Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 1162–1169

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Pseudo-labeling with transformers for improving Question Answering systems

Karolina Kuligowska[a,*], Bartłomiej Kowalczuk[a]

[a]*Faculty of Economic Sciences, University of Warsaw, Dluga St. 44/50, 00-241 Warsaw, Poland*

Abstract

Advances in neural networks contributed to the fast development of Natural Language Processing systems. As a result, Question Answering systems have evolved and can classify and answer questions in an intuitive yet communicative way. However, the lack of large volumes of labeled data prevents large-scale training and development of Question Answering systems, confirming the need for further research. This paper aims to handle this real-world problem of lack of labeled datasets by applying a pseudo-labeling technique relying on a neural network transformer model DistilBERT. In order to evaluate our contribution, we examined the performance of a text classification transformer model that was fine-tuned on the data subject to prior pseudo-labeling. Research has shown the usefulness of the applied pseudo-labeling technique on a neural network text classification transformer model DistilBERT. The results of our analysis indicated that the model with additional pseudo-labeled data achieved the best results among other compared neural network architectures. Based on that result, Question Answering systems may be directly improved by enriching their training steps with additional data acquired cost-effectively.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of KES International.

Keywords: Natural Language Processing; Question Answering systems; pseudo-labeling; neural networks; transfer learning; knowledge distillation

1. Introduction

Question Answering (QA) systems provide a way to find an answer to a question given some context - the knowledge environment they are fed during training. Despite the intensity of research and technology development

* Corrsponding author email address: kkuligowska@wne.uw.edu.pl

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of KES International.
10.1016/j.procs.2021.08.119

for natural language understanding, text processing, and information retrieval, present QA systems still lack a deep, multidimensional understanding of context when responding to a given question. According to recent researches of Bjerva [2] and of Muttenthaler [13], the reason behind these limitations is the dataset. Although current QA systems aim at high accuracy, labeled datasets used so far do not provide enough information about, among others, intent understanding, helpfulness of answers, or question correctness.

We have examined current state-of-the-art neural network architectures for language models. Despite great quality of recurrent network models, they still lack the knowledge that could be acquired by combining direct insight from both sides of the sequence simultaneously. Vaswani in 2017 [19] covered this issue by introducing a new architecture: the transformer, which is based solely on the attention mechanisms, and completely removes recurrence. Many new ideas using transformer architecture arose, among others, Bidirectional Encoder Representations from Transformers (BERT), Generative Pretrained Transformer version 2 (GPT-2) and Robustly Optimized BERT (RoBERTa) were proposed [5] [11] [16]. First of them, BERT, uses the encoder part of the transformer architecture, meaning that it processes the whole sequence at once, as opposed to GPT-2 where only the decoder part is utilized. RoBERTa uses the same tokenizer as GPT-2 and is based on the same architecture as BERT, but it uses a different pre-training scheme.

Among the above-mentioned transformer architectures, BERT has been the first fine-tuning based representation model that achieved state-of-the-art performance on a wide range of tasks in GLUE benchmark [5]. Nonetheless, achieving such results is strictly connected to the amount of time that transformers must be trained and the amount of data that they consume during training [17]. To compete with that, one has to use large clusters and huge amounts of data. Therefore, in this paper we are going to use DistilBERT model that utilizes process of knowledge distillation - a compression technique in which a small model (student) is trained to reproduce the behavior of a larger model (teacher) [7] [17].

DistilBERT has the same general architecture as BERT, but here the non-efficient elements were removed and the number of layers was reduced by a factor of 2. What is more, the student model is initialized with some of the layers from the original BERT model - exactly one out of two, which vastly helped it to converge. DistilBERT model used by us in this paper is 40% smaller and 60% faster than BERT, while maintaining 97% of language understanding capabilities. Thanks to its high performance, light-weight and short computation time, it constitutes a great candidate for the system that would evaluate results of other algorithms or index questions and answers in databases, providing new tailor-made solutions for knowledge evolvement in many scientific areas and business use cases.

