Bert Based Named Entity Recognition for the Albanian Language

Labehat Kryeziu[1]

Visar Shehu[2]

[1]Ph.D. Candidate, Faculty of Contemporary Sciences and Technologies
South East European University, Tetovo, North Macedonia
lk29054@seeu.edu.mk
[2]Full professor, Faculty of Contemporary Sciences and Technologies
South East European University, Tetovo, North Macedonia
v.shehu@seeu.edu.mk

Received: 21 September 2023 / Accepted: 29 October 2023 / Published: 23 November 2023

Doi: 10.56345/ijrdv10n302

Abstract

In this paper, we explore the application of the Bidirectional Encoder Representations from Transformers (BERT) model for Named Entity Recognition (NER) in the Albanian language. Despite the success of BERT in various NLP tasks across numerous languages, Albanian remains under-represented. Our approach leverages a transfer learning strategy with multilingual BERT, fine-tuned on a newly created, comprehensive Albanian language corpus. The corpus includes texts from various domains, thereby facilitating the identification of a broad range of named entities. We report on the challenges faced during corpus creation, model fine-tuning, and testing phases, such as dealing with dialectal variations and lack of existing resources for the Albanian language. This research not only contributes to the advancement of NER systems for low-resource languages, but also provides a robust foundation for further advancements in Albanian language processing.

***Keywords:** NER, BERT, Transfer Learning, Low-resource Language, NLP*

1. Introduction

Presently, the task of representing text within machines poses an increasingly significant difficulty, particularly for languages with limited available resources. However, this concern is now gaining traction and attracting the attention of numerous researchers. In this context, transformers have proven to be highly beneficial. These models rely on neural networks and exhibit promising outcomes when utilized for tasks involving sequential input and output [1]. Nonetheless, past endeavors in this domain have largely fallen short, leading to text comprehension systems confined to particular domains they were initially tailored for. These systems frequently struggled with novel or unfamiliar inputs, and their inflexibility posed challenges in grasping the intricacies of human language. Consequently, the realm of natural language processing is in a constant state of progression, marked by frequent and substantial shifts in neural network structures [2]. Recognizing that language serves as a primary conduit for human expression, it is imperative for natural language processing to attain a suitable level of human intelligence-like representation.

As indicated in the paper [2], within the last five years, deep neural networks have risen to prominence as a significant subset of machine learning models. They find extensive application in situations requiring data-driven learning. These models autonomously glean insights from layered hierarchical arrangements, reducing the necessity for manual feature crafting. As a result, human endeavor primarily revolves around choosing the suitable architecture for a specific task and training context.

Notably, pre-trained word embeddings were among the initial breakthroughs in transfer learning for natural language processing, as reported in [3]. Pre-trained models such as ELMo [4] and BERT [5] are trained on extensive collections of unlabeled text data, leading to superior performance across a range of tasks and datasets from various domains.

The utilization of transformers, exemplified by Bidirectional Encoder Representations by Transformers (BERT), stands as a widely acknowledged technique for acquiring insights into text portrayal. These models founded on transformers have exhibited remarkable accomplishments in addressing sequence-to-sequence challenges. Noteworthy achievements in natural language processing (NLP) have been demonstrated across various languages, encompassing English, French, German, and Chinese, owing to the abundance of extensive digital textual datasets [2]. Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language representation model, introduced by the Google AI team in 2018 [5]. The creators recognized that prior models could only be trained in a unidirectional manner, constituting a significant drawback in past language modeling approaches. This limitation implied that a token's encoding relied on information from either its left or right neighbors, without simultaneous integration of data from both tokens. In contrast to alternative language representation models, BERT generates deep bidirectional representations from input text. Bidirectional Encoder Representations by Transformers (BERT) has undergone training for diverse tasks like text classification and question answering, showcasing encouraging outcomes and fostering promise within the realm of Natural Language Processing [6].

