ISSN: 2395-1680 (ONLINE)
DOI: 10.21917/ijme.2024.0314

DOMAIN-SPECIFIC TOKEN RECOGNITION USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS AND SCIBERT

Nisha Varghese[1] and Shafi Shereef[2]
[1]Department of Computer Science, Christ University, India
[2]Department of Computer Science and Information Technology, Jain University, India

Abstract

Make machines to read and comprehend information from natural language documents are not an easy task. Machine reading comprehension is a solution to alleviate this issue by extracting the relevant information from the corpus by posing a question based on the context. The problem associated with this knowledge retrieval is in the correct answer extraction from the context with language understanding. The traditional rule-based, keyword search and deep learning approaches are inadequate to infer the right answer from the input context. The Transformer based methodologies are used to excerpt the most accurate answer from the context document. This article utilizes one of the exceptional transformer models - BERT (Bidirectional Encoder Representations from Transformers) for empirical analysis for Neural Machine Reading Comprehension. This article aims to reveal the differences between the BERT and the domain-specific models. Furthermore, explores the need for domain specific models and how these models outperform the BERT.

Keywords:

BERT, Transformers, Span Extraction, SciBERT, BioBERT

1. INTRODUCTION

Machine Reading Comprehension (MRC) has the potential to make the Human-Computer interaction a reality, communicating through question answering. MRC span extraction task takes the context and question as input and excerpts the most accurate answer meticulously by employing the Deep Neural Network (DNN) techniques. The problem of the existing models is in the quality of excerpting the exact answers. The traditional keyword search algorithms and typical deep learning methods have some limitations such as the keyword search algorithms extract the possible sentences as answers without a thorough understanding.

The typical deep learning and other methods (Long Short Term Memory (LSTM), Bi-LSTM, Gated Recurrent Unit (GRU), and Recurrent Neural Networks) have problems including Curse of high dimensionality, Slow and difficulty to Train the data, Not Bi-directionally connected, Long Sequences leads to vanishing or exploding gradient descent problem, Longer gradient paths for higher dimensions and Transfer Learning never really work. The methodologies that have been taken for the empirical analysis are Transformer, BERT, BERT-based domain-specific models. The need for domain-specific models is the generally trained BERT model is not suitable for specific domain applications for proper language understanding.

The objective of the article is two fold: influence of the word-subword token length in the context of a model and the role of capturing the semantics with token and subtokens in Generalised and Domain-specific models. The article examines how efficiently the domain-specific models understanding the context using the pre-training of the foundational model on the specific domain.

2. REVIEW OF LITERATURE

The review of Literature constitutes some of the domain-specific BERT-based models along with characteristics. SciBERT [7] belongs to the biomedical and computer science domain with 18% of computer science and 82% of the biomedical domain. This model was pre-trained on the scientific articles of the semantic scholar [6] with of textual size of 1.14M. BioBERT [10] is also a biomedical domain model, which is pre-trained on massive biomedical datasets (PubMed and PubMed Central) with a corpus size of 18 billion words. ClinicalBERT [12] is a BERT-based clinical domain flexible framework for the prediction of the readmission of patients in hospitals in 30-day. Prediction is based on the health records of the patient. Some other domain-specific models are FinBERT [4] for financial services, LegalBERT [8] for Legal-Case Documents pre-trained on the case reports from United States, United Kingdom, and other European Countries, PatentBERT [11] is pre-trained on the various kinds of patents and granting details for patent classification tasks, and DocBERT [3] for Document Classification. The forementioned models are trained with specific domains for improving the accuracy in specifc tasks in the domain.

For example, LegalBERT is pre-trained on the legal corpora such as case studies, laws, court pleadings and other legal contracts. All the people in the world are under laws, the research in this area can support the people in better understanding in the area of legal domain. Legal experts can analyse the legal documents and contracts, generate the explanations by identifying the obligations and major clauses, create the legal documents and reports based on the case history and also can take the wise decision based on the efficiency and the factual accuracy in the specific case.

