Textual Augmentation Techniques Applied to Low Resource Machine Translation: Case of Swahili

Gitau, Catherine
African Institute for Mathematical Sciences Ghana
cgitau@aimsammi.org

Marivate, Vukosi
Department of Computer Science, University of Pretoria, South Africa
vukosi.marivate@cs.up.ac.za

Abstract

In this work we investigate the impact of applying textual data augmentation tasks to low resource machine translation. There has been recent interest in investigating approaches for training systems for languages with limited resources and one popular approach is the use of data augmentation techniques. Data augmentation aims to increase the quantity of data that is available to train the system. In machine translation, majority of the language pairs around the world are considered low resource because they have little parallel data available and the quality of neural machine translation (NMT) systems depend a lot on the availability of sizable parallel corpora. We study and apply three simple data augmentation techniques popularly used in text classification tasks; synonym replacement, random insertion and contextual data augmentation and compare their performance with baseline neural machine translation for English-Swahili (En-Sw) datasets. We also present results in BLEU, ChrF and Meteor scores. Overall, the contextual data augmentation technique shows some improvements both in the $EN \rightarrow SW$ and $SW \rightarrow EN$ directions. We see that there is potential to use these methods in neural machine translation when more extensive experiments are done with diverse datasets.

Keywords: low-resource, data augmentation, machine translation

1 Introduction

There have been several advancements in machine translation and modern MT systems that can achieve near human-level translation performance on the language pairs that have significantly large parallel training resources. Unfortunately, neural machine translation systems perform very poorly on low-resource language pairs where parallel training data is scarce. Improving translation performance on low-resource language pairs could be very impactful considering that these languages are spoken by a large fraction of the world's population.

According to Lisanza (2021), only about 5-10 million people speak Swahili as their native language, but it is spoken as a second language by around 80 million people in Southeast Africa lingua franca, making it the most widely spoken language of sub-Saharan Africa.

Despite the fact that the language is spoken by millions across the African continent, it accounts for less than 0.1% of the internet whereas 58.4% of the Internet's content is in English according to W3Techs (2020), making it a low-resourced language. Even though Swahili is spoken by so many people, there is little extensive work that has been done to improve translation models built for the language. Data that is needed to produce high quality neural machine translation systems is unavailable resulting in poor translation quality.

In computer vision, data augmentation techniques are used widely to increase the robustness and improve the learning of the objects with very little training examples. In image processing, the trained data is augmented by, for example, horizontally flipping the images, random cropping, tilting etc. Data augmentation has now become a standard technique to train deep neural networks for image processing and it is not very common practice in training networks for natural language processing (NLP) tasks such as machine translation. Applying data augmentation techniques in text is not as straightforward as in computer vision because in computer vision, the label and content of the original im-

age is preserved as for natural language processing (NLP) tasks, there is need to retain the context of the sentence after augmentation. There are however several data augmentation methods that have been proven to improve performance on various NLP tasks such as text classification but it is not common practice to apply these data augmentation methods for tasks such as machine translation.

Neural machine translation (NMT) as presented in the work of Sutskever et al. (2014);Cho et al. (2014) is a sequence to sequence task that uses a bidirectional recurrent neural network known as an encoder to process a source sentence into vectors called the decoder which then predicts words in the target language. To be able to train a model that is able to produce good translations, these networks require a lot of parallel data or sentence pairs with words that are occurring in diverse contexts which is not available in low-resource language pairs therefore making the performance of the models quite low. One of the solutions to this problem is to manually annotate the available monolingual data which is time consuming and expensive or to perform unsupervised data augmentation techniques.

In this work we explore some data augmentation techniques that are widely used to improve text classification tasks and investigate their impact on the performance on low resource neural machine translation models for English-Swahili(En-Sw). The most popular text augmentation techniques applied to text classification tasks consist of four powerful operations; synonym replacement, random insertion, random swap and random deletion. These methods have been shown to improve performance in text classification tasks as shown in Wei & Zou (2019). And to better gauge the effect of our data augmentation methods, we compare the results with a baseline model trained on En-Sw datasets.

