English-Russian Data Augmentation for Neural Machine Translation

Nikita Teslenko Grygoryev n.grygoryev@pangeanic.com
Pangeanic BI Europa, S.L., Valencia, Spain
Mercedes García Martínez m.garcia@pangeanic.com
Pangeanic BI Europa, S.L., Valencia, Spain
Francisco Casacuberta Nolla fcn@prhlt.upv.es
Universitat Politècnica de València, Valencia, Spain
Amando Estela Pastor a.estela@pangeanic.com
Pangeanic BI Europa, S.L., Valencia, Spain
Manuel Herranz m.herranz@pangeanic.com
Pangeanic BI Europa, S.L., Valencia, Spain

Abstract

Data Augmentation (DA) refers to strategies for increasing the diversity of training examples without explicitly collecting new data manually. We have used neural networks and linguistic resources for the automatic generation of text in Russian. The system generates new texts using information from embeddings trained with a huge amount of data in neural language models. Data from the public domain have been used for experiments. The generation of these texts increases the corpus used to train models for NLP tasks, such as machine translation. Finally, an analysis of the results obtained evaluating the quality of generated texts has been carried out and those texts have been added to the training process of Neural Machine Translation (NMT) models. In order to evaluate the quality of the NMT models, firstly, these models have been compared performing a quantitative analysis by means of several standard automatic metrics used in machine translation, and measuring the time spent and the amount of text generated for a good use in the language industry. Secondly, NMT models have been compared through a qualitative analysis, where generated examples of translation have been exposed and compared with each other. Using our DA method, we achieve better results than a baseline model by fine tuning NMT systems with the newly generated datasets.

1 Introduction

The use of large, quality datasets to train neural network models for specific NLP tasks such as machine translation (MT), summarization, paraphrasing, text generation or dialogue systems is essential to achieve good quality results. Data augmentation (DA) has recently seen increased interest in Natural Language Processing (NLP) due to the lack of data in low-resource domains or new NLP tasks, and the popularity of large-scale neural networks that require large amounts of training data. Despite this recent upsurge, this area is still relatively underexplored. Perhaps this is due to the challenges posed by the discrete nature of language data which makes it challenging to make significant DA.

Although current progress in the areas of NLP and MT allows for the analysis, understanding, and automatic generation of increasingly accurate and fluid text, such amounts of data with

good quality are hard to find. Sometimes, they are too scarce for use during the training of a neural network model.

These techniques are often investigated in Computer Vision (Perez and Wang, 2017) and DA's adaptation for NLP seems secondary and comparatively underexplored, especially in MT task.

There are several works that have been done previously such as (Fadaee et al., 2017) which is a data augmentation approach that targets low-frequency words by generating new sentence pairs containing rare words in new, synthetically created contexts; and (Sánchez-Cartagena et al., 2021) which present a multi-task DA approach in which they generate new sentence pairs with transformations, such as reversing the order of the target sentence, which produce unfluent target sentences. During training, these augmented sentences are used as auxiliary tasks in a multi-task framework with the aim of providing new contexts where the target prefix is not informative enough to predict the next word.

In this work we introduce a new DA technique based on words substitution of a specific type (noun, adjective or adverb) in a sentence using a language model (LM) which generates a new word according to the context of the sentence. In addition, we check the new generated word, in order to maintain the main quality of the sentence.

In the rest of this paper, our DA approach is presented in Section 3, Section 4 shows the results of the experiments, and Section 5 outlines our conclusions and proposals for future work.

2 Background

We present how neural machine translation models can translate sentences and how neural language models can generate new words. Moreover, we explain some state-of-the-art data augmentation techniques for text.

2.1 Neural Machine Translation

NMT aims to estimate an unknown conditional distribution $P(\mathbf{y}|\mathbf{x})$ where $\mathbf{x}$ and $\mathbf{y}$ are random variables that represent the source (input) and target (output) sentences (Bahdanau et al., 2015).

We assume that the input sentence is $\mathbf{x} = (x_1, \ldots, x_S)$ and the output sentence is $\mathbf{y} = (y_1, \ldots, y_T)$, S corresponds to the total number of input words and T to the total number of output words. Using the chain rule, the conditional distribution could be described as Equation 1.

$$\hat{\mathbf{y}}_1^{\hat{T}} = \arg\max_{T, \mathbf{y}_1^T} \prod_{t=1}^{T} Pr_\theta(y_t | y_1^{t-1}, c(x_1^S)) \tag{1}$$

where y_t represents the current translated word, which is generated from the previous translated words y_1^{t-1} using a type of representation denoted by c function of the input sentence $\mathbf{x}_1^S$ and using the parameters of the model θ estimated from a training dataset D.