Since 1990, significant transformations have unfolded within NLP (Natural Language Processing) tools, including enhanced data accessibility. Despite a global array of 7000 languages, the majority of NLP research has centered around only 20 of them, leaving an extensive array of languages with scant attention or none at all [7]. These languages, which have received limited or negligible scrutiny in the realm of Natural Language Processing, are referred to as Low Resource Languages (LRLs). LRLs signify languages with reduced study, limited computerization, fewer available resources, less proficiency, and other comparable attributes, as outlined in papers [7], [8], and [9]. The term LRL denotes languages lacking adequate data for the application of statistical methods. Among these languages is Albanian, which falls into the LRL category, necessitating persistent research in this direction. Despite the achievements of BERT across diverse languages in various NLP tasks, the Albanian language remains inadequately represented. Our objective is to positively impact NLP implementation, specifically advancing Named Entity Recognition (NER) systems within the context of Albanian. A member of the Indo-European language family, Albanian is recognized for its intricate morphological structure. Its alphabet comprises 36 letters, encompassing 7 vowels and 29 consonants [10]. Hence, the core focus of this paper revolves around addressing challenges encountered in implementing NER for the Albanian language.

2. Related Work

Named Entity Recognition (NER) is a branch of natural language processing (NLP) aimed at extracting vital details from written content. NER's main goal is to pinpoint and categorize specific data within a text, known as named entities. Such entities include topics like names, companies, products, percentages, financial figures, places, dates, and so forth. Basically, NER works by spotting, separating, and pulling out key information from unorganized text, eliminating the need for manual analysis. This approach is especially helpful in quickly drawing out crucial details from large data sets, streamlining the data extraction process and saving significant time. NER comes into play when there's a need to pull out main data points, or "entities", and then categorize each one found in the text. To give you a clearer picture, here's an example showcasing how NER works using the Albanian language:

Apple do të lansoj iPhonin e ri në Shtator me çmim prej 1050 dollar

The entities that are extracted from this information are:

Apple $_{\text{ORGANIZATION}}$ do të lansoj iPhonin e ri në Shtator $_{\text{DATE}}$ me çmim prej 1050 dollar $_{\text{MONETARY VALUE}}$

Apple – ORGANIZATION

Shtator – DATE

1050 dollar - MONETARY VALUE

Ongoing research continues to demonstrate advancements in this field. In certain instances, authors have achieved superior results using BERTje in comparison to multilingual BERT for various NLP tasks like part-of-speech tagging and Named Entity Recognition (NER) [11]. Since the initial introduction of BERT, several models have emerged, drawing inspiration from it. The first three models, namely BERT-base, BERT-large, and multilingual BERT, were closely influenced by BERT. Another model, AraBERT (Arabic Biomedical Bidirectional Encoding Representation Transformer), is tailored for Arabic language. Arabic is linguistically intricate but lacks the exploration and resources comparable to English in the realm of natural language processing. Consequently, conducting common NLP tasks such as Sentiment Analysis and Question Answering in Arabic proves to be immensely challenging. According to the insights from [12], AraBERT retains the core BERT architecture but outperforms BERT when applied to Arabic. AraBERT's fine-tuning phase encompasses Named Entity Recognition, Question Answering, and Sequence Classification [13], [14]. Even the Albanian language, characterized by rich morphology, holds promise for achieving notable outcomes in Named Entity Recognition (NER) applications.

Citing [15], Named Entity Recognition (NER) involves identifying named entities (NEs) and categorizing them into predefined classes like organization, location, or person. Leading NER systems employ neural architectures pre-trained for language modeling. BERT [5] is a well-known model for language tasks, which we'll use in this study. Albert [16], RoBERTa [17], and ELMo [4] share similar principles. Despite its apparent simplicity, NER is complex due to context-dependent categories and varying definitions across languages [15].

Initially, neural network NER systems gained popularity for minimal feature engineering and enhanced domain confidentiality [18]. CharWNN [19] extended Collobert et al. [20] work by using convolutional layers to extract unique character-level features for words. These features combined with pre-trained word embeddings for classification. Other studies explored contextual embeddings from linguistic models with LSTM-CRF architecture. Santos et al. [21], harnessed Flair Embeddings [22] for contextual word embeddings from bidirectional LMs in Portuguese corpora. Embeddings combined with pre-trained characters were fed into a BiLSTM-CRF model. Castro et al. [23] employed ELMo embeddings that fuse word-level features from CNNs and layers from a bidirectional LM (BiLM) composed of a BiLSTM model.