FinBERT is pre-trained on a corpus of financial texts, including news articles, reports, and other financial documents. This specialization enables FinBERT to better understand and analyze financial language, jargon, and context that general-purpose models like BERT might not handle as effectively. Furthermore, some models (Multilingual) are representing various languages such as FlauBERT for French [5], BETO for Spanish [15], BERTje for Dutch [17], FinBERT for Finnish [2], BERTimbay for Portuguese [14], RuBERT for Russian [18], and mBERT [16] for multiple languages. Multilingual models are accelerating the effective communication in multiple languages. Each model developed in their specific language to improve the accuracy in various tasks for efficient communication and interaction.

3. METHODS AND METHODOLOGY

3.1 TRANSFORMER

The transformer [1] is the most influential Neural Network (NN) that has shown exceptional performance on various NLP tasks. Transfer Learning, Attention mechanism, and parallelization are the prominent features of the transformers. Transfer Learning is the process of utilizing pre-trained data. Consequently, the models can execute and learn in a faster and efficient manner even with fewer amounts of data. Transfer Learning is the combined procedure of Pre-training and Fine-tuning. The pre-trained data can be later fine-tuned and tailor for other similar tasks in Natural Language Processing (NLP).

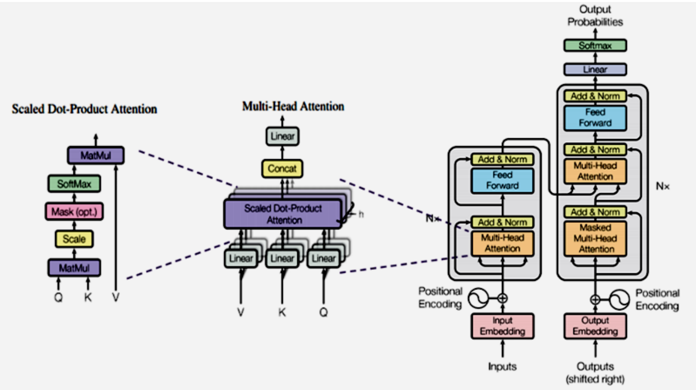


Fig.1. Architecture of Transformers [1]

The attention mechanism is the process of focusing the most relevant word in the context and parallelization is the process of the execution of the NN in a parallel manner. The Fig.1 represents the architecture of the transformer, which comprises other components such as Multi-head Attention, positional encoding, and Linear and softmax layers. Transformer executes based on a six-layered encoder-decoder stack. Encoder extracts the relevant information from the input vectors and passes it to the decoder layers.

3.2 BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model developed by Google [9]. BERT is an outstanding model that utilizes the prominence of transfer learning, which is designed to pre-train bidirectional representations from the unlabeled text. Pre-training brings magic to the model and that facilitates the model for the meticulous understanding of corpus. The model has pre-trained on the massive unlabelled corpus including whole Wikipedia articles and Toronto Book corpus [18] that will be approximate ~16 GB of data. BERT is the first model that has shown a bidirectional feature, which means that the model can learn information from both sides in a context. Pre-trained parameters can be later fine-tuned for other NLP downstream tasks. The BERT is the top of all other traditional models, and the model outperforms all other existing models by breaking new scores on leader boards of various renowned benchmark datasets. BERT is trained with the general Wikipedia text in English. Consequently, the model is not suitable for other types of domain-specific corpus such as clinical, legal, Bio-Medical, financial, Scientific, or any other specific corpus.

Domain-specific models can be generated by training the BERT from the initial stage with the domain-specific data to crack this problem. The newly developed domain-specific models with their vocabulary will be providing significant improvement in language understanding than the generally pre-trained BERT. BERT is a multi-layered Transformer model with encoder-decoder stacks and dramatically accelerated the performance of NLP and NLU Tasks. BERT empowered with unsupervised pre-training and supervised fine-tuning. BERT utilizes the Masked Language model and Next Sentence Prediction to understand the relation between sentences.

