Improving text simplification by corpus expansion with unsupervised learning

Akihiro Katsuta and Kazuhide Yamamoto
Nagaoka University of Technology
Nagaoka, Japan
{*katsuta, yamamoto*}*@jnlp.org*

***Abstract*—Automatic sentence simplification aims to reduce the complexity of vocabulary and expressions in a sentence while retaining its original meaning. We constructed a simplification model that does not require a parallel corpus using an unsupervised translation model. In order to learn simplification by unsupervised manner, we show that pseudo-corpus is constructed from the web corpus and that the corpus expansion contributes to output more simplified sentences. In addition, we confirm that it is possible to learn the operation of simplification by preparing large-scale pseudo data even if there is non-parallel corpus for simplification.**

***Keywords*-unsupervised machine translation; Japanese simplification; corpus expansion;**

I. Introduction

The number of foreigners in Japan has reached approximately 2.64 million and has been increasing[1]. Guide plates and official documents are often written in both Japanese and English to accommodate foreigners. However, foreigners in Japan are not always able to understand English. As such, they often experience inconveniences or are otherwise disadvantaged in their daily life. There are many more non-English speaking foreigners (56%) than foreigners who do not speak Japanese (37%), in Japan[6]. Simplification is one way of addressing the problem whereby foreigners are unable to effectively access the required information. In this investigation, a practical system is desired.

Automatic sentence simplification aims to reduce the complexity of vocabulary and expressions in a sentence while retaining its original meaning. Current approaches often consider the simplification process as a monolingual text generation task such as machine translation[18; 19; 15; 23; 21; 9]. The translation model learns simple rewriting operations from a parallel corpus consisting of complex sentences and simplified sentences. Neural Machine Translation (NMT) can more effectively exploit large parallel corpora, although Statistical Machine Translation (SMT) is still superior when the training corpus is not big enough. The only corpus available in Japanese is a Japanese Simplified Corpus with Core Vocabulary (hereinafter referred to as SNOW T15+T23)[2]. In this corpus, there are 85,000 sentences including a corpus of 50,000 sentences given by students (SNOW T15)[14] and a corpus of 35,000 sentences given to an anonymous person via crowdsourcing (SNOW T23)[8].

[1] https://www.e-stat.go.jp/
[2] http://www.jnlp.org/SNOW

We built a Japanese text simplification model trained by large corpus automatically expanded from web text using an unsupervised translation model. We first construct a large-scale non-parallel corpus of complex and simplified sentences collected automatically. Next, we learn simplification operation by unsupervised translation. In fact, in order to automatically collect simplified sentences, it is necessary to estimate the readability of sentences and classify them into simplified sentences or complex sentences. One of the most commonly used methods for estimating readability in English is the Flesch Reading Ease Formula[4]. However, in this corpus, simplification task is focused on the compression of the vocabulary, and the operation of manual rewriting to the predetermined 2000 words (core vocabulary) is performed. Therefore, it is easy to collect simplified sentences automatically given that fundamentally, the corpus only checks to examine the vocabulary composed of core vocabulary.

II. Related works

A. Unsupervised machine translation

Various neural models have been devised for unsupervised learning, presenting the possibility of machine translation techniques that do not require a parallel corpus. Artetxe et al. [3] and Lample et al. [11] have managed to train a standard attentional encoder-decoder NMT system from monolingual corpora alone. For that purpose, they use a shared encoder for both languages with pretrained cross-lingual embeddings, and train the entire system using a combination of denoising, back-translation and, in the case of Lample et al., adversarial training. This method was further improved by Yang et al. [20], who use a separate encoder for each language, sharing only a subset of their parameters, and incorporate two generative adversarial networks. However, Artetxe et al. [2] adapted the cross-lingual n-gram embeddings from monolingual corpora based on the mapping to train an unsupervised SMT model, obtaining large improvements over the original unsupervised NMT systems. It was argued that the modular architecture of phrase-based SMT was more suitable for this problem.

More recently, for additional improvement is attempted to combine both SMT and NMT to build hybrid unsupervised machine translation systems. This idea was already explored by Lample et al. [12], who aided the training of their unsupervised NMT system by combining standard back-translation with synthetic parallel data generated by unsupervised SMT.

