# Investigation of text data augmentation for transformer training via translation technique

**Dominykas Šeputis**

Vilnius University, Faculty of Mathematics and Informatics,
Institute of Computer Science,
Didlaukio str. 4, LT-08303 Vilnius
*dominykas.seputis@mif.stud.vu.lt*

**Abstract.** Data augmentation can improve model's final accuracy by introducing new data samples to the dataset. In this paper, text data augmentation using translation technique is investigated. Synthetic translations, generated by Opus-MT model are compared to the unique foreign data samples in terms of an impact to the trans- former network-based models' performance. The experimental results showed that multilingual models like DistilBERT in some cases benefit from the introduction of the addition artificially created data samples presented in a foreign language.

**Keywords:** Data Augmentation, Transformer, Fine-tuning, Machine Translation, DistilBERT, Opus-MT

## 1 Introduction

Over-paramterised models like neural networks, tend to benefit from large datasets in computer vision [1, 2] and natural language processing fields [3]. The work of Zhu *et al.* [4] and alike suggest that given the size of existing datasets, it appears that the current state-of-the-art will need significant additional data (perhaps exponentially larger sets) to continue producing consistent improvements in performance.

Since the publication of the work of Vaswani *et al.* [5] a substantial improvement has been achieved by scaling up the size of the transformers as well as the size of the training data used for training [6, 7, 8, 9, 3]. GPT-2 [10], one of the largest NLP model was trained on over 8 million documents for a total of 40 GB of text. The creation of public datasets like Pile [11] - an 800 GB text dataset confirms the mentioned growth of data samples size. There also exists a critics for the enormous (in term of parameters) networks.


DOI: https://doi.org/10.15388/LMITT.2021.11

Work of Bender *et al*. [12] suggests that weighing the environmental and financial costs, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web should be done and encourages the research directions beyond ever larger language models.

In the work of Ciolino *et al.* [13] acknowledged the problem of data-hungry transformer network-based models and suggested a back-translation process of translating text from English to another language and then back to English. In the mentioned work it is concluded that back translation shows a significant ability to move the various Natural language processing (NLP) metrics in many transformer architectures and the text augmentation technique empirically shows back-translation acts as a generalizable strategy.

Motivated by the text augmentation problem, in this work the idea of using text translations as a data augmentation technique is explored. First it is tested if adding an alternative language samples to the training data enhances the performance. Then a technique for fine-tuning multilingual transformers is presented. In contrast to Ciolino *et al.* [13], back-translations is not used and the traslations are kept as additional samples in the training data.

## 2 Datasets Configuration

### 2.1 Data Samples Selection

Before making any augmentations, it has be made sure that appending different languages to the dataset can improve the metrics. For this purpose a dataset that already has multilingual samples in it has to be used. "The Multilingual Amazon Reviews Corpus" [14] is the dataset that meets the requirements and is therefore the dataset used for the experiments. The dataset's corpus contains reviews in English, Japanese, German, French, Spanish, and Chinese, which were collected between 2015 and 2019. Each record in the dataset contains the review text, the review title, the star rating, an anonymized reviewer ID, an anonymized product ID, and the coarse-grained product category (e.g., 'books', 'appliances', etc.) The corpus is balanced across the 5 possible star ratings, so each rating constitutes 20% of the reviews in each language.

In order for the dataset to be suitable for our experiments, the following changes were made. As the training task was set to classify reviews to posi-

tive and negative ones, one and two star reviews got a “negative” label and four-star and five-star reviews were assigned a “positive” label. Three-star reviews were removed from the dataset for being in between the positive and negative review type. As the dataset was balanced across all the 5 possible ratings, removing one of them did not change the balance.

The next step was selecting the main and alternative languages of the review. English was selected as the main one and German, French, Spanish languages were selected as the alternative ones. We did not use Japanese and Chinese languages due to the alphabet and language structure differences.

As the computational resources were limited (one Nvidia Tesla P100 GPU) and finite time was allotted, the experiment was made doable by capping the maximum reviews count. The final number of samples was selected: 30 000 reviews (15 000 positive, 15 000 negative). To see if the alternative languages had any impact on the overall accuracy, 30 000 reviews in three alternative languages (10 000 for each language) were selected. These samples were selected at random, keeping the same item category distribution. Validation dataset was changed to have only English text (4000 samples), and did the same to the test set (4000 samples).

### 2.2 Text Translation

Once the dataset was selected, the next step was to make translations. There are two ways to translate text:

- Online cloud solutions (provided by Google, Microsoft, etc.)
- Local neural machine translation models

For this research, models were selected for the price per word efficiency. The exact implementation used was EasyNMT tool - wrapper for neural machine translation models. The Opus-MT [15] model was selected for being the fastest and the most accurate solution to choose from. The translations were applied to the English dataset, translating 30 000 unique reviews.

### 2.3 Final Datasets

The final data selection resulted in three different datasets: English language (30 000 samples), English and alternative languages (60 000 samples), English language and translated text (60 000 samples). The graphical representation of datasets configuration is given in Fig. 1.