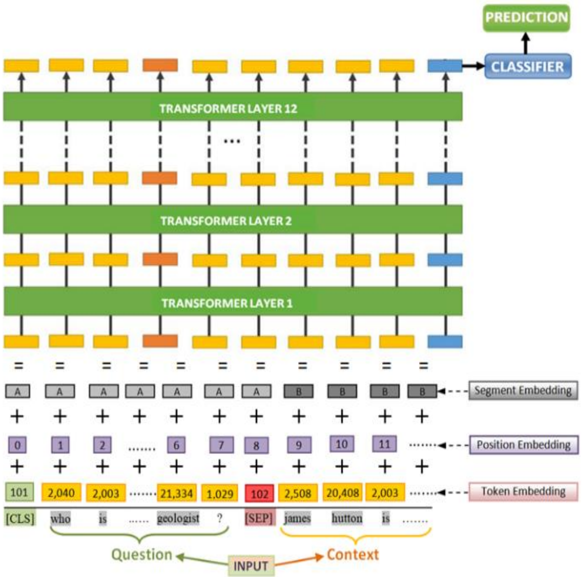


Fig.2. BERT Embedding and Layers [9]

As shown in Fig.2 BERT base model has 12 layers (BERT Large 24 Layers with 340M Parameters) of encoder-decoder stacks with 110 million parameters. These huge amounts of parameters present the incredible performance for the BERT-based models. For MRC Span extraction BERT takes the input context and question. There are three input embeddings in BERT including Token embedding, Position embedding, and Segment embedding. The token embedding converts the text to vectors, position embedding carries the position of each token and the segment embedding facilitates distinguishing between question and context sentences. Finally, predicts the answer followed by the encoder-decoder mechanism.

4. COMPARISON OF BERT AND SCIBERT

SciBERT has taken for the comparison of the BERT and Domain-specific models. The basic difference is in the vocabulary of the models. The basevocab is the tokenizer of the BERT and the scivocab is the tokenizer of the SciBERT and both have approximately ~30000-word vocabularies. Nevertheless, there is 42% overlap tokens between the basevocab and the scivocab, which reveals the substantial difference in the frequency of tokens between the general and domain-specific texts. Fig.3 and Fig.4 represent the BERT Vocabulary token lengths and SciBERT Vocabulary token lengths respectively.

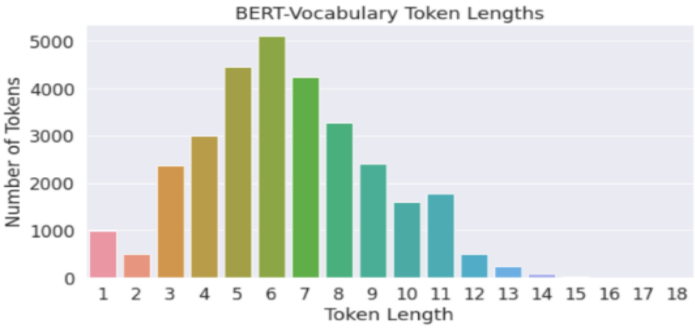


Fig.3. BERT Vocabulary token lengths

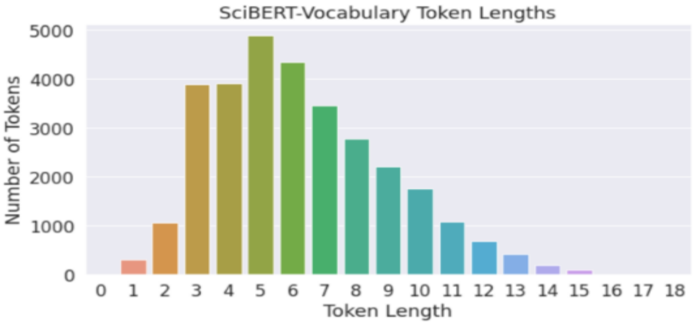
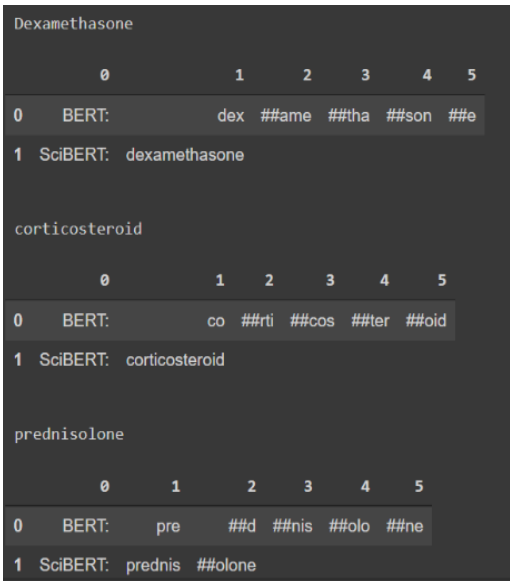