This work builds off of existing work in the unsupervised machine translation, based on phrase-based SMT[2]. This work acquires phrase translation pairs by mapping pre-trained n-gram embedding between two languages to shared cross-lingual space and enables unsupervised learning. Unsupervised mapping by the method of Artetxe et al. [1] was used to map the n-gram embeddings to shared cross-lingual space. Further, they used iterative back-translation to train two unsupervised translation systems in both directions in parallel and to generate a synthetic source to construct a development set for tuning the parameters of their unsupervised statistical machine translation system. Iterative back-translation is a joint training algorithm to enhance the effect of monolingual source and target data by iteratively boosting the source-to-target and target-to-source translation models[22; 5].

B. For Text simplification

In the unsupervised paradigm, Paetzold and Specia [17] proposed an unsupervised lexical simplification technique that replaces complex words in the input with simpler synonyms, which are extracted and disambiguated using word embeddings. For another work as an approach for simplification that does not require a parallel corpus, Kajiwara and Komachi [7] proposed an unsupervised method in which a large-scale pseudo-parallel corpus is automatically constructed for text simplification based on the word similarity among monolingual corpora. They computed the sentence similarity for all pairings of normal and simple sentences using maximum alignment. Alignment was performed for only word pairs with a word similarity equal to or greater than 0.49 and only sentence pairs with a sentence similarity equal to or greater than 0.53 were aligned. As a result, 492,993 sentence pairs were obtained from 126,725 article pairs of English Wikipedia and Simple English Wikipedia. Given that this investigation is on text simplification in Japanese, the result of Maruyama and Yamamoto [13] is shown compared to that using NMT or SMT models, and the accuracy was comparable to SNOW T15 corpus.

III. Method for Unsupervised learning

This system was trained on monolingual corpus only and thus complex and simplified sentences were extracted from Nihongo Web Corpus 2010 (NWC 2010)[3] and a monolingual corpus was constructed for each. This corpus was created from approximately 100 million web pages collected from June to September 2010.

A. building Simplification corpus

For unsupervised translation learning, training corpora are used as the independent monolingual corpora by eliminating the alignment of the bilingual corpus. However, in the simplified corpus, this condition causes the task advantageous in an unintended way. In a simplified corpus, there are pairs of sentences where complex sentences and complex words are simply rewritten. Therefore, there is a high possibility that the same peripheral words of the operation target word are shared when a sentence is simplified using only a local substitution operation. This inevitably results in getting close to between correct substitute pairs in word embedding that has learned co-occurrence in the window.

The ultimate goal of simplification that does not require a parallelized simplification corpus is to learn simplification operations from non-parallel monolingual corpora. We show this advantage and investigate whether a large unclean corpus is more effective than a small clean corpus.

As briefly mentioned in section I, the aim is simplification in SNOW T15+T23 as lexical compression by substitution to the core vocabulary (the named entity and symbols are exceptionally permitted), so that simplified sentences can be easily collected from a monolingual corpus by examining whether the sentences contain anything other than the core vocabulary and named entity and symbols. We check the words each sentences and classify NWC 2010 to simplified and complex sentences and build a pseudo-corpus. It should be noted that when we use NWC 2010, we remove three or more consecutive equivalent tokens and symbols for cleaning text.

B. Unsupervised SMT

The method of Artetxe et al. was used to train an unsupervised simplification system from monolingual corpora using their open-source implementation monoses[4]. We trained an unsupervised SMT system from monolingual corpora using the default settings in the implementation.

1) Cross-lingual mapping: Complex sentences were prepared and simplified sentences were built for each n-gram embedding. The method applies a frequency-based vocabulary cut-off by learning the mapping over the 20,000 most frequently used words in each language. We kept this cut-off to learn the mapping over the most frequent 20,000 uni-grams and then applied the resulting mapping to the entire embedding space including

2) induce phrase table: The extracted phrase translation pairs for every n-gram in the simplified corpus could be taken as a potential translation candidate for each n-gram in the complex corpus. We limit the translation candidates for each complex phrase to its 100 nearest neighbors of simplified phrase.