Having clarity on the chosen transformer model, we propose fine-tuning on the dataset that includes labels measuring human-like opinion-making and inference. In 2019 Google LLC shared a publicly experimental set of data (https://www.kaggle.com/c/google-quest-challenge/data) addressing quality aspects of question-answering. This dataset is formulated in English and includes 6072 distinct answers and 3582 distinct questions with adequate two target labels categories (answer or question). We use this dataset to build predictive algorithms for different subjective aspects of QA system and to further improve QA systems scoring. We present our novel approach of semi-supervised learning pseudo-labeling technique based on Google LLC dataset relying on a neural network text classification transformer model light-weight DistilBERT. Obtained results may positively impact the QA research area and contribute to obtain relevant human-level performance of intelligent QA systems.

This paper is organized as follows. Section 2 presents limitations related to present Question Answering systems, and Section 3 describes our proposed solution. Section 4 explains pseudo-labeling initial steps that we took as well as evaluation metrics that we applied. Section 5 presents the characteristics of our dataset and describes preprocessing. Next, in Section 6, we conduct modeling, describe the best model, and discuss the final evaluation. The paper ends with conclusions in Section 7, including further research ideas in Section 8.

2. Question Answering systems and their limitations

In recent years, Question Answering systems have evolved into sophisticated architectures, achieving state-of-the-art described often in publications concerning conversational artificial intelligence [4] [6] [14]. Nevertheless, algorithms used so far in QA systems are relatively poor in a multidimensional understanding of context. It seems they could notably benefit from the transfer learning and models trained on large datasets. As improving QA systems

requires a huge amount of data, here outflows the need to handle the lack of labeled datasets. This need is even more pressing because labeling data is an expensive and time-consuming process.

While transfer learning has become the new standard of dealing with NLP problems [17], the main flaw of transformer models is still their size and computation time, both for training (or fine-tuning) and inference. They are challenging when it concerns implementing them into production and even more problematic to maintain under low latency constraints.

In response to that issue, the DistilBERT model, that uses the process of knowledge distillation, was selected [7] [17]. DistilBERT provides a way to efficiently train and integrate transformer-based solutions into a wide range of applications. Again, the need for a workaround for the small amount of labeled data emerges. What if we could fine-tune our model on the data that we already have labeled and then use it to label new, unlabeled data and finally re-train the initial model on the combined dataset? The novel pseudo-labeling technique is the answer.

First proposed by Lee [9], pseudo-labeling requires the initial model to be trained on a small set of labeled data, followed by training on the combination of labeled and pseudo-labeled data, which are the results of logits (outputs) of the initial model. It is based on the assumption that to improve generalization, the decision boundary of clusters should lie in low-density regions. It increases the size of the dataset by adding the noise from pseudo-labels which may further separate the classes and highlight low-density regions due to the possible boost in the population of higher-density regions. Adding pseudo-labeled data also minimizes the entropy for unlabeled data, which reduces the overlap of class probability distribution by favoring low-density separation [3]. This approach (and its further modifications) recently achieved extraordinary results in the computer vision domain [8] [21]; moreover, some other researches [1] [10] also proved its legitimacy in the NLP domain.

3. Proposed solution

We propose a novel approach to overcome the lack of large volumes of labeled data by applying a pseudo-labeling technique that takes advantage of a neural network transformer model architecture DistilBERT. According to our proposed solution, the whole process of modeling consists of the following steps:

1. Model fine-tuning on a training dataset.
2. Computing pseudo-labels for the unlabeled dataset.
3. Adding pseudo-labeled data to the initial training dataset.
4. Further, fine-tuning of the initial model on combined datasets.