Fig.4. SciBERT Vocabulary token lengths

Fig.5 shows the examples for BERT and SciBERT tokenizations and the tokens are some of the biomedical terminologies. The tokens “dexamethasone”, “corticosteroid” and “prednisolone” are embedded in SciBERT. The tokenizer of SciBERT and BERT will be segmented the “unknown” words into subtokens. So the BERT segmented these words into five subwords each and the ## symbol is used to denote a token as a subword but it will not be included in the first subword. To recapitulate, if the models are not able to distinguish the single token as a word, then the models split the tokens into subwords or tokens.



```
Dexamethasone

          0                 1     2     3     4     5
0   BERT:                 dex ##ame ##tha ##son ##e
1   SciBERT:  dexamethasone

corticosteroid

          0                 1    2     3     4     5
0   BERT:                  co ##rti ##cos ##ter ##oid
1   SciBERT:  corticosteroid

prednisolone

          0         1        2     3     4     5
0   BERT:          pre      ##d ##nis ##olo ##ne
1   SciBERT:  prednis  ##olone
```

Fig.5. BERT and SciBERT Comparison

Furthermore, some of the individual characters, articles and subwords are more frequent than the most common words, such as “a”, “an”, “for, “the”, and models also comprise digits. According to statistics, the BERT has 1,109 tokens or 3.63% of text include a digit and the tokens are whole integers and in the SciBERT has about 3x as many tokens with digits with 3,345 tokens or 10.76% of the text. BERT uses the “WordPiece” model, whereas SciBERT uses the “SentencePiece” model for tokenization, but on the whole, the difference is superficial. The Sub word length comparisons of BERT and SciBERT are depicted in Fig.6 and Fig.7.

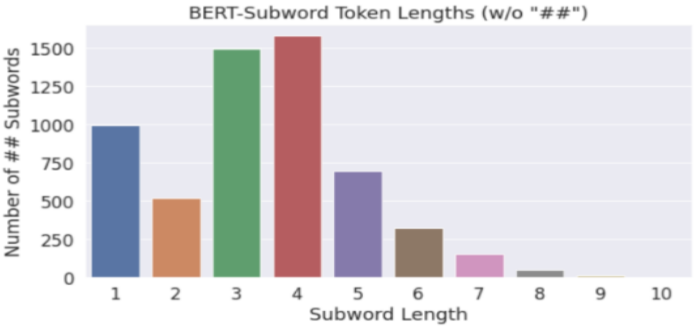


Fig.6. BERT Sub word token lengths

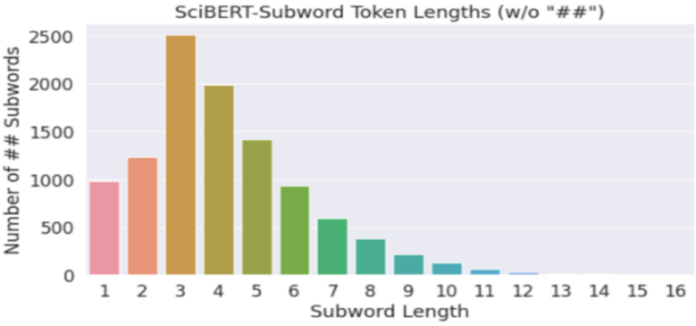
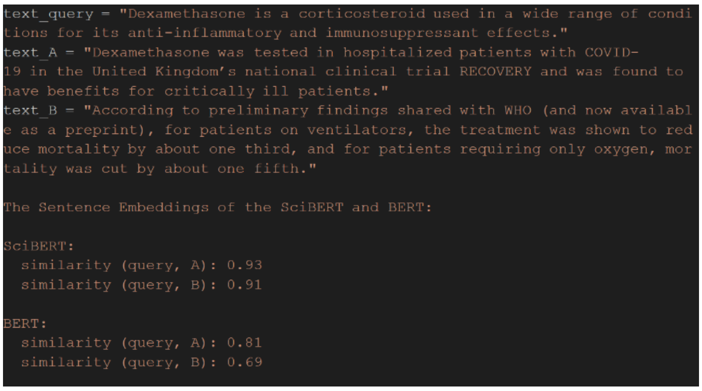