Training is performed on a parallel corpus with stochastic gradient descent. For translation, a beam search with 5-10 size range is employed.

The Transformer (Vaswani et al., 2017) is the state-of-the-art architecture in NMT that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. It relies entirely on self-attention to compute representations of its input and output without using sequence-aligned Recurrent Neural Networks (RNNs).

2.2 Language models based on Transformer architecture

Nowadays, for most NLP tasks aimed at encoding text sequences, language models based on Transformer architecture are the state-of-the-art. In addition, these models can be specialized in a specific task by fine-tuning the weights on a different task than the one they have been trained

for. For that, the models are trained using supervised labelled data obtaining the best results until now. This methodology, where a model is first pre-trained and then specialized in a specific task, is called transfer learning. One of the first language models to use Transformer architecture is *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) that is based on the encoder of the Transformer. Then, *Robustly optimized BERT approach (RoBERTa)* (Zhuang et al., 2021) was developed which deletes the *Next Sentence Prediction* task, which is one of the objective functions that is used for training BERT and modifies the second objective function defined as the *Masked Language Model* by applying a dynamic approach instead of a static one. In addition, *Generative Pre-trained Transformer* (GPT) with its latest release GPT-3 (Floridi and Chiriatti, 2020) (Brown et al., 2020) was developed, which is formed of multiple Transformer decoder layers. For instance, the development of XLNet (Yang et al., 2019) brought a new approach by combining auto-regression, such as GPT-2 does, and found an alternative way to introduce bidirectional context as BERT or other similar architectures.

2.3 Data Augmentation

DA encompasses methods of increasing the size of training data without the necessity of manually collecting more data. Most strategies either add slightly modified copies of existing data or create synthetic data, aiming for the augmented data to act as a regularizer and reduce over-fitting when training machine learning models. The most popular technique of DA in NLP is Back-translation (Sennrich et al., 2016) which is used to generate parallel synthetic data starting from a monolingual corpus. Another technique is the substitution of words using pre-trained language modeling (Kumar et al., 2020). Also, Easy Data Augmentation (EDA) (Wei and Zou, 2019), which consists of four simple techniques (synonym replacement, random insertion, random swap and random deletion) that produce small changes in the original text, has been demonstrated to be as useful as other more complicated methods of DA.

In this work we present a more precise approach of substitution of words by word type and using pre-trained language modeling and POS-tagging, where we care about the quality, coherence and variety of new generated sentences. This approach is hypothetically more precise and accurate than other DA techniques presented earlier in this paper such as EDA techniques or back-translation, typically used for the generation of synthetic texts. Therefore, we aim to create not only a bigger but also a more diverse and domain-orientated training dataset.

3 Automatic generation of a parallel English-Russian corpus

In this section, we present our approach of generating a synthetic corpus in Russian using a pre-trained language model specialized in Russian called ruRoBERTa-large. This model performed better than the multilingual BERT language model in preliminary experiments.

The United Nations dataset (UN) (Ziemski et al., 2016) dataset for English to Russian version 1 has been used. We apply a cleaning process which consists in filtering out sentences that

- are shorter than 5 words because they are no relevant.
- are longer than 50 words in order to fit the maximum capacity of tokens in ruRoberta encoder.
- have Latin characters in the Russian side.
- have more punctuation marks than characters.
- have more than 10 words in the absolute difference comparing source and target sentence length.

- have more numbers than letters, both in Cyrillic and Latin alphabets.

After this cleaning process, we save 1 million sentences for DA application, 7.9 million sentences for baseline model training, 2000 sentences for validation and 2000 sentences for test.

We use a Part-Of-Speech tagger in order to select the type of word that is generated by the pre-trained LM. For that purpose, we have used a model called $ru\_core\_news\_lg$ trained with the Spacy (Honnibal and Montani, 2017) toolkit. This model determines if the type of the selected word (noun, adjective or adverb) to be changed is the same as the generated one, which guaranties a similar level of quality from the original sentence. In addition, we have used the Python library *pymorphy2*, which provides us the gender, person and number of the selected word. It also provides a lot of useful morphological information about the words which may be used for another data augmentation methods, like verbs substitution where the grammatical time is relevant.