We decided to examine the performance of the proposed text classification solution against the following four various neural network model architectures:

- Recurrent Neural Network model (Bidirectional GRU) with GloVe embeddings and one-dimensional convolutional neural network (CNN) encoder
- DistilBERT model with linear layer ([CLS] token)
- DistilBERT model with CNN and linear layer
- DistilBERT model with RNN (Bidirectional GRU) and linear layer

An important difference between the DistilBERT with linear layer and the other two DistilBERT variations how they utilize DistilBERT outputs. The first classifies targets from the [CLS] token representation, and the other two classify token level hidden states from the DistilBERT's output layer. We presume that the DistilBERT model with CNN and DistilBERT model with RNN will take advantage of each token representation in the output layers. In contrast, the DistilBERT with linear layer will rather base on the DistilBERT abilities to aggregate representations in the [CLS] token.

4. Pseudo-labeling data and evaluation metrics

The pseudo-labeling process requires additional data that is not yet labeled. Data shared by Google LLC was fully scraped from Stack Exchange - an online network that comprises almost two hundred QA communities focused on various general topics, among others such as technology, culture, science, life arts, and others. Collected

data contains observations that consist of a question title, a question content and subsequent answers (plus a few additional features), along with the 30 various target labels - all with continuous values in the range [0,1]. We further expanded the dataset using the appropriate data archive (https://archive.org/download/stackexchange), and we selected the 10 most frequent categories (hosts, domain names) included in the training dataset, and we filtered unlabeled data to extract only those categories - it covered 53882 distinct questions.

We decided to randomly extract samples from this dataset: 10%, 30%, 50% of data to use in our experiment - we wanted to prevent the model from overfitting to pseudo-labels. That situation could happen if we used too much of the additional dataset. Further model tuning showed that a random sample of 30% of data improved the performance best. Thus, we will focus on this sample and omit descriptions of other samples. Proceeding this way, we extracted 16160 unique questions for the pseudo-labeling phase. We also verified if any of the questions in the extracted dataset is not doubled and did not appear in our initial dataset to prevent any potential data leakage.

Label values in the modeled dataset are not binary, and they are continuous in the range [0,1]. Thus, for training purposes, we use binary cross-entropy loss. Models trained on this particular dataset should be evaluated with Spearman's rank correlation coefficient. Choice of this metric can be argued by the fact that the target labels are human-rated based on common sense. Therefore, we focus on the order of those variables and not necessarily on the magnitude of difference between them – and that is why we are using rank correlation.

The formula presented below is computed as the covariance of the ranks of the variables (rg_x, rg_y) divided by the multiplied standard deviations of the rank variables. Coefficient should be calculated for each target label and then average among all:

$$r_s = \rho\,(rg_x, rg_y) = cov(rg_x, rg_y) / \sigma rg_x \;\; \sigma rg_y$$

This metric is defined as the Pearson correlation coefficient between rank variables, meaning that we measure a monotonic relationship between two variables.

5. Dataset description and preprocessing

The examined dataset includes 6072 distinct answers and 3582 distinct questions; all data text is formulated in English. Target labels split into two categories based on what they evaluate - question or answer. The first category focuses on question-related aspects, such as fact seeking, multi-intent, or reasonable explanation. The second category focuses on answers and scores the following aspects: relevance, level of information, or instructions.

Questions concern five different categories with following quantities: technology - 1495 distinct questions, StackOverflow - 759 questions, culture - 527, science - 412, life arts - 389. Such unbalanced categories may influence the generalization capabilities of the model. However, our experiment is not designed to examine the impact of this data property.

As the sequence input for the DistilBERT has a limited length (512 tokens), we had to adjust the sequences to that shape. Analysis of the length distribution of each dataset concluded in such division: question_title is set to a maximum length of 22 tokens, question_body 230 tokens, and answer 260 tokens. We set the same lengths for the Recurrent Neural Network model and all DistilBERT models for increased comparability.

We did not remove any stopwords from the dataset during the preprocessing step since the initial BERT model was trained on text corpus containing them. By doing so, we could deteriorate the 'model's performance—the same concerns special characters included in BERT's vocabulary (identical vocabulary is used in DistilBERT). The only transformation performed was to turn each text into lowercase to decrease the chance of unknown words due to the uppercase form (we used the uncased version of the DistilBERT model, which was trained on lowercase texts).