Fig.7. SciBERT Sub word token lengths

Both models SciBERT and BERT models have a maximum token length of 18. The number of subwords in the BERT is 5,828 out of 30,522 tokens, which is 19.1% of the entire tokens and in the SciBERT. The BERT model is pretrained in Wikipedia and Torento book corpus in English language. This represent the inefficiency of the model in capturing the subword recognition, eventhough the model are showing respectable accuracy in general tasks. The number of subwords is 10,491 out of 31,090 tokens, which is 33.7% of the whole tokens. The SciBERT is pre-trained on the biomedical and computer science corpus, due to this reason the SciBERT can identify the biomedical terminologies as a single token (as shown in Fig.5). For the SciBERT, scibert_scivocab_uncased model is used for the empirical evaluation with model type: <class 'transformers.models.bert.modeling_bert.BertModel'> and the tokenizer type: <class 'transformers.models.bert. tokenization_bert_fast.BertTokenizerFast'>. The pre-trained SciBERT model has taken from the AutoTokenizer class of the Transformer using the following code fragments:

scibert_tokenizer = AutoTokenizer.from_pretrained(“allenai/ scibert_scivocab_uncased”) scibert_model = AutoModel.from_pretrained(“allenai/scibert _scivocab_uncased”)



```
text_query = "Dexamethasone is a corticosteroid used in a wide range of condi
tions for its anti-inflammatory and immunosuppressant effects."
text_A = "Dexamethasone was tested in hospitalized patients with COVID-
19 in the United Kingdom’s national clinical trial RECOVERY and was found to
have benefits for critically ill patients."
text_B = "According to preliminary findings shared with WHO (and now availabl
e as a preprint), for patients on ventilators, the treatment was shown to red
uce mortality by about one third, and for patients requiring only oxygen, mor
tality was cut by about one fifth."

The Sentence Embeddings of the SciBERT and BERT:

SciBERT:
  similarity (query, A): 0.93
  similarity (query, B): 0.91

BERT:
  similarity (query, A): 0.81
  similarity (query, B): 0.69
```

Fig.8. Sentence Embedding Similarity of SciBERT and BERT

The Fig.8 shows an empirical estimation of the Scientific Text that shows the embedding comparison of the BERT and SciBERT by employing the Semantic Similarity. There is a text query and two contexts for the comparison they are: text_query, text_A, and text_B. The SciBERT outperforms the BERT in the proximity recognition between the query and the scientific context. SciBERT has shown the similarity of 93% and 91% for text_A and text_B respectively with text_query and BERT has shown the similarity of 81% and 69% for text_A and text_B respectively with text_query.

The empirical analysis of the MRC span extraction, here consider the question as “What is Dexamethasone?” and the text_query (from Fig.8) as the input context. While implementing, the total number of input tokens is 44 in BERT and the total number of input tokens is 28 in SciBERT. This variation of input embedding scores resembles the differences in the BERT and domain-specific models to understand the words or tokens in a context. After tokenization the input context and query pass to the segment embedding to distinguish and extract the features in the context and the question. The model predicted the final answer as “a corticosteroid”, which is in the position of the 16th tensor to the 21st tensor in the token labels as shown in Fig.9.