In summary, our contributions include:

1. We explore and evaluate on NMT task, three data augmentation techniques currently only being used in text classification tasks; Word2Vec based-augmentation which does synonym replacement in the sentences, TF-IDF-based augmentation to insert words in random positions in the sentence as well as use of Masked Language Model based-augmentation that does contextual data augmentation on the text.

2. We show how these data augmentation techniques can be used in NMT tasks.

3. We also extended the textaugment library Marivate & Sefara (2020) to use Fasttext's pre-trained models.

4. We present baseline NMT results in BLEU, Meteor and ChrF scores.

This paper is organised as follows; We look at the work that is been done by past authors on data augmentation for low-resource languages first. We also look at the data augmentation techniques and the approaches we used in our study which is described in Section 3. Section 4 describes the model settings that were considered for every data augmentation approach, Section 5 discusses the experimental results and then we conclude with stating limitations and future work as well as conclude at Sections 6 and 7.

2   Literature Review

Work on machine translation to improve machine translation quality on low resource languages is a widely studied problem. In natural language processing (NLP), data augmentation is a popular technique that is used to increase the size of the training data.

One promising approach is the use of transfer learning Zoph et al. (2016). This method proved that having prior knowledge in translation of a different higher-resource language pair can improve translating a low-resource language. A NMT model is first trained on a large parallel corpus to create the *parent model* and continued to train this model by feeding it with a considerable smaller parallel corpus of a low-resource language resulting into the *child model* which inherits the knowledge from the parent model by reusing its parameters. The parent

and child language pairs shared the same target language(English). The use of data from another language can be seen as a data augmentation task in itself and large improvements have been observed especially when the high-resource language is syntactically similar with the low-resource language Lin et al. (2019).

The work of Sennrich et al. (2016) explores a data augmentation method for machine translation known as back translation where machine translation is used to automatically translate target language monolingual data into source language data to create synthetic parallel data for training and is currently the most commonly used data augmentation technique in machine translation tasks. The quality of the backward system while effective, has been shown to negatively affect the performance of the final NMT Model when the target-side monolingual data is limited. Back translation as a method for performing data augmentation in machine translation could deteriorate the Low resource Language - English(LRL - ENG) translation performance due to the limited size of the training data as shown in Xia et al. (2019).

Xia et al. (2019) augment parallel data through two methods: back-translating from ENG to low resource language (LRL) or high resource language (HRL) and converting the HRL-ENG dataset to a pseudo LRL-ENG dataset. They use an induced bilingual dictionary to inject LRL words into the HRL then further modify these sentences using modified unsupervised machine translation framework. Their method proved to improve translation quality as compared to supervised back-translation baselines however, the method requires access to a HRL that is related to the LRL as well as monolingual LRL.

There are other data augmentation methods which have been used in other NLP tasks such as text classification to improve performance. Wei & Zou (2019) show that simple word replacement using knowledge bases like WordNet Miller (1995) can improve performance of classification tasks. Marivate & Sefara (2020) also observe that Word2Vec-based aug-

mentation is also a viable solution when one does not have access to a knowledge base of synonyms such as the WordNet Miller (1995). Kumar et al. (2020) show that seq2seq pre-trained models can be effectively used for data augmentation and these provide the best performance. These data augmentation methods are currently only being used to improve classification tasks and have not yet been utilized in any neural machine translation task to improve performance. In this work we will be looking at how some of these methods can be used to also improve neural machine translation models where the data is low-resourced. In particular, we will explore three data augmentation methods which include: 1) Word2Vec based-augmentation, 2) Tf-idf based augmentation, 3) Masked Language Model based-augmentation and use the additional data to train the NMT model.

3 Methodology

Our goal is to compare different data augmentation methods that are used in text classification tasks with the aim of identifying whether the methods can be used to improve the baseline NMT score. The results are compared across two different datasets and uses in-domain test sets to demonstrate the generalization capability of the models. These experiments are useful to help other researchers gain insights as they work on building better neural machine translation models for low-resourced languages. First, we describe the data that was used to train the models, then the data augmentation methods that we will be using and finally give details of the experiments we performed to test these methods together with the results obtained.

