Proceedings of International Conference on Educational Discoveries and Humanities
Hosted online from Plano, Texas, USA.
Date: 1st August - 2024
ISSN: 2835-3196 **Website:** econferenceseries.com

"ANALYSIS AND RETRAINING OF THE BERT MODEL FOR THE UZBEK LANGUAGE: METHODS AND RESULTS"

Rajabov J.
National University of Uzbekistan, Tashkent, Uzbekistan
j.rajabov@nuu.uz

Abstract

This paper discusses the use of the BERT model for processing texts in the Uzbek language. BERT (Bidirectional Encoder Representations from Transformers), being one of the most advanced models in the field of natural language processing (NLP), demonstrates high efficiency when working with various languages. The study analyzes the main aspects of adapting BERT for the Uzbek language, including data collection and preparation, model training and evaluation of its performance.

Introduction

BERT, developed by Google AI, is a deep, bidirectional model trained on a large volume of text. Its ability to take into account the context of words on both the left and right makes it especially effective for natural language processing tasks [1]. Applying BERT to the Uzbek language is important for several reasons. First, the Uzbek language has a unique morphological structure, which poses special challenges for NLP [2]. Secondly, the lack of large text corpora in the Uzbek language complicates training and evaluation of models.

Data preparation

To successfully use BERT in the Uzbek language, it is necessary to create an extensive corpus of texts [3]. This study used texts from news sites, social networks and books. This data was denoised, tokenized and normalized. The data preparation process included:

- Collection of texts from open sources.
- Removing duplicates and irrelevant information.
- Tokenization taking into account the peculiarities of the Uzbek language.
- Lemmatization to reduce words to their base form.

Adaptation of the BERT model

BERT requires additional training on language-specific data [4]. The model adaptation process included the following stages:

- Pre-training a model on an Uzbek text corpus.
- Using data augmentation techniques such as text augmentation and translation from other languages.
- Additional training on problems of classification and extraction of named entities.

Pre-training formula:

$$L = -\sum_{i=1}^{N} [y_i \log(p_i) + (1 + y_i) \log(1 - p_i)]$$

where L is the loss function, N is the number of training examples, y_i is the true label, p_i is the predicted probability.

Application examples

The BERT model, adapted for the Uzbek language, can be used in various NLP tasks, such as:

- *Text classification:* The model can identify the topics of news articles with high accuracy.

Example:

Input text: "Uzbekistan national team won the Asian football championship"("O‘zbekiston terma jamoasi futbol bo‘yicha Osiyo chipionligini qo‘lga kiritdi").

Classification: "Sport"

- *Named Entity Extraction:* The model can extract names, dates, and other important entities in text [5].

Example:

Input text: "President Shavkat Mirziyoyev visited Tashkent on July 20, 2023" (" Prezident Shavkat Mirziyoyev 2023-yil 20-iyul kuni Toshkentga tashrif buyurdi").

Extracted Entities: ["Shavkat Mirziyoyev" (Person), " Toshkent" (Location), " July 20, 2023" (Date)]

- *Text translation*: improving the quality of machine translation using the BERT model.

Performance Evaluation

The model's performance was assessed using several metrics, including precision, recall, and F1-score. The results show that the adapted BERT model performs well on the test data.

F1-measure formula:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The graph below shows the improvement in text classification performance compared to the baseline model.

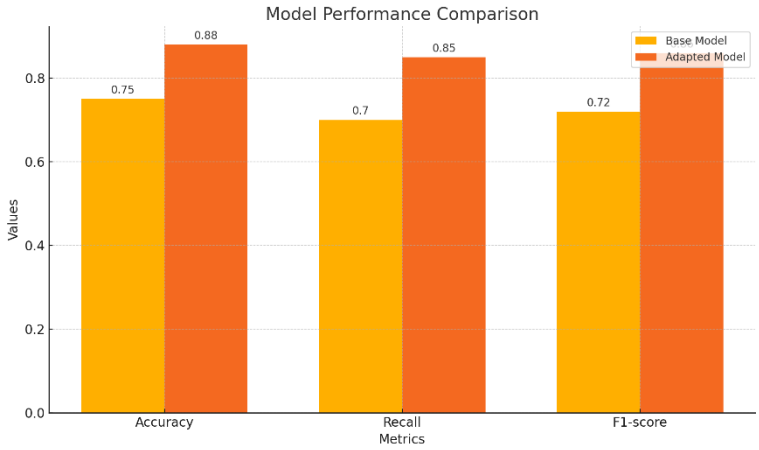

Image 1-Comparison of Model Performance Metrics

Conclusion

Using the BERT model for text processing in the Uzbek language opens up new opportunities for automating NLP tasks. Despite the challenges associated with the morphological features of the Uzbek language and the limited amount of available data, the adapted model demonstrates high efficiency. Future research may be aimed at further improving the model and expanding its application areas, such as machine translation, sentiment analysis and other tasks.

List of used literature

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
3. Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12).
4. Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 328-339).
5. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.