original sentence
Jes was recruited to be on the Rock Of Love show while she was bartending in downtown Chicago at a bar called Rizzo's.

Original split-and-rephrased sentences
Jes was recruited to be on the Rock Of Love show. <::::> **She** was bartending in down town Chicago at a bar called Rizzo's.

Regarding the part-of-speech classification, we adopted the Spacy POS tagger[5]. Contrary to some approaches that seek training the model with an extensive vocabulary, we generalize this learning solely by grammatical classes and by their respective recurrences. This way, we obtained a small set of attributes capable of optimizing training times. It took approximately two hours for the training

[1] https://github.com/google-research-datasets/wiki-split

[2] https://creativecommons.org/licenses/by-sa/4.0/

[3] https://www.nltk.org/

[4] https://spacy.io/

[5] https://spacy.io/api/tagger/

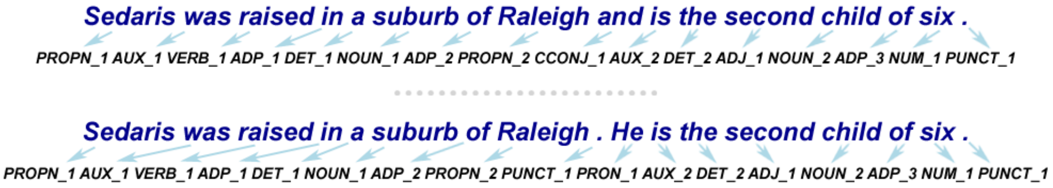
Sedaris was raised in a suburb of Raleigh and is the second child of six .
PROPN_1 AUX_1 VERB_1 ADP_1 DET_1 NOUN_1 ADP_2 PROPN_2 CCONJ_1 AUX_2 DET_2 ADJ_1 NOUN_2 ADP_3 NUM_1 PUNCT_1
Sedaris was raised in a suburb of Raleigh . He is the second child of six .
PROPN_1 AUX_1 VERB_1 ADP_1 DET_1 NOUN_1 ADP_2 PROPN_2 PUNCT_1 PRON_1 AUX_2 DET_2 ADJ_1 NOUN_2 ADP_3 NUM_1 PUNCT_1

Fig. 1. Sample of the assigned parts-of-speech tags used to train the encoder-decoder neural model.

process in a multi-user computer environment. By using POS tags, the learned model may also split sentences in other languages than the training dataset itself, as the Brazilian Portuguese.

3.2 Model specification

We follow the proposal of Bahdanau and colleagues [Bahdanau et al. 2014], implementing an encoder-decoder neural network model based on the sequence-to-sequence (seq2seq) architecture, composed of recurrent neural networks with GRU (Gated Recurrent Unit) gating mechanisms [Cho et al. 2014]. This seq2seq composition is an appropriate architecture for training an aligned corpus comprising original sentences and their simplified, split versions.

Unlike traditional sequence-to-sequence approaches, which promote the compression of original sequences into a fixed context vector, the method proposed by Bahdanau an co-authors [Bahdanau et al. 2014] presents the attention mechanism that suggests associations between specific context vectors at each of the output time steps. This mechanism makes it possible to establish references at particular points in the original sequences, and enable the transmission of these instances to the decoder outputs.

In this experiment, the proposed architecture configuration had an attention layer connected to encoder-decoder GRU layers, both composed of 50 units. We used a batch size of 200 and trained the model in 1,000 epochs. One hundred thirty-nine (139) different labels generated by the part-of-speech tagger represented the vocabulary size. They were composed of grammatical classes and numbers (see Figure 1 example), together with the wildcard character '*' for padding. We applied Adam optimization algorithm [Kingma and Ba 2015] to update the weights of the networks iteratively and a categorical-cross entropy loss function. The Keras library[6] was used.

4. RESULTS

We trained the model using 63,000 random sentence pairs from the selected data (around 99% of the alignments). We recall that these selected sentences are composed of a main clause and a coordinate clause. We validated the method in the remaining 623 sentence pairs following a similar approach adopted by Botha and colleagues [Botha et al. 2018]. Even though this validation set seems small, there were enough predictions to analyze the expected split behavior.

Although some studies consider human judgments on grammaticality, meaning preservation and simplicity the most reliable method for evaluating the sentence simplification task, it is a common practice to use automatic metrics [Alva-Manchego et al. 2019]. Following the WikiSplit work, we adopted the BLEU (Bilingual Evaluation Understudy) method [Papineni et al. 2002] to validate the results. This metric was originally created to analyze results obtained by translation algorithms, but nowadays is also used by the literature [Vu et al. 2018; Nisioi et al. 2017; Štajner et al. 2015] to evaluate text simplification.