We randomly excluded 15% of the data for evaluation/testing purposes for all models, which resulted in the following quantities: 5165 observations in the training set and 912 in the test set. For further validation (and development) purposes, 10% of the training dataset was extracted.

Preprocessing for the Recurrent Neural Network model was initiated with tokenizing. BERT's input needs a specific preprocessing that includes tokenizing using WordPiece embeddings [20], therefore we derived the most frequent 30522 tokens to match the size of BERT's vocabulary. We combined questions and answers into sequences and padded them to the length of 512 tokens - meaning that we added zeros to the shorter sequences to unify the

lengths. Finally, tokens were matched to pre-trained embedding vectors from GloVe [15], a version of uncased 100-dimensional word vectors.

Data preparation for DistilBERT began with tokenization of the input using DistilBertTokenizer in base and uncased version [18]. It was followed by adding special tokens [CLS] at the beginning of the sequence and [SEP] between question and answer and at the end of the sequence. All sequences were also padded, and the attention masks were added to inform the model which tokens to omit during attention computation (due to the padding).

Pseudo-labeled data followed the same preprocessing scenario as the one for the DisitlBERT described above (stopwords and special characters are left, lowercase transformation, tokenization) with additional removal of HTML tags like '<p>' or ' ' due to the raw form of the data downloaded from the archive.

6. Modeling and final evaluation

All following training processes were done using hardware resources that included a single NVIDIA T4 GPU with 16GB GDDR6 memory. First, we characterize the application of the recurrent neural network model (BiGRU) + GloVe + CNN. Then we focus on DistilBERT model fine-tuning and its results. Further, we discuss the pseudo-labeling process, and we present the final evaluation of our findings.

Recurrent Neural Network model tuning, using early stopping on minimum validation loss (with 3 epochs patience) and Adam optimizer resulted in a model with the following hyperparameters: 512 batch size, 27 epochs, one-dimensional convolution layer with 256 filters and kernel of size 5, two Bidirectional GRU layers with 256 units separated by dropout of 0.1, one-dimensional Global Max Pooling layer and dense layer with 128 units and Rectified Linear activation function (ReLU).

All DistilBERT model variations were tuned on batch sizes: 16 and 32, learning rate: 5e-5, 3e-5, 2e-5, and epochs: 2,3,4 as recommended by BERT authors [5]. Model optimizing was done using Adam optimizer with a 2e-5 learning rate. Further training (e.g., more than the recommended number of epochs) may result in a phenomenon called catastrophic forgetting, in which the fine-tuned model replaces the complete knowledge of the former model by overfitting the training data [12]. All three DistilBERT models performed best on a validation set on batch size 16, with a learning rate of 2e-5 and 4 epochs.

In two DistilBERT models, an additional tuning of output layers was also performed, resulting in the following hyperparameters sets:

- DistilBERT + CNN: two one-dimensional convolution layers with 16 and 8 filters, and kernels of size 3,
- DistilBERT + RNN: Bidirectional GRU layer with 32 units in each direction, one-dimensional Max Pooling layer of pool size 8;

Results obtained after processing on the validation dataset are presented in Table 1.

Table 1. Models evaluation on the validation dataset.

Model	Spearman's rank correlation coefficient (validation data)
BiGRU + GloVe + CNN	0.2817
DistilBERT ([CLS] token)	0.3677
DistilBERT + CNN	0.3221
DistilBERT + RNN	0.3395

Source: Authors own elaboration.

Based on the results obtained from this phase, we chose the best performing model: DistilBERT with a linear layer on top of [CLS] token representation. We will use this model in the pseudo-labeling phase.

Computing predictions on the unlabeled dataset initiated the Pseudo-labeling phase, and then these results were joined to the set as labels. This new dataset was concatenated with the initial training dataset, and the best-performing model was further fine-tuned for two additional epochs on the combined data. We want to point out that the validation set was unchanged through all phases, and no data from the pseudo-labeling phase was included in it.

To further validate the performance of the pseudo-labeled model, we trained the best-performing model on two additional epochs (like the pseudo-labeled one). But only on the initial training set - this way, we may find evidence that the results obtained are not only impacted by the training length.