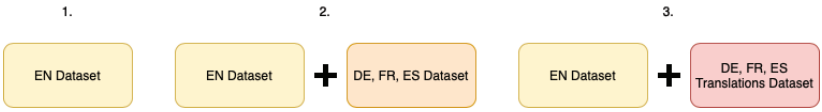


**Figure 1:** Datasets configuration

## 3. Training Pipeline

### 3.1 Making a Fare Comparison

To test the validity of the set hypothesis variability in between training runs has to be accounted. It is done so by averaging the results from multiple training runs with varying hyperparameters as follows:

- Variations of model's hyperparameters:
  - Batch size: 16, 32
  - Max sequence length: 128, 256, 512
- Data sample variations:
  - English dataset:
    - 5 000 samples took three times with three unique random states
    - 30 000 samples
  - English and alternative languages dataset:
    - 5 000 samples sampled three times with three unique random states
    - 30 000 samples sampled three times with three unique random states
    - 60 000 samples
  - English and translations dataset:
    - 5 000 samples took three times with three unique random states
    - 30 000 samples took three times with three unique random states
    - 60 000 samples

In the end 114 training runs with varying hyperparameters were run (each dataset variation was combined with each possible max sequence length and batch size parameter). The training pipeline's scheme is presented in the Fig. 2.

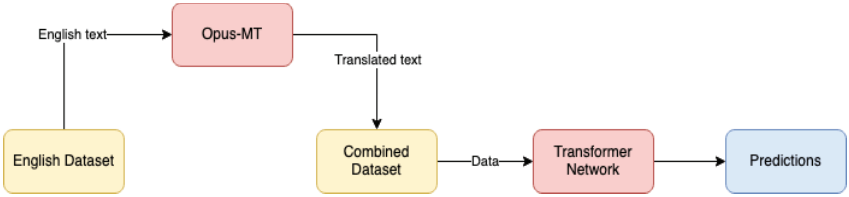


**Figure 2:** Training pipeline with appended translated text to the dataset

### 3.2 Transformer Network Architecture

For this experiment a DisilBERT [16] transformer network implementation was selected. The implementation offers a pre-trained multilingual model weights (distilbert-base-multilingual-cased), is a reduced size (40% smaller) of a BERT model but retains 97% of its language understanding capabilities and is 60% faster.

## 4 Experiment Results

An experimental investigation consisted of evaluation of text augmentation techniques for transformer training. To evaluate the model performance, results were grouped by samples count and the averaged F1 score of each group was calculated. The obtained results are given in Table 1.

**Table 1:** Experimental investigation's results. Mean F1 score of different group of samples count presented.

| Dataset name | Samples count | F1 score (mean) |
|---|---|---|
| English dataset | 5 000 | 86.2% |
| English dataset | 30 000 | 89.2% |
| English + alternative languages dataset | 5 000 | 84.2% |
| English + alternative languages dataset | 30 000 | 88.8% |
| English + alternative languages dataset | 60 000 | 90.8% |
| English + translations dataset | 5 000 | 84.3% |
| English + translations dataset | 30 000 | 89.2% |
| English + translations dataset | 60 000 | 90.6% |

The results showed that replacing English language samples with alternative language samples or translation samples did not improve the F1 score. The better results can be seen only by appending (30 000 English samples + 30 000 samples in different language) alternative language or translation samples to the dataset: the data augmentation technique moved the F1 score metric by 1.77% when appending alternative languages to the dataset and by 1.58% when appending translations to the dataset.

The biggest findings were that translated text addition action performs exactly the same as the alternative language addition. A graphical representation of the results is given in Fig. 3. This finding can suggest that if alternative language addition can improve performance of the model, translated text addition can do it too.

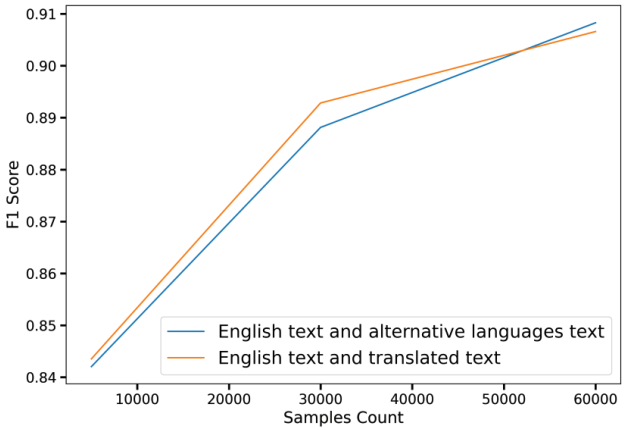


**Figure 3:** F1 score comparison between model trained with English language and translated text and model trained with English and alternative languages text.

## 5 Limitations of the Study

**Model selection**. The work focused only on one transformer type network. Bigger (in term of parameter number) or different architecture models could be impacted by the augmented data differently.