3.1 Training Data

Small amounts of parallel data are available for Swahili-English. The data was received from the work of Lakew et al. (2020) where they released standardized experimental data and test sets for five different languages(Swahili, Amharic, Tigirinya, Oromo and Somali). They collected all available parallel corpora for those five languages from the

Opus corpus Tiedemann (2012) which consists of a collection of translated texts from the web. For this work, we utilized data that includes JW300 Agić & Vulić (2019) and Tanzil Tiedemann (2012) which provides a collection of Quran translations to compare with the baseline results from the work of Lakew et al. (2020).

Table 1 shows the amount of parallel data that was collected. The data was split into train, dev and test sets as in Lakew et al. (2020). We then segmented the data into subword units using Byte Pair Encoding Sennrich et al. (2016) where we learned $20K$ byte pair encoding tokens.

3.2 Baseline

In this approach the Transformer NMT model is trained using Jw300 and Tanzil data combined then tested on different datasets from two different domains (Jw300 and Tanzil). The model is trained with no modifications throughout with standard preprocessing steps such as tokenization, lowercasing and cleaning. This model in this approach serves as a baseline for comparison.

3.3 Data Augmentation methods

We augmented the data using three types of augmentation methods: Word2Vec-based augmentation(synonym replacement), Tf-idf based augmentation(random insertion) and Masked Language Model(MLM)-based augmentation(context based augmentation). We combined the first two augmentation methods and used the Masked Language Model-based augmentation on its own. The Word2Vec and Tf-idf augmentations were done on the source language such that when training an En-Sw model, we augment the English language and when training a Sw-En model we augment the Swahili language. In Masked language modeling the augmentation was only done on the English language.

3.3.1 Word2vec-based augmentation

Word2vec is an augmentation technique mostly used in classification tasks that uses a word embedding model Mikolov et al. (2013) that is trained on publicly available datasets to find the most similar words for a given input word. We use Word Vectors pre-trained on Common Crawl and Wikipedia on both English and Swahili data using fastText Joulin et al. (2017) a library for text representation and classification. We load the pre-trained fastText model for each language into our algorithm to augment the texts by randomly selecting a word in the text to determine their similar words using cosine similarity as a relative weight to select a similar word that replaces the input word as done in Marivate & Sefara (2020). Our algorithm is as illustrated in Algorithm 1 . It receives a string which is the input data and augments the text into five different augmented texts then we use cosine similarity to select the best sentence that is at least 0.85 closer to the original text. The reason for this is that we'd like to retain the contextual meaning of a sentence even after augmentation. We compare the five different augmented sentences and pick the sentence that has a cosine similarity score that is highest. To prevent duplicated augmentations, we drop the sentences that are 100% similar to the original sentence. This augmentation was done on the source language where the corresponding target language sentences remained constant and unchanged. Examples of the augmented sentences can be seen in Table 2.

3.3.2 Tf-idf based augmentation

We created another set of augmented data that uses Tf-idf Ramos (2003). The concept of Tf-idf is that high frequency words may not be able to provide much information gain in the text. It means that rare words contribute more weights to the model. In this case, words that have low Tf-idf scores are said to be uninformative and thus can be replaced or inserted in text without affecting the ground truth labels of the sentence. Here, the words that are chosen to be inserted at a random position in the sentence are chosen by calculating the Tf-idf scores of

Language Pair	Split	Domain		
		JW300	Tanzil	Total
Sw-En	train	907842	87645	1024717
	dev	5179	3505	8684
	test	5315	3509	8824

Table 1: Data statistics showing number of examples/sentences available across four domains

Method	Sentence
English	
Original	The quick brown fox **jumps** past the lazy dog
Word2Vec + tfidf	The quick brown fox **leaps over retrorsum** the lazy dog
Swahili	
Original	Baba na mama yako ni wazuri sana
Word2Vec + tfidf	**Kizee** baba na mama yako ni **wema waar** sana

Table 2: Table showing example of augmented sentences

words over all the sentences and then taking the lowest ones. We therefore insert a new word at a random position according to the Tf-idf calculation. This was also done on the source language only and the corresponding target language sentences remain unchanged. Tf-idf was applied after performing the Word2Vec based augmentation method. This is illustrated in Algorithm 1.