As noted in the article [12], several NLP tasks grounded in machine learning often grapple with insufficient training data, particularly when concerning low resource languages (LRLs) like Arabic. This inadequacy hampers effective model generalization. To address this, transfer learning emerges as a solution by leveraging knowledge from prior domains or tasks. Transfer learning aims to extract insights from one or multiple tasks, which can then be applied to future tasks. According to [5], contextual language models exhibit notable performance in various NLP tasks. Among these models, Bidirectional Encoder Representation by Transformer (BERT) stands out as a fine-tuned contextual language model. At its core lies a multi-layer bidirectional transformer encoder that takes both left and right context into account within a sentence. The study by [24] scrutinized several deep learning models, particularly within the biomedical NER domain, demonstrating an advantage for two-way models over one-way ones across various architectures. Word embedding models have been extensively explored in NER research. Works like [25], [26], [27] achieved success with word2vec, including Skip-gram and CBOW models. In [28], [29], pre-trained models with BLSTM-CRF were employed. Fasttext saw application in Japanese NER by [30] and [31]. Given BERT's proficiency as a contextual language model [5] and its success in NLP tasks [32], numerous deep learning-based architectures have been introduced to adapt BERT for NER. The paper [50] underscores the effectiveness of adding a BLSTM-CRF layer atop pre-tuned BERT models, outperforming standalone fine-tuned BERT.

Recently, an upsurge in research endeavors by numerous scholars has centered on delving into Natural Language Processing (NLP) for the Albanian language, particularly in terms of tools and applications. These studies encompass part-of-speech and morphological labeling [33], [34], the Albanian lexicon for NLP [35], and more, signifying various attempts at crafting NLP tools for Albanian. In a study by the authors of [33], a basic morphological tagger was introduced for standard Albanian. This served as an initial building block within the Albanian Corpus Initiative, with encouraging results from the evaluation of the tagger on a 1,000-word corpus. Subsequently, dictionaries and transducers were devised to automatically process Albanian, drawing from [36] and [37]. These works elucidated distinct features of the Albanian language and elucidated how FST and NooJ graphs facilitate their manipulation within linear text segments.

Initial steps in sentiment analysis for Albanian were taken in [38], where a machine learning model was constructed to classify text documents into positive or negative opinions on specific topics. To train this model, a 400-document corpus comprising political news and diversified topics was created. Each topic comprised 80 documents classified as positive or negative, totaling five topics.

Further progress emerged with the introduction of the name entity recognition (NER) model for Albanian in [39]. Employing a high entropy strategy via the Apache OpenNLP tool, they ventured into NER. For evaluation, they compiled an annotated corpus with historical and political text. Authors in [40] proposed their NER approach, incorporating deep learning models, notably a deep neural network employing LSTM and CRF. Both word and character labeling were leveraged. The paper's outcomes underscored the potential for refining NER efficiency through larger training corpora.

These exemplify select endeavors within the realm of NLP for Albanian. The imperative is clear: an emphasis on Albanian integration into NLP is vital. The application of diverse NLP tasks, specifically advancements in NER systems, holds promise for this language's evolution.

3. Methodology

For the utilization of the BERT model in Named Entity Recognition (NER) for the Albanian language, we employed our custom dataset (alb_dataset) from 2022 [10]. This dataset comprises a vast collection of texts, gathered using a web crawler that autonomously sourced content from Wikipedia, Assembly of Albania records, and assorted reports from the ministries of Kosovo, Albania, and North Macedonia. Authored by various writers, these texts are in standardized Albanian, and we've ensured to include only those with a minimal likelihood of grammatical and spelling mistakes.