[6] https://keras.io/

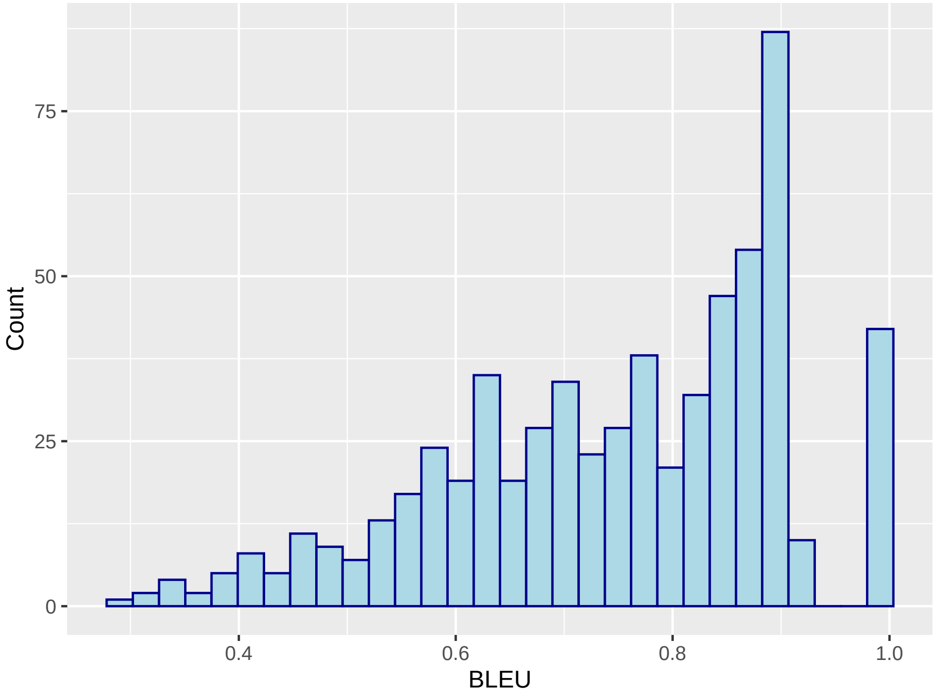

Fig. 2. Histogram of the BLEU scores for the 623 experiment sentences.

Using the proposed model against the validation set, we obtained a BLEU score of 74.72 ± 15.17. In Table I, we present some example results. The 'Input' corresponds to the original sentence from WikiSplit. The 'Reference' corresponds to the split version from WikiSplit, and the 'Predicted' is the result of the proposed approach. Figure 2 show a histogram of the BLEU scores for all the 623 validation sentences in the experiment. One may notice that there are many sentences with a perfect BLEU score. The complete log containing all the 623 predictions is publicly available[7].

The extracted examples in Table I show some behaviors that kept a high BLEU score and others that decreased its count. In Example 1, the predicted sentence was strictly the same as the expected reference sentence, thus keeping the BLEU score at its maximum. In Example 2, the word 'also' does not appear in the input sentence, then the proposed model could not predict this word according to the expected reference sentence. Despite that, it produced a perfect meaning sentence, close to the generated sentence. In Example 3, we have a similar problem since the neural model predicted a pronoun that could not be found in the input sentence. Thus it was not able to resolve which pronoun would best fit the sentence. In Example 4, we can observe a mix of both situations described in Example 2 and Example 3, with the prediction skipping a pronoun and also skipping some information present in the reference sentence like 'capital Quito'.

4.1 Experiments in Brazilian Portuguese

We also examined the ability of the proposed method to split-and-rephrase sentences written in Brazilian Portuguese, reusing the neural model trained with the part-of-speech labels gained from the English sentences of the WikiSplit corpus. We loaded the compiled model and made some predictions using the Spacy POS tagger [Honnibal and Montani 2017] for Brazilian Portuguese.

[7] `https://github.com/pauloberlanga/split-and-rephrase/`

Example 1	
Input	He was first elected in 2005 and represents the British Columbia New Democratic Party .
Reference	He was first elected in 2005 . He represents the British Columbia New Democratic Party .
Predicted (100.0%)	He was first elected in 2005. He represents the British Columbia New Democratic Party.
Example 2	
Input	They enrolled at New York Central College , an interracial institution in Cortland , New York , and worked as cleaning servants to support themselves .
Reference	They enrolled at New York Central College , an interracial institution in Cortland , New York . They also worked as cleaning servants to support themselves .
Predicted (90.0%)	They enrolled at New York Central College, an interracial institution in Cortland, New York. They worked as cleaning servants to support themselves.
Example 3	
Input	GTS Technologies is a paint finishing and mechanical handling systems company , registered in Wolverhampton , England .
Reference	GTS Technologies is a paint finishing and mechanical handling systems company . It is registered in Wolverhampton , England .
Predicted (85.7%)	GTS Technologies is a paint finishing and mechanical handling systems company. PRON_1 is registered in Wolverhampton, England.
Example 4	
Input	Pichincha is an active volcano in the country of Ecuador and gives its name to the entire province .
Reference	Pichincha is an active volcano in the country of Ecuador directly beneath its capital Quito . It gives its name to the entire province .
Predicted (63.4%)	Pichincha is an active volcano in the country of Ecuador. PRON_1 gives its name to the entire province.

Table I. Examples predicted by the model using WikiSplit sentences. See the BLEU score under the 'Predicted' tag.