During the empirical experimentation, we observed that the language model returns a score of the pertinence of generated word to the bidirectional context of the sentence. With this in mind, we established an acceptance threshold for each generated word, which is set to 0.07. This also reduces significantly the selection of useless and not significant words like one character words or punctuation marks. Thus, we can make a more precise and accurate word substitution by the generation of the new sentences in Russian. In the last step, we generate by high quality Pangeanic MT model the English part of the new sentences to obtain parallel corpus. Finally, we have created four completely new datasets which have been generated by four different methods: (1) substitution of adverbs, (2) substitution of nouns, (3) substitution of adjectives and (4) a mixture of the three methods where the selection of the methods is done equitably.

For basic methods (noun, adverbs and adjectives substitution) we select all the words of the selected type from the sentence using the POS-tagger and save the position of that words in the sentence, next we iterate over all the saved words, so in each iteration we took one word, save its gender and number using the *pymorphy2* library and change that word for the <MASK> token so the ruRoberta-large model can predict a new word which is evaluated by checking that the generated word

- is not the same word than original one.
- is not a useless word such as punctuation marks or prepositions.
- is the same in terms of gender and number than original one.

So when we haven't more words to be changed from the current sentence, we firstly check if at least one of the selected words have changed and if not we discard the sentence. Then, we select the next sentence to be augmented and repeat the process until there are no sentences left.

For the mix method we add an additional step before the selection of all the words of a specific type in a sentence. This step consists in selecting which method will be applied for the following sentence. This previous selection is done in a way that the final corpus has similar amount of new texts of each DA method used.

For all the methods, we stop when there is not more original sentences to be augmented.

Table 1 presents the number of new number of words, sentences and the mean of new words per sentence per method showing the number of computing days.

As we can see in Table 1, substitution of nouns is the method that produces the largest amount of words, sentences and gives the highest mean of words per sentence. This is due to Russian has a very rich vocabulary, especially when it comes to nouns and it is relatively easy to find an appropriate or accurate new noun that can fit in the context of the sentence.

Likewise, the adjectives and mix methods have also generated a significant amount of data. As nouns method substitution, adjectives are also a very heterogeneous and permissive when it

Method	New #words	New #sentences	New #words/sentence	Comp. days
Adverbs	275K	327K	1.2	5
Nouns	4.8M	922K	5.2	8
Adjectives	2.2M	800K	2.7	6
Mix	2M	504K	4	11

Table 1: Statistics of adverbs, nouns, adjectives and mixed substitution methods where the original number of sentences (1 million), generated new number of words, generated new number of sentences, mean of generated new words per sentence and the time used for each DA method using parallelized in 10 CPUs.

comes to substitution of another adjective that fits in the context of the sentence. As we can see, the mixed method combines the two most generative methods. This is because the mix method generates almost the same amount of data substituting the three types of words. The method that produces the least amount of data is adverbs because adverbs have less options of substitution. However, it took less time (5 days) than the rest of the methods.

Although the adjective method generated a similar amount of new sentences as nouns, its mean of words per sentences is the smallest of all four methods. Due to the number of adjectives in a sentence is significantly less than nouns.

By contrast, the mixed method generated fewer sentences than the nouns or adjectives methods, but has a significantly high mean of words per sentence. Although this method generates a dataset with high variety because of the combination of the type of words, it takes more time to compute due to the equality of the generation of type of words.

4 Experiments

Experiments have been performed in order to study how our DA method for a English-Russian dataset can improve NMT models.

The architecture of the machine translation models used is composed of 6 layers of encoder and 6 layers of decoder with 8 multihead attention units in both of them. This configuration is the same as the standard Transformer architecture.

On other hand, the maximum length of input and output sentences was established to 400 and batch size was set to 4096 tokens. For baseline model, we trained the model with 7.9 million samples of parallel data in 100000 training steps and for each model trained with extra data, we retrain the baseline model with an extra 50000 train steps using for each retrained model the data reflected in the Table 2. For each training process we have used the first 8000 steps for warm up, NOAM as a decay method and SGD as the optimizer method with $\epsilon = 0.05$. We have trained this model using the OpenNMT-py framework.