For the pre-training data, we utilized the corpus (alb_dataset) that is 950 MB in size and contains a total of 149,473,629 words. Beyond its volume, the alb_dataset is composed of complete documents. Its methodological approach guarantees a broad domain variety and superior content quality—traits that are beneficial for BERT pre-training and the application of Named Entity Recognition (NER) in the Albanian language.

3.1 Training BERT model

In our study, we employed the HuggingFace library for model training [41] and used Google Collab with TPU-backed virtual machines. The model we refined is based on "bert-base-multilingual-cased". We leveraged the HuggingFace Transformers and API pipeline to load models, then implemented a BERT model named "mbert-base-albanian-cased-ner", which was tailored using the alb_dataset for the Named Entity Recognition (NER) task in the Albanian language.
In the following, we will mention some of the stages of the application of the model of Bidirectional Encoder Representations from Transformers (BERT) for Named Entity Recognition (NER) in the Albanian language:

- Split the dataset into manageable files for the beginning of the application of NER
- Using a model from HuggingFace for training
- Using the spaCy to visualize Python dictionary
- Extracting entities for our text using that model

Next, we extract the entities for our text using the model mentioned above, and as output we get:

Table 1: Extracting entities

[{'entity': 'B-LOC', 'score': 0.99528414, 'index': 8, 'word': '**Kosova**', 'start': 21, 'end': 27},	{'entity': 'I-LOC', 'score': 0.99677736, 'index': 25, 'word': '**Ball**', 'start': 87, 'end': 91},	{'entity': 'I-LOC', 'score': 0.996743, 'index': 26, 'word': '##**kan**', 'start': 91, 'end': 94},
{'entity': 'I-LOC', 'score': 0.9964747, 'index': 27, 'word': '##**ik**', 'start': 94, 'end': 96},	{'entity': 'B-LOC', 'score': 0.99915445, 'index': 40, 'word': '**Shqipëri**', 'start': 132, 'end': 140},	{'entity': 'B-LOC', 'score': 0.99719185, 'index': 49, 'word': '**Mali**', 'start': 162, 'end': 166},
{'entity': 'B-PER', 'score': 0.99911326, 'index': 208, 'word': '**Has**', 'start': 657, 'end': 660},	{'entity': 'B-PER', 'score': 0.9989893, 'index': 209, 'word': '##**him**', 'start': 660, 'end': 663},	{'entity': 'I-PER', 'score': 0.9979571, 'index': 210, 'word': '**Tha**', 'start': 664, 'end': 667}, …etc

As illustrated in table 1, the output comprises a list of dictionaries detailing the start and end locations of the entity within the text, the prediction score, the word in question, its index, and the name of the entity.

The designated entities in this dataset include:

- I-LOC: Denotes a location.

- B-LOC: Marks the start of a new location immediately following a different location.
- I-ORG: Represents an organization.
- B-ORG: Signals the start of a new organization immediately after a different one.
- I-PER: Refers to a person's name.
- B-PER: Indicates the beginning of one person's name immediately following another person's name.

4. Results

As detailed in the preceding section, we successfully extracted named entities from our dataset. Following this extraction, we implemented a function leveraging spaCy to display this Python dictionary visually.

We employed the spacy.displacy.render() function to showcase the text associated with a named entity, with the manual=True parameter indicating a hands-on visualization. Furthermore, we set jupyter to True given our use of a Jupyter notebook or Colab. The primary aim of the for loop is to form a list of dictionaries specifying the start and end points, along with the label of the entity. Additionally, we assess if similar entities are in proximity and, if so, merge them, executing:

# get HTML representation of NER of our text
get_entities_html(text, doc_ner)
resulting with:

Figure 1: Albanina NER results from alb_dataset

Next, we condense these findings and display them in table 2. The annotations are structured in the Inside Outside (IO) format, encompassing the six tags previously elaborated upon:

Table 2: NER tag and example

NER tag	Example
I-LOC	Ballkanik
B-LOC	Kosova
B-ORG	Kombeve
I-ORG	të Bashkuara
O-PER	Hashim
I-PER	Thaçi

5. Conclusions

This study explores different authors' opinions on including the Albanian language in the Natural Language Processing (NLP) field, specifically in Named Entity Recognition (NER) systems. The paper contributes to advancing NER systems for low-resourced languages (LRL) and offers a clear overview of the progress made so far, with a focus on Albanian.
To train and build our models we use Google Colab, with which we achieve acceptable results for smaller datasets. However, Google Colab, functioning as a cloud platform, provides just 25.5 GB of RAM to its subscribed users (only 12 GB to non-paying users).

