Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5588–5594
Marseille, 11–16 May 2020

CombiNMT: An Exploration into Neural Text Simplification Models

Michael Cooper, Matthew Shardlow
Department of Computing and Mathematics
Manchester Metropolitan University
mikejcooper90@gmail.com, m.shardlow@mmu.ac.uk

Abstract

This work presents a replication study of *Exploring Neural Text Simplification Models* (Nisioi et al., 2017). We were able to successfully replicate and extend the methods presented in the original paper. Alongside the replication results, we present our improvements dubbed CombiNMT. By using an updated implementation of OpenNMT, and incorporating the Newsela corpus alongside the original Wikipedia dataset (Hwang et al., 2016), as well as refining both datasets to select high quality training examples. Our work present two new systems, CombiNMT995, which is a result of matched sentences with a cosine similarity of 0.995 or less, and CombiNMT98, which, similarly, runs on a cosine similarity of 0.98 or less.

By extending the human evaluation presented within the original paper, increasing both the number of annotators and the number of sentences annotated, with the intention of increasing the quality of the results, CombiNMT998 shows significant improvement over any of the Neural Text Simplification (NTS) systems from the original paper in terms of both the number of changes and the percentage of correct changes made.

Keywords: Text Simplification, Replication, Neural Machine Translation

1. Introduction

Neural Machine Translation (NMT) has brought significant improvements to the field of Natural Language Processing (NLP) especially when results from these systems are compared to those from Rule Based Statistical Machine Translation (Weiss, 1995), Statistical Machine Translation Systems (Hwang et al., 2015) and Neural Phrase Based Machine Translation (Wu et al., 2016). NMT has become a widely used technique in Machine Translation (MT), as well as a well-regarded approach for other tasks, including dialogue generation, parsing and summarization.

This paper is a replication study of the Neural Text Simplification (NTS) system by Nisioi et al. (2017) as part of the REPROLANG shared task at LREC 2020. As such, we employed the same methods and configurations as the original paper. We have built on this replication by employing an extended version of the human evaluation described in the original paper, as well as employing more recent technologies to improve on the results from the original paper.

In addition to the technologies used in the original paper, we refined the original corpus, using RoBERTa's large model (Liu et al., 2019), finetuned on MNLI [1], to run the Semantic Text Similarity function as well as introducing further parallel data from Newsela derived in the same way. We also removed the sentences which were the most similar, ensuring the system preferred more bold rewrites, learnt from the differences found in the sentences.

2. Background

2.1. Neural Text Simplification

This paper is a replication study of Nisioi et al. (2017), in which the research team presented an Automated Text Simplification (ATS) system to address the applicability of Neural sequence to sequence models. ATS systems are designed to change original texts into simpler variants which would be understood by a wider audience and more easily processed by various NLP tools. By making use of advances in NMT, the researchers adapted existing architectures for their task.

The resulting system was named the Neural Text Simplification (NTS) model which used the OpenNMT framework to train and build an architecture with two LSTM layers. They had an RNN Encoder-Decoder pair, connected by an attention mechanism layer, the RNNs had hidden states of size 500 and 500 hidden units.

In an attempt to reduce the likelihood of the system overfitting, NTS has a dropout probability rate of 0.3. The researchers trained the model for 15 epochs over the data, with plain Stochastic Gradient Descent (SGD) optimization and the vocabulary size set to 50,000. After the 8th pass over the data, the learning rate of the system is halved. The learning rate is a configurable hyper-parameter, which dictates the amount of change to a model during the discovery of the 'weights' of the neural network. The learning rate has a small value, usually between 0.0 and 0.1. In the original paper, the parameter is set over 'uniform distribution with support [-0.1, 0.1]' meaning each outcome is initially equally likely.

On top of this architecture, the researchers employed global attention in combination with input feeding for the decoder. Input feeding in this case is the approach of concatenating the representation of the previous output with the context vector of the next input, forcing the model to keep track of important encoder-decoder alignment decisions. The researchers refer to this model as NTS.