In order to assess the performance of all models, we compared scores that each model achieved on the test data. Results of the model evaluation are presented in Table 2. Spearman's rank correlation was rounded to four decimal places for increased readability, but it is not relevant for evaluating the final order of results.

Table 2. Models evaluation on the test dataset.

Model	Spearman's rank correlation coefficient (test data)
BiGRU + GloVe + CNN	0.2817
DistilBERT ([CLS] token)	0.3785
DistilBERT + CNN	0.3489
DistilBERT + RNN	0.3611
DistilBERT ([CLS] token) - additional 2 epochs	0.3787
DistilBERT ([CLS] token) with pseudo-labels	0.3866

Source: Authors own elaboration.

Further training of additional two epochs of the DistilBERT model on training data resulted in an almost imperceptible change in the score, whereas the fine-tuning on pseudo-labeled data made a significant change. It turned out that using a larger dataset, that includes pseudo-labeled data, improved the results of the same architecture DistilBERT model. We can finally conclude that the DistilBERT model with used pseudo-labeling method outperformed all other examined architectures, thus proving the legitimacy of the proposed solution.

7. Conclusions

The paper's aim was accomplished, and a novel approach for overcoming the lack of labeled datasets for improving Question Answering systems was examined. We described applying pseudo-labeling technique into a neural network text classification transformer model and proved its legitimacy. The DistilBERT model with additional pseudo-labeled data was compared against four other neural network architectures. Research has shown that the DistilBERT model with additional pseudo-labeled data achieved the best result among compared architectures. Therefore indeed, the lack of large volumes of labeled data may be overcome by applying a pseudo-labeling technique. Based on that finding, the training stages of Question Answering systems may be improved with additional data in training steps, and the process can be reproduced using limited capital and time.

We want to emphasize that the value of improvement Question Answering systems derives from the pure performance of text classification predictions and the involved dataset that included subjective aspects that usually do not appear in widespread popular datasets. Comprehensive, subjective, and contextual aspects of evaluating connections between questions and answers are crucial for developing NLP algorithms towards producing more human-like solutions.

We hope this paper will constitute an incentive for business representatives to introduce techniques like pseudo-labeling and architectures like DistilBERT into their Question Answering systems and architectures. Considering obtained results and low costs of fine-tuning and maintenance, it seems a perfect solution to explore daily for every

company. Companies with a small volume of labeled data may change their approach to pseudo-labeling and eliminate the need for additional time and capital-consuming human labeling process, not risking computation costs when using neural network architectures, and especially distilled version of BERT like DistilBERT. Such solution would eventually contribute to any domain or science as a whole, providing evolution by assigning appropriate answers to questions or even selecting crème de la crème of all questions from a specific domain.

8. Further research - discussion

We expect the pseudo-labeling process to work the best if we have a few thousand observations already labeled and a lot more unlabeled data. This will allow the initial classifier to learn the data structure and sufficiently label additional data. We could imagine an example scenario: a company has conversational data from help desks where clients, at the end of the talk, label the conversation whether it was satisfactory or not. The labeling is not necessary, so only a relatively small part of the data is used. The company may try to build a satisfaction classifier based on this labeled data and additionally experiment with pseudo-labeling on the rest of the data. This way, the classifier could explore previously unseen data and adjust its form to generalize better.

As further research, we propose to explore the following ideas:

- Apart from testing other transformer model architectures like TinyBERT, RoBERTa, or XLNet, we would highly encourage experimenting with linear combinations of the model's predictions in an ensemble modeling approach.
- Another way to further improve the Question Answering model could also be to pre-train it on a Language Modelling method basing on the data used in the training phase. This way, the model would acquire knowledge from the data that might have been omitted in the initial pre-training.
- It could also be promising to experiment with the use of different hidden states from transformer models, e.g., instead of using the last hidden state of [CLS] token, and one could compute a weighted average of all of them.