ONU dataset[1]		**# Samples**
Train	Original	7.9M
	Adverbs	271K
	Nouns	916K
	Adjectives	792K
	Mix	501K
Validation		2K
Test		2K

Table 2: Statistics of generated and original datasets.

As we can see in Table 2, after the cleaning process, we split the remaining of the original UN dataset for DA techniques, train, validation and test. For baseline models training set we picked 7.9 million samples. For DA techniques, we got 1 million samples which generated data used for the retraining of baseline models. Then, we save 2000 samples for validation set. Finally, we picked 2000 samples for testing set of the models.

4.1 Quantitative analysis

We have performed a quantitative analysis, for both Russian-English and English-Russian models, by comparing the corresponding baseline NMT model with the corresponding four methods trained with augmented data.

Therefore, six automatic evaluation metrics have been selected to automatically measure the NMT outputs: (1) *BiLingual Evaluation Understudy* (BLEU) (Papineni et al., 2002) which is the standard score used in machine translation evaluation, (2) *Translation Error Rate* (TER) (Snover et al., 2006) which looks for the total amount of edits needed to get the reference sentence from the hypothesis sentence, (3) ChrF-2 (Popović, 2015), which is the F-2 score but at char level, (4) NIST (Doddington, 2002), similar to BLEU but also calculates how informative a particular n-gram is, (5) *Better Evaluation as Ranking* (BEER) (Stanojević and Sima'an, 2014), which is a sentence level metric that can incorporate a large number of features combined in a linear model and (6) *Crosslingual Optimized Metric for Evaluation of Translation* (COMET) (Rei et al., 2020), which uses multilingual sentence embedding and the source sentence. The main goal of using a larger quantity of evaluation metrics than usual is to get a more precise information of the quality of the translations done by the NMT models.

Method	BLEU	TER	ChrF-2	NIST	BEER	COMET
Baseline	59.6 ± 1.4	31.4 ± 1.1	78.3	2.9	.73	.89
Adverbs	70.2 ± 1.4	21.4 ± 1.0	84.6	3.2	.79	.97
Nouns	68.8 ± 1.4	21.6 ± 1.0	83.9	3.2	.80	1
Adjectives	69.2 ± 1.4	22.1 ± 1.0	83.8	3.2	.79	.98
Mix	$\mathbf{70.9 \pm 1.3}$	$\mathbf{20.5 \pm 1.0}$	84.8	3.2	.80	1

Table 3: Results of the DA methods used (substitution of adverbs, nouns, adjectives and mix) applying a set of automatic evaluation metrics for machine translation models in English to Russian language direction. From left to right are BLEU, TER, ChrF-2, NIST, BEER and COMET. The value that follows the symbol $\pm$ is the confidence interval calculated using the bootstrap resampling technique.

The evaluation of models that were trained in the English-Russian direction is presented in Table 3. All DA substitution methods perform better compared to the NMT model in all the evaluation metrics. The mixed substitution method yields the best results of all the automatic evaluation metrics that have been calculated. As we can see in Table 3, we obtain a good scores when evaluating with COMET our models with augmented data where each one of them outperforms the baseline model.

This fulfills the hypothesis that a richer and more diverse dataset make translation models more accurate and which produces a more fluent translations.

However, as we can see in Table 4, best results were produced by the adjectives substitution method. Nevertheless, if we focus on the values of BLEU and TER and their confidence intervals, we can see that the values of the adjective and mixed substitution methods are similar. Therefore, we can deduce that mixed substitution method is also useful and produces good results. Furthermore, in both language directions (Russian to English and English to Russian) all the models that were retrained with augmented data perform better than the baseline model.

Method	BLEU	TER	ChrF-2	NIST	BEER	COMET
Baseline	40.9 ± 1.1	46.4 ± 1.1	62.3	2.6	.62	.28
Adverbs	47.1 ± 1.3	42.7 ± 1.1	71.9	2.7	.66	.25
Nouns	48 ± 1.2	43.1 ± 1.1	72.6	2.7	.67	.25
Adjectives	**48.3 ± 1.3**	**41.8 ± 1.1**	72.7	2.7	.67	.29
Mix	48 ± 1.3	42.5 ± 1.1	72.7	2.7	.68	.27

Table 4: Results of the DA methods used (substitution of adverbs, nouns, adjectives and mix) applying a set of automatic evaluation metrics for machine translation models in the Russian to English language direction. From left to right are BLEU, TER, ChrF-2, NIST, BEER and COMET. The value that follows the symbol ± is the confidence interval calculated using the bootstrap resampling technique.