Another model, which is referred to as NTS-w2v came about because the researchers were interested in whether 'large scale pretrained embeddings' improved text simplification models. This was constructed using pretrained word2vec embeddings from Google News Corpus concate-

[1] available from `https://github.com/pytorch/fairseq/blob/master/examples/roberta`

nated with locally trained embeddings using word2vec with hierarchical SoftMax and a window of 10 words. There were two sets of embeddings used in this model; one for the encoder, which used 'word2vec trained on the original English texts combined with Google News' whilst the decoder was trained using 'word2vec trained on the simplified version of the training data combined with Google News.' When concatenated, these embeddings create representations of size 500, as stated at the start of the description of the NTS model. If there was a word missing from the embeddings, it was replaced 'with a sample from a Gaussian distribution with mean 0 and standard deviation of 0.9.' All other parameters are unchanged from the NTS model.

To ensure best predictions and therefore the best simplified sentence, the researchers used the inbuilt beam search to find the output with the highest likelihood from the set of probabilities over potential output sequences.

For the first evaluation, the output from each system had the total number of changes counted, which included counting a change of an entire phrase as one change. If the change preserved the original meaning and grammaticality whilst making the sentence easier to understand, they are marked as 'correct.' If two annotators did not agree, the contentious sentence was given to a third annotator to obtain the majority vote. The second saw three native English speakers rate the grammaticality and meaning preservation of each sentence with at least one change on a 1-5 Likert scale. Third, three non-native English speakers were asked how easy to understand the simplified sentence was in comparison to the original sentence.

2.2. Related Work

Text simplification has existed as an open problem in NLP for the past 20 or so years, starting out trying to address issues of technical manual writing (Hoard et al., 1992) and assisting stroke survivors to read (Carroll et al., 1998). Text simplification is typically considered in two strands, lexical simplification and syntactic simplification.

In lexical simplification a pipeline approach (Shardlow, 2014b) is typically adopted, consisting of complex word identification (Yimam et al., 2018), substitution generation, word sense disambiguation and re-ranking to select a replacement synonym. Lexical simplification is a difficult task and sometimes errors may cause the resulting text to be more difficult than the original text (Shardlow, 2014a).

Syntactic simplification, on the other hand, focuses on rewriting the grammatical structure of a text to transform difficult to understand constructs such as the passive voice, or long lists into more understandable structures (Siddharthan, 2014). Syntactic simplification is typically rule-based (Siddharthan, 2006), although rules to transform tree structures may be learnt from corpora.

Uniting these two approaches, machine translation software can be used to identify lexical and syntactic simplifications at the same time. Early attempts used Phrase=based statistical machine translation software (Wubben et al., 2012), whereas newer efforts have used neural machine translation (Nisioi et al., 2017). In this context of translation, the source language is complex English and the target language is simple English.

Since the advent of Neural Text Simplification, there has been a growing interest in this area from the wider NLP community. Advances have focused on adapting new types of generation networks for simplification such as Pointer Generator (Li et al., 2018) and Neural Semantic Encoders (Vu et al., 2018). Other strands of research have, in parallel, sought to control the difficulty level of the output of neural text simplification systems (Nishihara et al., 2019; Marchisio et al., 2019; Agrawal and Carpuat, 2019).

3. Data

In this study we use the dataset from the original study, an edited version of the publicly available Hwang et al. (2015) dataset comprising alignments between standard English Wikipedia and Simple English Wikipedia, as well as the Newsela Corpus, comprising of 1.9k English and simplified English news articles, which has been shown to be useful for simplification. We used Standard English articles and those articles graded at simplicity 3 on a 1-5 scale. This was the last grading level which included all 1.9k articles in the standard English set.

3.1. EW-SEW

The dataset used is the publicly available dataset released by Hwang et al. in 2015, based on manual and automatic alignments between English Wikipedia and Simple English Wikipedia. Only the matches above 0.45 similarity threshold were used, which came to 284K sentences (around 150K fully matched sentences and 130K partial matches.) The NTS authors also used the SARI dataset (Xu et al., 2016) containing 2000 sentences for tuning and 359 for testing. The sentences used for tuning included 8 different simplified versions of the same original sentence. The first 70 sentences for testing were subjected to three different types of human evaluation. The edited version is publicly available on the original study's GitHub release [2].