3.3.3 Masked Language Model (MLM) augmentation

Since the above methods do not consider the context of the sentence, we decided to use Masked Language Modeling(MLM) where we used RoBERTa Liu et al. (2019) a transformer model that is pretrained on a large corpus of English data in a self-supervised fashion. It is used to predict masked words based on the context of the sentence. You can find the algorithm used in Algorithm 2. Taking the sentence, the model randomly masks 15% of the words in the input then runs the entire masked sentence through the model and predicts the masked words which helps the model to learn a bidirectional representation of the sentence. In this work, a sentence is passed through our algorithm which then predicts the masked word creating a new aug-

---

Algorithm 1 Word2Vec and Tf-idf based augmentation

Input: s: a sentence
Output: $\hat{s}$: augmented sentence
1: *Step 1: get similar words of each word in s :*
2: **procedure** AUGMENT(s)   Augmentation of sentence
3:      $t \leftarrow$ sentence s tokenized
4:      $u \leftarrow$ unique words from t
5:      **for** $w\ in(u)$ **do**
6:          $\vec{w} \leftarrow$ find five similar words for w
7:      **end for**
8: *Step 2: replace random words in s with similar words and insert Tf-idf word :*
9:      $n \leftarrow 5$
10:      **for** _ *in range*(n) **do**
11:          $w_i \leftarrow$ randomly select a word from s
12:          $w_0 \leftarrow$ randomly select one similar word for w_i from $\vec{w}$
13:          $\hat{s} \leftarrow$ replace w_i with similar word w_0
14:          $ss \leftarrow$ insert Tf-idf word in random position in s
15:          $\hat{s} \leftarrow$ merge $\hat{s}$ and ss
16:      **end for**
17:      **return** $\hat{s}$ Augmented sentence
18: **end procedure**

mented sentence. Note that this augmentation method was only done on the English language due to lack of enough resources to train a good MLM for the Swahili language.

Algorithm 2 Masked Language Model augmentation algorithm

Input: s: a sentence
Output: $\hat{s}$: augmented sentence
 1: **procedure** MLMAUGMENT(s)
   Augmentation of sentence
 2:   $n \leftarrow 15\%$ of words in the sentence
 3:   **for** _ *in range*(n) **do**
 4:    $w_i \leftarrow$ randomly select a word from s
 5:    $\hat{s} \leftarrow$ mask word w_i
 6:    $\hat{s} \leftarrow$ replace masked word with predicted mask
 7:   **end for**
 8:   **return** $\hat{s}$ Augmented sentence
 9: **end procedure**

For our experiments we combined the augmented sentences for Word2Vec based-augmentation and Tf-idf based data augmentation producing almost triple the original sentences. The MLM-based augmentation methods produced almost double the original parallel sentences. The total training data that was used is as shown in Table 2.

Method & Language Pair	Total
Word2Vec + tfidf (EN-SW)	2952864
Word2Vec + tfidf (SW-EN)	2774186
MLM (EN-SW)	2048613

Table 3: Data statistics showing total data used for training after augmentation

4 Experiments

This section explains in detail the learning and the model settings that were considered for every data augmentation approach.

4.1 Model Settings

All the models were trained using the transformer architecture of Vaswani et al. (2017) using the open-source machine translation toolkit joeyNMT by Kreutzer et al. (2019). The model parameters were set to 512 hidden units and embedding dimension, 4 layers of self-attentional encoder decoder with 8 heads. The byte pair encoding embedding dimension was set to 256. Adam optimizer is used throughout all experiments with a constant learning rate of 0.0003 and dropout was set at 0.3. All the models were trained on 40 epochs.

4.2 Evaluation Metrics

The models were evaluated using in-domain test sets. The performance of the different approaches was evaluated using different translation evaluation metrics: BLEU Papineni et al. (2002), METEOR Banerjee & Lavie (2005) and chrF Popović (2015). BLEU(Bilingual Evaluation Understudy) is an automatic evaluation metric that is said to have high correlation with human judgements and is used widely as the preferred evaluation metric. METEOR(Metric for Evaluation of Translation with Explicit Ordering) is based on generalized concept of unigram matching between the machine translations and human-produced reference translations unlike BLEU and is calculated by getting the harmonic mean of precision and recall. ChrF is a character n-gram metric, which has shown very good correlations with human judgements especially when translating to morphologically rich languages. The higher the score of these metrics means that the system produces really good translations.