Table II shows some rephrased sentences in Portuguese. We did not apply the BLEU metric here since we have no correct references for these sentences. In Example 5, we see that the word 'e' from the input sentence was successfully replaced by the '.' (period) character in the prediction sentence. The period was also followed by an expected pronoun. We consider it an exciting result although the skipped pronoun, since the model intercepted the correct terms just like for the English predictions. In Example 6, we see the skipped pronoun with repetitive words, sampling one of our wrong predictions.

Example 5	
Input	João era o melhor aluno de matemática e tirava nota 10 em todas as provas .
Predicted	João era o melhor aluno de matemática. PRON_1 tirava nota 10 em todas as provas.
Example 6	
Input	A seleção brasileira de futebol representa uma das equipes mais gloriosas da história futebolística, já tendo vencido 5 vezes a Copa do Mundo Fifa e diversos outros torneios.
Predicted	A seleção brasileira de futebol representa uma das equipes mais gloriosas da história . . PRON_1 tendo vencido vezes a copa do mundo diversos outros outros torneios.

Table II. Examples of some Brazilian Portuguese sentences predicted by the model.

5. DISCUSSIONS

The results confirm that the proposed encoder-decoder neural model could split a complex sentence into shorter sentences, most of the time preserving the meaning of the original sentence successfully. More than that, it also showed the potential to simplify sentences written in Brazilian Portuguese. On the other hand, some of the predictions skipped familiar words by the reference sentence and brought the grammatical classes instead. This is justified by the fact that our model does not treat lexical issues. We consider this as a minor problem but an essential question to solve.

As the model generates each word of the prediction sentence considering a soft-search on a set of references and all the previously created words, it showed the capability to align and learn this split-and-rephrase behavior simultaneously.

Regarding the BLEU score, the skipped words and some mistaken repetitive words reflected a low score for some predictions. Additionally, a certain number of the reference sentences in the WikiSplit corpus contain new information that does not effectively simplify the text, as observed in the reference sentence from the Example 4, Table I. We view that that extra information also contributed to harm the BLEU score.

6. CONCLUSIONS

We proposed a novel encoder-decoder neural model for sentence simplification through the use of the split-and-rephrase method. The model relies on the recurrence of the part-of-speech tags to automatically learn this split behavior. We also showed that the model trained against an English dataset, can split and rephrase sentences in Brazilian Portuguese with exciting results, performing a simple form of transfer learning.

One of the challenges left for the future is constructing a lexical approach to properly place skipped words in the predicted sentences, like the skipped pronouns according to their referenced constituents. Another exciting experience would be to promote the model evaluation under a proper Brazilian Portuguese corpus and also by human judgment assessments.

Acknowledgements

This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), under process number 2018/03129-8.

REFERENCES

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., Caseli, H. M., and Fortes, R. P. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*. ACM, New York, NY, United States, pp. 15–22, 2008.

Alva-Manchego, F., Martin, L., Scarton, C., and Specia, L. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, pp. 49–54, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2014.

Botha, J. A., Faruqui, M., Alex, J., Baldridge, J., and Das, D. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pp. 732–737, 2018.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014.

Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese. In *Computational Processing of the Portuguese Language*, A. Villavicencio, V. Moreira, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H. Gonçalo Oliveira, and G. H. Paetzold (Eds.). Springer International Publishing, Cham, pp. 272–283, 2018.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *In Proceedings of the International Conference on Learning Representations (ICLR)*. Curran Associates, Inc., San Diego, CA, USA., 2015.

Munkhdalai, T. and Yu, H. Neural semantic encoders. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Vol. 1. Association for Computational Linguistics, Valencia, Spain, pp. 397, 2017.

Narayan, S., Gardent, C., Cohen, S. B., and Shimorina, A. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pp. 606–616, 2017.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pp. 85–91, 2017.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318, 2002.

Scarton, C., Oliveira, M., Candido Jr., A., Gasperin, C., and Aluísio, S. SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In *Proceedings of the NAACL HLT 2010 Demonstration Session*. Association for Computational Linguistics, Los Angeles, California, pp. 41–44, 2010.

Scarton, C., Paetzold, G. H., and Specia, L. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Elsevier Inc., Miyazaki, Japan, pp. 3504–3510, 2019.

Shardlow, M. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications* 4 (1): 58–70, 2014.

Siddharthan, A. A survey of research on text simplification. *ITL–International Journal of Applied Linguistics* 165 (2): 259–298, 2014.

Štajner, S., Calixto, I., and Saggion, H. Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. "INCOMA Ltd. Shoumen, BULGARIA", Hissar, Bulgaria, pp. 618–626, 2015.

Tajner, S. and Glava, G. Leveraging event-based semantics for automated text simplification. *Expert Systems with Applications* vol. 82, pp. 383 – 395, 2017.

Vu, T., Hu, B., Munkhdalai, T., and Yu, H. Sentence Simplification with Memory-Augmented Neural Networks, 2018.

Wang, T., Chen, P., Amaral, K., and Qiang, J. An experimental study of LSTM encoder-decoder model for text simplification, 2016.