References

[1] Ahmed, M. S., Khan, L., and Oza, N. C. (2011) "Pseudo-Label Generation for Multi-Label Text Classification", *Proceedings of NASA Conference on Intelligent Data Understanding*, CIDU 2011, pp. 60-75.

[2] Bjerva, J., Bhutani, N., Golshan, B., Tan, W., and Augenstein, I. (2020) "SubjQA: A Dataset for Subjectivity and Review Comprehension". ArXiv, abs/2004.14283, pp.1-12.

[3] Chapelle, O., and Zien, A. (2005) "Semi-Supervised Classification by Low Density Separation*". Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, AISTATS 2005, pp.57-64.

[4] Chung, Y-A., Lee, H-Y., Glass, J. (2018) "Supervised and Unsupervised Transfer Learning for Question Answering", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pp. 1585–1594. DOI: https://doi.org/10.18653/v1/N18-1143

[5] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *NAACL-HLT 2019*, vol.1, pp.4171-4186.

[6] Gao, J., Galley, M., and Li, L. (2019), "Neural Approaches to Conversational AI", *Foundations and Trends in Information Retrieval*, Vol. 13: No. 2-3, pp. 127-298. DOI: http://dx.doi.org/10.1561/1500000074

[7] Hinton, G.E., Vinyals, O., and Dean, J. (2015) "Distilling the Knowledge in a Neural Network". *Proceedings of the 29th Conference on Neural Information Processing Systems*, NIPS 2015, pp.1-9.

[8] Iscen, A., Tolias, G., Avrithis, Y., and Chum, O. (2019) "Label Propagation for Deep Semi-Supervised Learning". *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2019, pp.5065-5074.

[9] Lee, D. (2013) "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". *ICML 2013 Workshop: Challenges in Representation Learning (WREPL)*, pp.1-6.

[10] Lee, S., Dai, D., Li, S., and Ahrens, K. (2011) "Extracting Pseudo-Labeled Samples for Sentiment Classification Using Emotion Keywords". *2011 International Conference on Asian Language Processing*, IALP 2011, pp.127-130. DOI: https://doi.org/10.1109/IALP.2011.61

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach", International Conference on Learning Representations (ICLR 2020). ArXiv, abs/1907.11692, pp.1-13.

[12] McCloskey, M., and Cohen, N. J. (1989) "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem". *Psychology of Learning and Motivation*, vol. 24, Elsevier 1989, pp.109-165. DOI: https://doi.org/10.1016/S0079-7421(08)60536-8

[13] Muttenthaler, L. (2020) “Subjective Question Answering: Deciphering the inner workings of Transformers in the realm of subjectivity”. *Computer Science Computation and Language, ArXiv*, abs/2006.08342, pp.9-12.

[14] Pendharkar, D., and Gupta, G. (2019) “An ASP Based Approach to Answering Questions for Natural Language Text”. *Proceedings of 21th International Symposium,* PADL 2019*,* pp. 46–63. DOI: https://doi.org/10.1007/978-3-030-05998-9

[15] Pennington, J., Socher, R., Manning, Ch. (2014) “GloVe: Global Vectors for Word Representation”. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014. Association for Computational Linguistics, pp.1532-1543. DOI:https://doi.org/10.3115/v1/D14-1162

[16] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). “Language models are unsupervised multitask learners”. *Technical report, OpenAI*, pp.1-24.

[17] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019) “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. *ArXiv*, abs/1910.01108, pp.1-5.

[18] Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., and Lin, J. (2019) “Distilling Task-Specific Knowledge from BERT into Simple Neural Networks”. *ArXiv*, abs/1903.12136, pp.1-8.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). “Attention is All you Need”. *Proceedings of the 31st Conference on Neural Information Processing Systems, NIPS 2017*, pp.1-15.

[20] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., and Dean, J. (2016). “Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. *Computing Research Repository (CoRR) at ArXiv*, abs/1609.08144v2, pp.7.

[21] Xie, Q., Hovy, E., Luong, M., and Le, Q.V. (2020) “Self-Training With Noisy Student Improves ImageNet Classification”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2020, pp. 10687-10698.