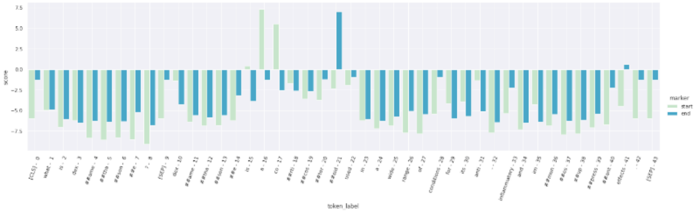


Fig.9. Start and End Scores of the extracted answer in BERT

5. CONCLUSION

This research article focused on the domain-specific transformer models in a machine Reading Comprehension perspective and also incorporates the difference between the General BERT and domain-specific models, Transformer architecture, BERT functionalities, the need for domain-specific models. A comprehensive analysis is included in the article in comparison with BERT and SciBERT with the empirical pieces of evidence. Furthermore, a detailed study on the vocabularies, word length, and semantic similarity is also incorporated with the comparison. Finally, the MRC model takes the input context and question and excerpts the most accurate answer meticulously. The future enhancement of the models is train the model with extensive data in specific domain and reduce the number of parameters by utilizing the Small Language models (SLMs) instead of using the Large Language Models with millions of parameters to improve the accuracy of the model.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is All you Need”, *Proceedings of International Conference on Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.

[2] A. Virtanen, F. Ginter and S. Pyysalo, “Multilingual is not enough: BERT for Finnish”, *Proceedings of International Conference on Advances in Neural Information Processing Systems*, pp.1-14, 2019.

[3] A. Adhikari, A. Ram, R. Tang and J. Lin, “*DocBERT: BERT for Document Classification*”, Springer, 2019.

[4] D.T. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, *Proceedings of International Conference on Artificial Intelligence*, pp.1-11, 2019.

[5] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, L. Besacier and D. Schwab, “FlauBERT: Unsupervised Language Model Pre-training for French”, *Proceedings of International Conference on Artificial Intelligence*, pp.1-12, 2020.

[6] I. Beltagy, K. Lo and A. Cohan, “SCI BERT: A Pretrained Language Model for Scientific Text”, *Proceedings of International Conference on Natural Language Processing*, pp. 1-10, 2019.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras and I. Androutsopoulos, “*LEGAL-BERT: The Muppets Straight Out of Law School*”, Wiley Publisher, 2020.

[8] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of International Conference on Computation and Language*, pp. 1-8, 2019.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim and J. Kang, “Bio-BERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining”, *Bioinformatics*, Vol. 78, pp. 1-7, 2019.

[10] Jieh Sheng Lee and Jieh Hsiang, “*Patent Classification by Fine-Tuning BERT Language Model*”, Ios Press, 2019.

[11] K. Huang, J. Altosaar and R. Ranganath, “Clinical BERT: Modeling Clinical Notes and Predicting Hospital Readmission”, *Proceedings of International Conference on Computation and Language*, pp. 1-14, 2020.

[12] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine and M. Boeker, “GottBERT: A Pure German Language Model”, *Proceedings of International Conference on Computation and Language*, pp. 123-132, 2020.

[13] R.C. Rodrigues and Anderson da Silva Soares, “Multilingual Transformer Ensembles for Portuguese Natural Language Tasks”, *Proceedings of International Conference on Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, pp. 1-6, 2020.

[14] S. Wu and M. Dredze, “Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT”, *Proceedings of International Conference on Computation and Language*, pp. 1-18, 2019.

[15] T. Pires and E. Schlinger Dan Garrette, “How Multilingual is Multilingual BERT?”, *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 1491-1493, 2019.

[16] W. Vries, A.V. Cranenburgh, A. Bisazza, T. Caselli, G.V Noord and M Nissim, “*BERTje: A Dutch BERT Model*”, *Proceedings of International Conference on Computation and Language*, pp. 81-98, 2019.

[17] Y. Zhu and S. Fidler, “Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books”, Available at https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Zhu_Aligning_Books_and_ICCV_2015_paper.pdf, Accessed in 2015.

[18] Y. Kuratov and M. Arkhipov, “Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language”, *Proceedings of International Conference on Computation and Language*, pp. 85-94, 2019.