This dataset was chosen by the original researchers because it was one of the largest publicly available datasets at the time which allowed the system to learn how to shorten sentences due to the fact it had full and partial matched sentences in them. We also adopted this dataset to ensure proper replication of the original study.

3.2. Newsela

We also evaluated the Newsela corpus, which includes around 1.9K standard English news articles and then simplified version of these articles. We used standard English and then those articles graded at 3 on a 1-5 scale. This level was chosen because it was the last grading which included all the same articles as the standard English set. We chose to use the Newsela corpus due to the fact it had gradients of simplification. In theory, this means the system can be trained to simplify the text to different levels, depending on the comprehension level of the reader. The different reading complexity are professionally levelled to ensure that the complexity is standardised across the individual 1-5 levels. In practice, we found that at the level of simplification we

[2] https://github.com/senisioi/NeuralTextSimplification/

employed, the number of matching articles and matched sentences was significantly reduced, which made it untenable to simplify to different levels. We were able to show however that incorporating these matches from Newsela led to an improvement in the simplification system.

3.3. Data Quality

We ensured that only files, from the Newsela corpus, with matching titles were evaluated. These files were read in line by line to find only parallel sentences, which were analysed to make sure they were not identical. The leftover sentences were then run through a Semantic Text Similarity function freely available on the RoBERTa large model, finetuned on MNLI. This left us with 4 files. These 4 files contained sentences which were identical matches, those which were simplified, those which were contradictions of each other, and those where one tailed on from the other. This meant that only non-identical, parallel, simplified sentences were used in the combined dataset. We ran the Hwang et al. (2015) dataset through the same evaluation process to ensure that the resulting corpus was of the highest possible quality, although the size was significantly reduced. We ranked the sentence pairs according to cosine similarity. Sentences which have a higher cosine similarity will, in theory, not teach a system much about text simplification, there could be a cut off on the lower end of the scale too, where the sentence are too dissimilar that overfitting of the system becomes a real possibility. The resulting combined dataset contained 124k high quality sentences in comparison to the 180k sentences in the EW-SEW dataset. The dataset used to train CombiNMT995 used 6.5k sentences from Newsela, and 102k from EW-SEW.

4. Replication Notes

Re-implementing the system proved to be rather straight forward. The complications encountered mainly arose from incompatible versions of software used. This re-implementation used the Python implementation of OpenNMT rather than the torch version. Since the original paper was released, OpenNMT have changed 'epochs' to training steps. As (Nisioi et al., 2017) saved the model after each epoch, and reduced the learning rate at the end of epoch 8, it was necessary to calculate the length of each epoch (total length of dataset / batch size) to properly replicate the results.

The decision to train models finetuned to improve SARI and BLEU scores seemed odd, due to the way which the two metrics lean favourably to different effects from the output system. There is a correlation between the number of changes, and the SARI score. The BLEU score usually correlates with a high score on the human evaluations of Grammaticality and Meaning Preservation. In our replication we decided not to fine tune our system to these metrics and we compare our results solely on the pre-tuned version of NTS.

The quality of the data included in the github repository [3], and the general quality of the documentation [4] were both incredibly helpful in the re-implementation, and in the experiments which created the CombiNMT systems.

5. Results

In this section, we present the results from both human evaluation and automated evaluation of our systems (table 1). The output from the default NTS system presented by (Nisioi et al., 2017) were also annotated, to see how the team of annotators compares to those used by the original team. This section also shows a few examples from each of our CombiNMT systems.

In contrast to the original system, we only used English-speaking voluntary annotators, who hold at least a Bachelor's degree, to ensure a good level of written English comprehension. We presented our annotators with the first 120 sentences from the (Xu et al., 2016) test set[5], as described earlier in the paper, alongside the output from each system running the same sentences.

CombiNMT995 was trained using the configuration described in (Nisioi et al., 2017) and earlier in this paper, with the combined dataset after subtracting any sentence pair with a cosine similarity of less than 0.995, with the development set used in (Nisioi et al., 2017). CombiNMT98 was trained in a similar manner, except with a cosine similarity value of less than 0.98.