**Alternative language selection**. As the main language of the dataset selected was English and alternative ones Spanish, French and German. In the work of Aharoni *et al.* [17] suggests that different language families are located in a different places of encoded representation's space, so different main and alternative languages combinations could be explored.

**Dataset diversity and data sample selection issues**. The whole study was done only on one dataset. If a different dataset would be selected, there are multiple factors that could change, thus influencing the results: content of the text (during the experiments only the item reviews were presented in the text), different language groups (see Fig. 4), count of data samples. Furthermore, the selected dataset was not fully explored as only a fraction of the available data was explored.

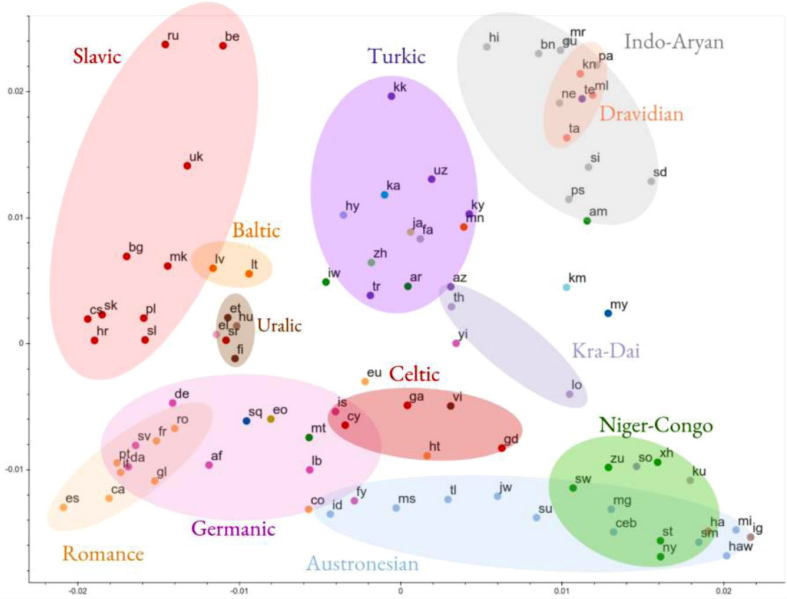


**Figure 4:** Visualization of the clustering of the encoded representations of different languages, based on representational similarity. Languages are color-coded by their linguistic family. Product, by, Google AI Blog (https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html), 2019

**Text translation efficiency and technique variety**. Text translations for the data augmentation were translated using only one tool - Opus-MT. The whole text translation process was not efficient and required a lot of computing power. Different types of available translation models should be tested too. Cloud-based services should be also investigated as they could offer a more accurate and faster translation process.

## 6 Conclusions and Future Work

In this paper data augmentation via text translation technique for the NLP tasks was investigated. The main insight of the experiment is that multilingual transformer architecture based models in some cases can acquire a slight advantage from the addition of alternative language samples. Another insight is that translation models like Opus-MT can perform high quality translations of simple sentences that can be used as an alternative to the generated text by native speakers.

Checking the hypothesis of beneficial data augmentation technique via text translations on one transformer-based models is only the first step. To fully complete the work, all the mentioned shortcomings of the paper would have to be addressed. Coulombe [18] and the works alike suggest that other types of text augmentation like textual noise injection, spelling errors injection, word replacement using a thesaurus, and paraphrases generation using a regular expression, paraphrases generation using syntactic tree transformations can help to overcome the fact of not having enough data, so a combination of multiple augmentation techniques could be also explored.

### References


[1] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision, pages 843–852, 2017.

[2] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In Proceedings of the European Conference on Computer Vision (ECCV), pages 181–196, 2018.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

[4] Xiangxin Zhu, Carl Vondrick, Charless C. Fowlkes, and Deva Ramanan. Do We Need More Training Data? International Journal of Computer Vision, 119(1):76–92, August 2016.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[7] VictorSanh,LysandreDebut,JulienChaumond,andThomasWolf.Distilbert,adistilledversion of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

[8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.
[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.
[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.
[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027 [cs], December 2020. arXiv: 2101.00027.
[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?; In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 610–623, Virtual Event, Canada, March 2021. Association for Computing Machinery.
[13] Matthew Ciolino, David Noever, and Josh Kalin. Multilingual Augmenter: The Model Chooses. arXiv:2102.09708 [cs], February 2021. arXiv: 2102.09708.
[14] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus, 2020.
[15] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.
[16] VictorSanh,LysandreDebut,JulienChaumond,andThomasWolf.Distilbert,adistilledversion of bert: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108, 2019.
[17] Ankur Bapna. Exploring Massively Multilingual, Massive Neural Machine Translation. http://ai.googleblog.com/2019/10/exploring-massively-multilingual.html. Accessed: 2021- 03-28.
[18] Claude Coulombe. Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs. arXiv:1812.04718 [cs], December 2018. arXiv: 1812.04718.