To estimate their corresponding phrase translation probabilities, we applied the SoftMax function over the cosine similarities of their respective embeddings. More concretely, given the source language phrase $\bar{e}$ and the translation candidate $\bar{f}$, their direct phrase translation probability was computed as follows:

$$\phi(\bar{f}|\bar{e}) = \frac{\exp(\cos(\bar{e}, \bar{f})/\tau)}{\sum_{\bar{f}'} \exp(\cos(\bar{e}, \bar{f}')/\tau)} \quad (1)$$

In this formula, $\bar{f}'$ iterates across all target language embeddings and τ is a constant temperature parameter that controls the confidence of the predictions. For tuning, we

[3]https://www.s-yata.jp/corpus/nwc2010/

[4]https://github.com/artetxem/monoses

induce a dictionary over the cross-lingual embeddings with nearest neighbor retrieval and use maximum likelihood estimation. However, inducing the dictionary in the same direction as the probability predictions lead to a degenerated solution (SoftMax approximates the hard maximum underlying the nearest neighbor as τ approaches 0). Therefore, we induce the dictionary in the opposite direction at the same time and apply maximum likelihood estimation:

$$\min_{\tau} \sum_{\bar{f}} \log \phi(\bar{f}|NN_{\bar{e}}(\bar{f})) + \sum_{\bar{e}} \log \phi(\bar{e}|NN_{\bar{f}}(\bar{e})) \quad (2)$$

To compute the lexical weightings, we align each word in the target phrase with the one in the source phrase that most likely generating it, and take the product of their respective translation probabilities:

$$\mathrm{lex}(\bar{f}|\bar{e}) = \prod_{i} \max(\epsilon, \max_{j} w(\bar{f}_i|\bar{e}_i)) \quad (3)$$

The constant ϵ guarantees that each target language word also in out of vocabulary will yield a minimum probability mass, which is useful for modelling NULL alignments. In our experiments, we set ϵ = 0.001, which is the same experimental setting used by Artetxe et al.

3) Iterative back-translation: In general, standard SMT uses MERT over a small parallel corpus to tune the weights of the different scoring functions combined via its log-linear model. Given that we only have access to monolingual corpora in our scenario, we generate a synthetic parallel corpus through back-translation and apply MERT tuning iteratively, repeating the process in both directions. For this purpose, we reserved a random subset of 10,000 sentences from each monolingual corpora and ran the proposed algorithm over them for 10 iterations, which was sufficient for convergence. Thereafter, final tuning was performed on the data that was divided into training and validation data. Pseudo data was created by back-translation for each group and the phrase table was updated by applying MERT with validation. To accelerate our experiments, we use each monolingual corpus for training, in addition to the 10,000 separate sentences that were held out as a validation set for MERT tuning and performed a fixed number of 3 iterations of the aforementioned algorithm.

IV. Experiment

The systems are evaluated using BLEU scores computed by the "multi-bleu.perl" script included in Moses and SARI scores. In general, evaluation was performed to determine simplicity. In our experiment, tokenizer is MeCab[10] using dictionary of UniDic[5].

A. Comparison between supervised and unsupervised

In the previous work, Maruyama and Yamamoto [13] conducted experiments using a PB-SMT and a standard bi-LSTM seq2seq model without attention in only a T-15. In this experiment, a baseline was set up with similar experimental settings.

Table I
Detail Parameters of the machine translation

Parameter	Value
learning rate	0.25
clip-norm	0.1
dropout	0.3
max_tokens	4000
hidden size	512
Number of layers	2
Number of headers	8

First, we conducted to compare supervised manner and unsupervised manner. We calculated the accuracy of simplification using only SNOW T15+T23 by SMT and transformer as supervised learning and USMT as unsupervised learning. The difference between supervised one and unsupervised one is whether the sentences are aligned or not in simplification corpus. We extracted data from the corpus to 83000/1000/747 as train/valid/test. Given that some original sentences are originally simple, they included data whereby the original and simplified sentences match. The test data was dropped to omit such data. In our experiments, we use Moses[6] as SMT and transformer that were implemented in fairseq[16]. The transformer performs parameter tuning and as a result, the parameter is Tab.I.

B. Usefulness of pseudo-corpus

For unsupervised learning, we used two kinds of corpora. We experimented with the case of using the pseudo-corpus constructed from NWC 2010 and the case of combining it with the training data of SNOW T15+T23.

As a result of extracting each 50 million sentences from the NWC 2010 using the method described in section III-A, details of the corpus are shown in TableII. In order to train USMT in multiple data scales, we further reduced the scale of the extracted corpora from NWC 2010 by several stages to obtain several scales (5M, 500k, 50k) of the corpus. We used USMT to learn simplification with each scale. At the same time, we also try that with combining SNOW T15+T23.