To achieve our goals for applying NER systems in the Albanian language, we utilize the HuggingFace library for model training [41], the model we use is based on "bert-base-multilingual-cased", while the specific model we implemented to obtain the results was "mbert-base-albanian-cased-ner", which was fine-tuned with the alb_dataset for the Named Entity Recognition (NER) task in the Albanian language. From all this we achieve very good results regarding the application of NER systems in our dataset. The entities extracted from the dataset were very accurate in terms of grouping into 6 labels (I-LOC, B-LOC, I-ORG, B-ORG, O-PER, O-PER) with a high level of alignment between NER tags and extracted entities. This NER system applied to the Albanian language managed to accurately distinguish persons, organizations, locations, etc.

References

A. A. Darwish, K. Durrani, and H. Mubarak," Farasa: A fast and furious segmenter for arabic". In Proceedings, pp. 11–16, 2016.

A. K. Singh, "Natural language processing for less privileged languages: Where do we come from? where are we going?" In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, 2008.

A. Mageed, M. Alhuzali and M. Elaraby, "You tweet what you speak: A citylevel dataset of arabic dialects" In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638– 1649.

B. Kabashi and T. Proisl, "A proosal for a part-of-speech tagset for the Albanian language," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 4305–4310.

B. Kabashi, "A lexicon of albanian for natural language processing," The XVIII EURALEX International Congress, pp. 855–862, 2019.

Ch. Cieri, M. Maxwell, S. Strassel, and J. Tracey, "Selection criteria for low resource language programs", In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4543– 4549, 2016.

Cicero Nogueira dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. Computing Research Repository, arXiv:1505.05008. Version 2.

D. C. Wintaka, M. A. Bijaksana, and I. Asror, ``Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF,'' Procedia Comput. Sci., vol. 157, pp. 221-228, Jan. 2019.

Delvin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

E. Trandafili, E. K. Meçe, and E. Duka, "A named entity recognition approach for Albanian using deep learning," in Complex Pattern Mining, 2020, pp. 85–101.

H. Fei,Y. Ren, and D. Ji, ``Dispatched attention with multi-task learning for nested mention recognition,'' Inf. Sci., vol. 513, pp. 241-251, Mar. 2020.

H. Huggard, A. Zhang, E. Zhang, and Y. S. Koh, ``Feature importance for biomedical named entity recognition,'' in Advances in Artifcial Intelli-gence. Cham, Switzerland: Springer, 2019, pp. 406-417.

Hugging Face (2023). ("akdeniz27/mbert-base-albanian-cased-ner"). Retrieved from https://huggingface.co/akdeniz27/mbert-base-albanian-cased-ner?fbclid=IwAR2mH4KvTS1Jtw3Ts1_SyIiMVIl-SbvKqr7A-zn8vLetP27OIgEtqfYGV9g

J. Li, S. Zhao, J. Yang, Z. Huang, B. Liu, S. Chen, H. Pan, and Q. Wang, ``WCP-RNN: A novel RNN-based approach for bio-NER in Chinese EMRs,'' J. Supercomput., vol. 76, no. 3, pp. 1450 1467, Mar. 2020.

J. Trommer and D. Kallulli, "A morphological tagger for standard Albanian," in Proceedings of LREC, 2004, pp. 1-8.

Joaquim Santos, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini, and Renata Vieira. 2019a. Assessing the impact of contextual embeddings for portuguese named entity recognition. In 8th Brazilian Conference on Intelligent Systems, BRACIS, Bahia, Brazil, October 15-18, pages 437– 442.