We used the described configuration for (Nisioi et al., 2017)'s Neural Text Simplification (NTS) and Neural Text Simplification with word2vec embeddings (NTSw2v) replication systems. These systems were trained using the same dataset as in the original study.

5.1. Human Evaluation

For the measure of correctness, we presented the two sets of sentences to 2 annotators, asking them to count up the number of changes made, and then marking those sentences which successfully kept their grammatically, whilst preserving the meaning of the original sentence and creating a simpler to understand output. In the case that these annotators disagreed, there was a third on hand who was presented with only the contested sentences to provide a majority vote.

For the measures of both Grammaticality and Meaning Preservation, we presented the two sets of sentences to a total of 5 annotators, asking them to mark on a scale from 1–10 where 1 is poor grammar/ poor meaning preservation and 10 is perfect grammar / very good meaning preservation. We then calculated the mean value of the results from the annotations and halved it to match the original study's 1–5 ranking system.

When compared to the results of the output from the NTS system presented to our annotators, our CombiNMT995 system performed a significantly higher percentage of correct changes. The system also performed well on both the Grammaticality and Meaning Preservation measure. CombiNMT995 was outperformed by the replicated versions of the NTS version on the Grammaticality scores. In comparison to the SMT and Lexical Simplification systems used as

[3] https://github.com/senisioi/NeuralTextSimplification

[4] http://opennmt.net/OpenNMT-py/

[5] available as part of https://github.com/senisioi/NeuralTextSimplification/

Approach	Changes		Score		Evaluation Metric	
	Total	%	G	M	SARI	BLEU
CombiNMT995	**120**	**70.41**	4.08	**3.54**	**33.1**	76.02
CombiNMT98	83	4.21	3.54	2.74	30.81	77.04
NTS default (beam 5, hypothesis 1)	58	48.28	3.98	3.39	30.65	84.5
NTS replication	55	36.37	3.45	3.15	29.13	**87.46**
NTSw2v replication	88	50	**4.12**	3.37	30.28	80.75

Table 1: Scores for replicated systems without BLEU and SARI fine-tuning, the original default NTS system and CombiNMT systems

the benchmark in (Nisioi et al., 2017), our CombiNMT995 system performed almost double the percentage of correct changes that the best scoring model [6]. The system performed comparatively with the SMT systems on both grammaticality and meaning preservation and outperformed the LS system in terms of meaning preservation, whilst not performing as well as in terms of grammaticality.

5.2. Automated Evaluation

Alongside the human evaluation, the original study also presented the SARI and BLEU scores of the outputs from the systems. In the publicly available release of their study, (Nisioi et al., 2017) included an evaluation file, which calculated the SARI and BLEU scores. For the sake of consistency, we have used the same file to calculate the scores presented for our systems.

As can be seen in table 1, CombiNMT995 outperforms the other systems on the SARI score. CombiNMT995, however, is the worst performing system according to the BLEU scores. When comparing these scores to the systems in the original study, CombiNMT995 performs comparatively to the NTS and NTSw2v systems on the SARI scores, and comparatively to the SMT systems on the BLEU scores.

The replications which we produced performed less well than the original model on SARI score, however the reproduction of the default out performed it when comparing BLEU scores.

5.3. Example of Outputs

As can be seen in the well performing examples of CombiNMT995, the system performs both simplifications and reductions. The poorer scoring examples of that system also perform reductions; however they do not maintain the meaning of the original sentence. The well performing examples of CombiNMT98 do not perform so well. Although it still performs reductions, they are not performed quite so well. These reductions cut sentences off part way, so the meaning is completely lost, not dissimilar to the reductions performed by the poorer scoring examples from CombiNMT995. Where the sentence is not reduced, words are replaced which are not correct simplifications from the original sentence. The poorer scoring examples from CombiNMT98 seems to replace words at random, with no meaning retention, no care for grammaticality, nor correctness.

[6] tables available in the appendix

6. Error Analysis

To complete the error analysis, we have adopted the framework presented in (Shardlow and Nawaz, 2019). This framework uses the following 6 error categories, The results of which can be seen in Table 4.:

Type 1: A change has been made with no loss or alteration to meaning.

Type 2: No change has been made.