V. Results and Discussion

A comparison of each system using SNOW T15+T23 to learn text simplification is shown in Table III. The baseline gives no rewriting to the input sentence and thus the output is as same as the input. It is difficult to exactly compare the BLEU score of Maruyama and Yamamoto [14] with our score because the smoothing of BLEU calculation method is not clear in their experiment. However, it is possible to compare the results of the supervised approach that we conducted in SARI score. Thus TableIV shows that the transformer is the model that can generate the simplified sentence most because it is the highest SARI and reduces the vocabulary.

[5]https://unidic.ninjal.ac.jp/

[6]http://www.statmt.org/moses/

Table II
DETAIL OF DATASETS FOR TEXT SIMPLIFICATION (IN SNOW T15+T23, IT SHOWS EACH TRAIN/VALID/TEST.)

datasets	#sentences	Vocabulary	#tokens per sent.
SNOW T15+T23 (complex)	83,000/1,000/747	20,339/2,034/2,015	10.87/11.01/11.30
SNOW T15+T23 (simplified)	83,000/1,000/747	5,392/1,464/1,492	12.04/12.13/12.95
NWC 2010 (complex)	50,000,000	1,227,318	16.59
NWC 2010 (simplified)	50,000,000	81,925	10.25

Table III
SIMPLIFICATION RESULT IN SNOW T15+T23

model	BLEU	SARI	Vocabulary
baseline	48.22	22.19	2015
NMT (Maruyama et al.[14])	79.4	58.5	-
SMT (ours)	**60.88**	69.97	1372
transformer (ours)	60.46	**72.44**	**1223**
USMT (ours)	54.15	47.11	1755

Table IV
IMPROVEMENT ACCURACY WHEN EXTENDING CORPUS USING PSEUDO-CORPUS

data-size of NWC2010	BLEU	SARI	Vocabulary
without SNOW T15+T23 (83,000)			
50M	**52.78**	55.57	1685
5M	51.94	**59.82**	**1547**
500k	49.37	43.03	1722
50k	42.98	37.10	1766
with SNOW T15+T23 (83,000)			
50M	52.32	50.80	1677
5M	53.30	**56.39**	**1616**
500k	54.06	51.10	1699
50k	53.68	52.37	1728
5k	**54.65**	50.44	1715

As compared with the baseline, the result of the unsupervised learning shows that the score is improved and correct rewriting can be performed even if alignment information is lost.

Table IV shows the simplification accuracy by changing the learning data size of USMT. As a result, regardless of whether SNOW T15+T23 is included, when unsupervised SMT is learned, it is evident that the simplification accuracy improves and the vocabulary scale is reduced depending on the scale of the training data. However, using SNOW T15+T23, BLEU shows a different tendency. Although the SARI score improves, there is almost no difference in the BLEU score depending on the scale of the training data. It is considered that the vocabulary is further compressed because the case for which the solution is robustly solved is sufficiently learned by SNOW T15+T23 only, and expressions that the test data cannot consider are learned from the NWC 2010. Therefore, it is seen that even if the alignment information is dropped, it is advantageous for the unsupervised approach that the original was a parallel corpus. However, this result also shows that in its absence, comparable results can be obtained by expanding the scale of training data.

In Table V, we compares the output examples of each model. (a) shows the change with the size of the data, and (b) shows an error case often seen when expanding with a pseudo-corpus. As shown in (a), it can be seen that if the data is small, relatively many incorrect rewriting (for example, the "洪水, flood" has been rewritten as "話し合い, discussion" in the case of 500k) or cases are not performed, whereas rewriting quality improves as data is increased. In (b), there is a case where the meaning of rewriting becomes opposite by combining pseudo data. When USMT learned only SNOW T15+T23, "がっかりした, disappointed" is rewritten to "気を落とした, discouraged", but when expanded SNOW T15+T23 with a pseudo-corpus, it is rewritten to "驚いた, surprised" and when learned only with a pseudo-corpus, it is rewritten to "感動した, impressed". The inability to distinguish antonyms is due that word embedding has a problem, and in order to eliminate such (b), it is necessary to devise measures to keep antonyms away when learning word embedding.

VI. CONCLUSION

We constructed each corpus of complex sentences and simplified sentences automatically as a pseudo-corpus from a nonparallel corpus and showed that expanding simplification corpus in USMT promotes more positive simplification. When USMT is trained using a data set obtained by dropping the alignment from the parallel corpus, the enhancement of BLEU by the expansion of the corpus is not seen. However, since the score of SARI has been improved and the vocabulary of the output sentence has been reduced, it has become possible to rewrite it into a more simplified sentence. In addition, in the case of learning only with a pseudo-corpus with a small scale, sufficient word embedding learning can not be performed, and BLEU becomes worse than the baseline. Since such a tendency is not seen when using a parallel corpus, we can see that the parallel corpus is still advantageous even if we learn from the parallel corpus that dropped the alignment information.