Finally, as we can see in Table 4, the COMET values obtained using NMT models retrained with augmented data produce worse values but they are not statistically significant compared with the NMT baseline model. This is due to the fact that target (English) is synthetically generated using MT.

4.2 Qualitative analysis

We have randomly selected an example of the English to Russian models. The qualitative analysis showing the source sentence in English, the Russian translation reference and the machine translation sentences that have been generated by the NMT models are represented in Table 5. We can see the differences between the reference sentence and the different translations generated by the NMT models. The underlined words refer to the differences between each translation model. As we can see, the mixed method has more variability in the translation but keeps the overall meaning of the reference.

Method	Translations from English to Russian
Source	Their case also attracted the attention of the control commission in Geneva.
Reference	Их случай также привлек внимание комиссии по контролю в Женеве.
Baseline	Их дело также привлекло внимание комиссии по контролю в Женеве.
Adverbs	Их случай также привлек внимание контрольной комиссии в Женеве.
Nouns	Их дело также привлекло внимание комиссии по контролю в Женеве.
Adjectives	Их дело также привлекло внимание комиссии по контролю в Женеве.
Mix	Их дело также привлекло внимание контрольной комиссии в Женеве.

Table 5: Examples of translations obtained when translating using the models trained with the datasets generated using substitution of adverbs, nouns, adjectives and mixed methods where the source language is English and the target language is Russian.

Table 6 shows an example of a source sentence in Russian, a reference translation in English and the machine translation sentences in English that have been generated by the NMT models. We can see again that the mixed method has produced more variability in the translation. However, the adverb substitution method also has provided a rich and an accurate translation if we compare it with reference sentence. In general, we can see that the models retrained with augmented data have provided more fluent, accurate and clearer translations than the baseline model.

Method	Translations from Russian to English
Source	Их случай также привлек внимание комиссии по контролю в Женеве.
Reference	Their case also attracted the attention of the control commission in Geneva.
Baseline	**The case** also **brought** to the attention of the Geneva **Monitoring Commission**
Adverbs	The **other case** also **drew** the attention of the **control commission** in Geneva.
Nouns	The case also **attracted** the attention of the **control commission** in Geneva.
Adjectives	The case also **drew** the attention of the **control commission** in Geneva.
Mix	**Others also drew** the attention of the **control commission** in Geneva.

Table 6: Examples of translations obtained when translating using the models trained with the datasets generated using substitution of adverbs, nouns, adjectives and mixed methods where the source language is Russian and the target language is English.

5 Conclusions

In this work, we have automatically generated four new datasets using pre-trained neural language models in order to increase a Russian-English dataset for NMT systems. We have used DA methods by substituting nouns, adjectives and adverbs, or a mix of them observing the importance of selecting the correct type of generated word in order to generate a better-quality sentence.

The generated new datasets have been used for retraining the baseline models in both language directions of translation (from Russian to English and from English to Russian). In addition, the retrained NMT models have been compared performing a quantitative and qualitative analysis showing better results than the baseline models. In conclusion, it is worth noting that the presented DA methods are a viable way of improving NMT systems when there is not enough data or the quality of the data is low. However, there is a lot of work to do in this area in order to improve the method.

In terms of future work, we can also use the verb substitution method which will probably generate richer and broader datasets. However, this method seems to be more complex in terms of quality stability because the number, person and gender (Russian has three types of gender which makes it more complex: feminine, masculine and neutral) must correspond with the rest of the sentence. In addition, we propose the use of a statistic aligner which will significantly reduce the use of machine translation to create synthetic data by only translating a word instead of the full sentence.

References

Bahdanau, D., Cho, K., Montréal, U. D., Bengio, Y., and Montréal, U. D. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, page 138–145.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Floridi, L. and Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

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

Kumar, V., Choudhary, A., and Cho, E. (2020). Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11:1–8.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Sánchez-Cartagena, V. M., Esplà-Gomis, M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2021). Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Stanojević, M. and Sima'an, K. (2014). Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, volume 32.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*.