Type 3: A significant reduction in the information has been made, which has led to critical information being missed.

Type 4: A single lexical substitution has been made, which led to loss or alteration of the original meaning.

Type 5: An incorrect paraphrase or rewording of the sentence has been made, which led to loss or alteration of the original meaning.

Type 6: A single word from the original text is repeated multiple times in the resulting text.

These are the results from the same 120 sentences which were submitted to the human and automated evaluation. As can be seen in the table, 34.2% of the output was made up of sentences which had been successfully changed with no impact on the meaning. The next most frequent category is 'no change' at 20.8%. Many sentences will already be sufficiently simple and not require any form of simplification. This means that 55% of the sentences were successfully processed, with either a simplification, or no change being made. However, the remaining 45% of the sentences did represent errors that the system had made.

In 18.3% of cases, the sentences were vastly reduced, which affected the meaning of the sentence. More frequently than not, this system managed to maintain grammaticality, even when the meaning was affected. An error of this type may not negatively affect an end user, depending on the context of the sentence. For example, the following sentence:

> *They are castrated so that the animal may be more docile or may put on weight more quickly .*

Was simplified to give:

> *They are castrated .*

Original Sentence	CombiNMT98
SummerSlam (2009) is an upcoming professional wrestling pay-per-view event produced by World Wrestling Entertainment (WWE) , which will take place on August 23 , 2009 at Staples Center in Los Angeles , California .	SummerSlam (2009) is an upcoming professional wrestling pay-per-view event produced by World Wrestling Entertainment (2009 at Staples Center in Los Angeles , California .
This was demonstrated in the Miller-Urey experiment by Stanley L. Miller and Harold C. Urey in 1953 .	This was demonstrated in the Miller-Urey experiment by Stanley L. Miller and Harold C. Ga in 1953 .
Terms such as " undies " for underwear and " movie " for " moving picture " are oft-heard terms in English.	Aesthetics such as please undies Privy for underwear and Filli movie whenever for Selective moving picture " are oft-heard terms in English .

Table 2: CombiNMT98 - two well scoring outputs and one poorly scoring outputs

Original Sentence	CombiNMT995
On October 14 , 1960 , Presidential candidate John F. Kennedy proposed the concept of what became the Peace Corps on the steps of Michigan Union .	On October 14 , 1960 , President John F. Kennedy suggested what became the Peace Corps on the steps of Michigan Union .
They are culturally akin to the coastal peoples of Papua New Guinea .	They are culturally like the coastal peoples of Papua New Guinea .
Formal minor planet designations are number-name combinations overseen by the Minor Planet Center , a branch of the IAU .	It is a branch of the IAU .

Table 3: CombiNMT995 - two well scoring outputs, one poorly scoring output

Error Category	Total	%
1	41	34.1%
2	25	20.8%
3	22	18.3%
4	13	10.8%
5	19	15.8%
6	0	0%

Table 4: CombiNMT995 error category totals

Whilst the latter sentence is clearly still grammatical, and still retains a partial meaning from the original sentence, it is clearly missing some vital information from the original text. In this case, it may have been useful to split the former sentence into two sentences, where the sentence the system produced would be the first sentence and the second would be a simplified version of the remainder of the sentence.

Type 4 errors occurred 10.8% of the time. These errors replace or remove a single word, confusing the meaning of the sentence, or proving more difficult to read than the original. For example:

> *Oregano is an indispensable ingredient in Greek cuisine .*

Was simplified to give:

> ***The symbol** is an indispensable ingredient in Greek cuisine .*

Where it is clear that a key word has been negatively replaced, leading to the meaning of the output sentence becoming obfuscated to a reader.

Type 5 errors, similar to type 4 errors, but containing an extended phrase, occurred a further 15.8% of the time. Our system did not produce any Type 6 errors, indicating little overfitting was taking place.

7. Discussion

We matched the annotators used in the original study, for example, to ascertain correctness, we used 2 native English speaking participants, with a third on hand for a majority vote when needed. We increased the number of annotators for the grammaticality and meaning preservation from 3 to 4 to enable a higher majority. Rather than using a 1-5 Likert scale to measure the grammaticality and meaning preservation of the output, we used 1-10, with the intention of picking up greater nuances in the quality of the simplification.