5 Results and Discussion

This section describes the results of the three methods; The baseline (S-NMT), the word2vec-based + tfidf (Word2Vec) augmentation and masked language model augmentation(MLM). Table 4 shows the performance of the different data augmentation methods applied in machine translation. The

BLEU scores for the EN ↔ SW, domain specific best performing results are highlighted for each direction with the bolded scores displaying the overall best scores. We observe that in all the test domains, the models trained with the MLM-augmented data performed better than both the baseline and Word2Vec in most cases. These results are highly related to the fact that the MLM-based augmentations are based on contextual embeddings. The drop in performance in some cases can be due to the fact that the structure of the sentence is not necessarily preserved while doing word or synonym replacement thus making the translation not retain it's original meaning. We can also observe that there is a degradation of performance when translating into the low-resource language for the JW300 test data but for models tested on Tanzil, the degradation occurs mostly when translating into English. The Tanzil training data that was used to train the model was quite low compared to JW300 data which explains the low scores for Tanzil as compared to JW300 data. The Word2Vec + Tf-idf based augmentations do not lead to significant improvements of the baseline model, however, the results show there is potential in using these methods in NMT especially the Masked Language Model for augmentation which proved to perform better than the Word2Vec+Tf-idf model

6   Limitations and Future Work

One of the biggest challenges in machine translation today is learning to translate low-resource language pairs with technical challenges such as learning with limited data or dealing with languages that are distant from each other.

This paper shows that we could potentially use simple data augmentation methods in machine translation. In our experiments, we only augment the source language for the Word2Vec based augmentation method and only augment the English sentences for the MLM based augmentations. In Future work, we plan on exploring augmenting the target side of the parallel data in the Word2Vec-based augmentation and compare performance

to the source language augmentation as well as testing the model's ability to generalize by using out-of domain datasets. Another experiment that could be explored is the use of the Word2Vec data augmentation method only without the use of Tf-idf word replacement method as it adds more noise to the sentences. We plan on continuing this research and will make available the algorithms used in this paper at *https://github.com/dsfsi/translate-augmentationhttps://github.com/dsfsi/translate-augmentation*

6.0.1   Computational Considerations

Training time took about 1 hour running one epoch using NVIDIA Tesla V100 GPU on Google Cloud on the augmented texts. Running on Colaboratory took about 5 days to run 40 epochs and with limited time on our hands, there is only so much we could experiment. Running these experiments was quite expensive and there needs to be consideration of budgets as well as time so as to run these MT experiments.

7   Conclusion

In this work we proposed the use of different textual data augmentation tasks in neural machine translation using the low-resourced language Swahili. We also showed how one can perform data augmentation on the low resourced language using pre-trained word vectors and presented baseline results in ChrF and METEOR which have never been presented before. Our investigation shows that although the models trained on the augmented texts did not improve on the baseline model, there is still potential to using these methods in NMT tasks with enough compute and more experiments. We hope that this work will set the stage for further research on applying simple augmentation methods that don't require a lot of computation power in low-resource NMT modelling.

Model	Domain	BLEU		METEOR		ChrF	
		en-sw	sw-en	en-sw	sw-en	en-sw	sw-en
S-NMT(BPE)	JW300	45.30	46.54	66.32	62.21	65.92	**71.30**
	Tanzil	27.48	24.66	50.43	**50.41**	**52.51**	46.69
Word2Vec	JW300	45.23	45.52	65.90	66.56	65.02	61.76
	Tanzil	26.29	**25.80**	49.45	42.43	45.78	40.68
MLM	JW300	**45.32**	**46.98**	**66.56**	**70.55**	**65.94**	69.68
	Tanzil	**29.26**	24.86	**58.31**	49.31	48.23	**47.47**

Table 4: BLEU, ChrF, Meteor scores for Swahili ↔ English directions, domain-specific best performing results are in bold.