Joseph, T., Lev, R., & Yoshua, B. (2010). Word representations: A simple and general method for semi-supervised learning. 48th Annual Meeting of the Association for Computational Linguistics.

Kryeziu, L., Shehu, V., & Caushi, A. (2022). Evaluation and Verification of NLP Datasets for the Albanian Language. International Conference on Artificial Intelligence of Things. Istanbul, Turkey.

L. Kryeziu and V. Shehu, "A Survey of Using Unsupervised Learning Techniques in Building Masked Language Models for Low Resource Languages", in 11th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2022.

L. Kryeziu and V. Shehu, "PRE-TRAINING MLM USING BERT FOR THE ALBANIAN LANGUAGE", in SEEU Review, North Macedonia, 2023.

M. Biba and M. Mane, "Sentiment analysis through machine learning: An experimental evaluation for Albanian," in Recent Advances in Intelligent Informatics, 2014, pp. 195–203.

M. P. Skenduli and M. Biba, "A named entity recognition approach for Albanian," in 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Aug. 2013, pp. 1532–1537, doi: 10.1109/ICACCI.2013.6637407.

M. Xiaofeng, W. Wei, and X. Aiping, ``Incorporating token-level dictionary feature into neural model for named entity recognition,'' Neurocom- puting, vol. 375, pp. 43-50, Jan. 2020.

Matthew, E., Mark, N., Mohit, J., Matt, G., Christopher, C., Kenton, L., & Luke, Z. (2018). Deep contextualized word representations.

Monica Marrero, Juli´an Urbano, Sonia S´anchez- Cuadrado, Jorge Morato, and Juan Miguel G´omez- Berb´ıs. 2013. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5):482–489.

O. Piton and K. Lagji, “Morphological study of Albanian words, and processing with NooJ,” arXiv preprint arXiv:1002.0485, Feb. 2010.

O. Piton, K. Lagji, and R. Përnaska, “Electronic dictionaries and transducers for automatic processing of the Albanian language,” in Natural Language Processing and Information Systems, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 407–413.

Pedro Castro, Nadia Felix, and Anderson Soares. 2019. Contextual representations and semi-supervised named entity recognition for portuguese language.

Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of machine learning research, 12(Aug):2493–2537.

S. Long, R. Yuan, L. Yi, and L. Xue, ``A method of Chinese named entity recognition based on CNN-BILSTM-CRF model,'' in Data Science (Communications in Computer and Information Science). Singapore: Springer, 2018, pp. 161-175.

S. Misawa, M. Taniguchi, Y. Miura, and T. Ohkuma, ``Character-based bidirectional LSTM-CRF with words and characters for Japanese named entity recognition,'' in Proc. 1st Workshop Subword Character Level Models NLP, Copenhagen, Denmark, 2017, pp. 97-102.

Tjong Kim Sang, E. F,” Introduction to the CoNLL-2002 Shared Task: Language- Independent Named Entity Recognition”, In COLING-02: The 6th Conference on Natural Lan-guage Learning 2002 (CoNLL-2002), 2002.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2145–2158.

W. Antoun, F. Baly, H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding", 2020.

W. Zhang, S. Jiang, S. Zhao, K. Hou, Y. Liu, and L. Zhang, ``A BERTBiLSTM- CRF model for Chinese electronic medical records named entity recognition,'' in Proc. 12th Int. Conf. Intell. Comput. Technol. Automat. (ICICTA), Xiangtan, China, Oct. 2019, pp. 166-169.

Y. Jin, J. Xie, W. Guo, C. Luo, D. Wu, and R. Wang, ``LSTM-CRF neural network with gated self attention for Chinese NER,'' IEEE Access, vol. 7, pp. 136694-136703, 2019, doi: 10.1109/ACCESS.2019.2942433.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Y. Tsvetkov, “Opportunities and challenges in working with low-resource languages”, Carnegie Mellon University, 2017.

Z. Yuwen, and X. Zhaozhuo, “Bert for question answering on squad 2.0”, 2018.

Zh, Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942