It's easy to see that the CombiNMT995 system outperforms the CombiNMT98 system. The CombiNMT98 system showed classic signs of overfitting during the training phase. The accuracy prediction on the training data was much higher than on the development set. Although OpenNMT is designed to handle noise in the data, due to outliers such as dataset size/quality the system learns idiosyncrasies in the dataset and treats them as the true pattern.

The quality of the dataset, when any sentence pair with a cosine similarity of 0.98 or higher is disregarded, did not properly represent general patterns of simplification. As stated earlier, CombiNMT995 performs a higher percentage of 'correct' changes to the input sentences. We agree that this is the most important of the metrics, due to the fact it has been assessed in a real world setting by human evaluation rather than by an automated metric which undoubtedly favours one factor over another. It must be stated, however, that the overall quality of the system is found when combining the different metrics.

We sent a larger number of sentences to be annotated, the original study used the first 70 sentences from the (Xu et al., 2016) test set, where as we used the first 120 for each system. By looking at the human evaluations from the original study in comparison to our study, it can be seen that the annotators on this study marked the original outputs from the original NTS system more harshly than the original annotators. This could be due to the quality of the sentences annotated for this study which were not annotated for the original study.

NTSw2v performed comparatively with the CombiNMT995 system, which, considering the original NTSw2v had the most impressive results in the original paper, shows that a dataset of an increased quality can reduce the necessary size of dataset needed. Using the dataset at a cosine similarity below 0.995 gave the dataset the size it needed to reduce the chance of overfitting. It is difficult to propose the specific affect of Newsela over the system, and indeed to quantify any preposition put forth, however the filtered Wikipedia data alone would not have been of a size to create a system applicable to generalised text. The original study tuned their NTS model to find the best beam size and hypothesis number. This could have reduced the chance of overfitting on their models, due to the fact that beam search examines multiple alternatives in parallel and reduces the number of poor alternatives and reduce the size of its beam, whilst increasing the quality of the prediction.

We present an improvement on the NTS system (Nisioi et al., 2017) by using updated version of OpenNMT and incorporating new data from Newsela. In addition, we filtered the source sentences to remove very similar pairs, ensuring that the simplification system learnt true simplifications and not just minor edits. Our results show an improvement in human judgment scores, as well as SARI score over the original NTS Baseline.

8. Acknowledgments

Michael Cooper was funded via the EPSRC through the Project "Healtex: UK Healthcare Text Analytics Research Network" (EP/N027280/1) via a Feasibility Study: "Text Simplification for Clinicians — a Case Study in Cardiology"

9. Bibliography

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. 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)*, pages 1549–1564, Hong Kong, China, November. Association for Computational Linguistics.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98) Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Hoard, J. E., Wojcik, R., and Holzhauser, K., (1992). *An Automated Grammar and Style Checker for Writers of Simplified English*, pages 278–296. Springer Netherlands, Dordrecht.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217.

Li, T., Li, Y., Qiang, J., and Yuan, Y.-H. (2018). Text simplification with self-attention-based pointer-generator networks. In Long Cheng, et al., editors, *Neural Information Processing*, pages 537–545, Cham. Springer International Publishing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marchisio, K., Guo, J., Lai, C.-I., and Koehn, P. (2019). Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 193–203, Dublin, Ireland, 19–23 August. European Association for Machine Translation.

Nishihara, D., Kajiwara, T., and Arase, Y. (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, Florence, Italy, July. Association for Computational Linguistics.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada, July. Association for Computational Linguistics.

Shardlow, M. and Nawaz, R. (2019). Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the of the Associate for Computational Linguistics*, pages 380–389, 01.

Shardlow, M. (2014a). Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

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

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

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

Vu, T., Hu, B., Munkhdalai, T., and Yu, H. (2018). Sentence simplification with memory-augmented neural net-

works. 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 79–85, New Orleans, Louisiana, June. Association for Computational Linguistics.

Weiss, T. (1995). Translation in a borderless world. *Technical Communication Quarterly*, 4(4):407–425.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea, July. Association for Computational Linguistics.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.