In the future, we would like to improve the system by incorporating model improvements, such as using a neural network, and better training schemes in order to address complex simplification operations.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grants-in-Aid for Challenging Research (Exploratory) Grant ID 17K18481.

REFERENCES

[1] M. Artetxe, G. Labaka, and E. Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

Table V
EXAMPLES OF WORD REPLACEMENT

(a)	
Original Sentence	彼らは大雨といえば洪水を連想した。 Speaking of heavy rain, they associate floods.
USMT(50M)	彼らは雨といえば災害をイメージした。 Speaking of rain, they imaged a disaster.
USMT(5M)	彼らはまた雨といえば自然災害をイメージした。Speaking of rain, they imaged a natural disaster.
USMT(500k)	彼らは札幌といえば話し合いを連想した。 Speaking of Sapporo, they associated a discussion.
USMT(50k)	彼らは大雨といえば洪水を発展した。 Speaking of heavy rain, they developed floods.
SMT	彼らはひどい雨といえば被害を思い出させた。Speaking of heavy rain, they reminded the damage.
transformer	彼らはすごい雨といえば水が災害を想像した。Speaking of great rain, they imagined a disaster of water.
(b)	
Original Sentence	彼はその成績にがっかりした。 He was disappointed at the grade.
USMT(50M)	彼はその結果に感動した。He was impressed at the result.
USMT(50M + SNOW T15+T23)	彼はその結果に驚いた。 He was surprised at the result.
USMT(SNOW T15+T23 only)	彼はその評価に気を落とした。He was discouraged by the evaluation.
SMT	彼はその結果に残念な気持ちになった。He felt disappointed with the result.
transformer	彼はその試験の結果を残念に思った。He was disappointed with the results of the exam.

pages 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P18-1073.

[2] M. Artetxe, G. Labaka, and E. Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1399.

[3] M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, April 2018.

[4] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[5] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-2703.

[6] K. IWATA. The preference for english in linguistic services : apos;japanese for living: Countrywide surveyapos; and hiroshima(lt;special issuegt;changing japanese society and language issues). *The Japanese Journal of Language in Society*, 13(1):81–94, 2010. doi: 10.19024/jajls.13.1_81.

[7] T. Kajiwara and M. Komachi. Text simplification without simplified corpora. In *The Journal of Natural Language Processing*, volume 25, pages 223–249, 2018. doi: 10.5715/jnlp.25.223.

[8] A. Katsuta and K. Yamamoto. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1072.

[9] R. Kriz, J. Sedoc, M. Apidianaki, C. Zheng, G. Kumar, E. Miltsakaki, and C. Callison-Burch. 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, Volume 1 (Long and Short Papers)*, pages 3137–3147, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1317.

[10] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3230.

[11] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*, 2018.

[12] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[13] T. Maruyama and K. Yamamoto. Sentence simplification with core vocabulary. In *2017 International Conference on Asian Language Processing (IALP)*, pages 363–366, Dec 2017. doi: 10.1109/IALP.2017.8300618.

[14] T. Maruyama and K. Yamamoto. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1185.

[15] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu. 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 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2014. URL https://www.aclweb.org/anthology/P17-2014.

[16] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[17] G. H. Paetzold and L. Specia. Unsupervised lexical simplification for non-native speakers. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[18] S. Wubben, A. van den Bosch, and E. Krahmer. 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 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P12-1107.

[19] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4: 401–415, 2016. doi: 10.1162/tacl_a_00107. URL https://www.aclweb.org/anthology/Q16-1029.

[20] Z. Yang, W. Chen, F. Wang, and B. Xu. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1005. URL https://www.aclweb.org/anthology/P18-1005.

[21] B. Zhang, D. Xiong, J. Su, Q. Lin, and H. Zhang. Simplifying neural machine translation with addition-subtraction twin-gated recurrent networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4273–4283, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1459.

[22] Z. Zhang, S. Liu, M. Li, M. T. Zhou, and E. Chen. Joint training for neural machine translation models with monolingual data. *ArXiv*, abs/1803.00353, 2018.

[23] S. Zhao, R. Meng, D. He, A. Saptono, and B. Parmanto. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1355.