8 Bibliographical References

References

Agić, Ž. & Vulić, I. (2019), JW300: A wide-coverage parallel corpus for low-resource languages, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics.
URL: *https://www.aclweb.org/anthology/P19-1310*

Banerjee, S. & Lavie, A. (2005), METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, *in* 'Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization', Association for Computational Linguistics, Ann Arbor, Michigan, pp. 65–72.
URL: *https://www.aclweb.org/anthology/W05-0909*

Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. (2014), On the properties of neural machine translation: Encoder–decoder approaches, *in* 'Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation', Association for Computational Linguistics, Doha, Qatar, pp. 103–111.
URL: *https://aclanthology.org/W14-4012*

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2017), Bag of tricks for efficient text classification, *in* 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers', Association for Computational Linguistics, Valencia, Spain, pp. 427–431.
URL: *https://aclanthology.org/E17-2068*

Kreutzer, J., Bastings, J. & Riezler, S. (2019), Joey NMT: A minimalist NMT toolkit for novices, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations'.
URL: *https://www.aclweb.org/anthology/D19-3019*

Kumar, V., Choudhary, A. & Cho, E. (2020), Data augmentation using pre-trained transformer models, *in* 'Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems', Association for Computational Linguistics.

Lakew, S. M., Negri, M. & Turchi, M. (2020), 'Low Resource Neural Machine Translation: A Benchmark for Five African Languages', *arXiv e-prints* p. arXiv:2003.14402.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P. & Neubig, G. (2019), Choosing transfer languages for cross-lingual learning, *in* 'Proceedings of the 57th Annual Meeting of the Association for Compu-

tational Linguistics', Association for Computational Linguistics.

Lisanza, V. (2021), *Swahili gaining popularity globally*. Accessed: 2020–30.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv e-prints* p. arXiv:1907.11692.

Marivate, V. & Sefara, T. (2020), 'Improving short text classification through global augmentation methods', *Machine Learning and Knowledge Extraction* .
   URL: *http://dx.doi.org/10.1007/978-3-030-57321-8₂1*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems', Vol. 26, Curran Associates, Inc.

Miller, G. A. (1995), 'Wordnet: A lexical database for english', *COMMUNICATIONS OF THE ACM* **38**, 39–41.

Papineni, K., Roukos, S., Ward, T. & 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', Philadelphia, Pennsylvania, USA.
   URL: *https://www.aclweb.org/anthology/P02-1040*

Popović, M. (2015), chrF: character n-gram F-score for automatic MT evaluation, *in* 'Proceedings of the Tenth Workshop on Statistical Machine Translation', Lisbon, Portugal.
   URL: *https://www.aclweb.org/anthology/W15-3049*

Ramos, J. (2003), Using tf-idf to determine word relevance in document queries.

Sennrich, R., Haddow, B. & Birch, A. (2016), Neural machine translation of rare words with sub-

word units, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)'.
   URL: *https://www.aclweb.org/anthology/P16-1162*

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to sequence learning with neural networks', *CoRR* **abs/1409.3215**.
   URL: *http://arxiv.org/abs/1409.3215*

Tiedemann, J. (2012), Parallel data, tools and interfaces in OPUS, *in* 'Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA).
   URL: *http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_paper.pdf*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention Is All You Need', *arXiv e-prints* p. arXiv:1706.03762.

W3Techs (2020), *W3Techs. Usage of content languages for websites*. Accessed: 2020–30.

Wei, J. & 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)', Association for Computational Linguistics, Hong Kong, China, pp. 6382–6388.
   URL: *https://www.aclweb.org/anthology/D19-1670*

Xia, M., Kong, X., Anastasopoulos, A. & Neubig, G. (2019), Generalized data augmentation for low-resource translation, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 5786–5796.
   URL: *https://www.aclweb.org/anthology/P19-1579*

Zoph, B., Yuret, D., May, J. & Knight, K. (2016), Transfer learning for low-resource neural ma-

chine translation, *in* 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Austin, Texas, pp. 1568–1575. **URL:** *https://www.aclweb.org/anthology/D16-1163*