www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 9, No: 4, 2022 (1193-1200)

El-Cezerî Journal of Science and Engineering
Vol: 9, No: 4, 2022 (1193-1200)
DOI : 10.31202/ecjse.1131826

ECJSE

Research Paper / Makale

Keyword Extraction from Kazakh News Dataset with BERT

Aiman ABİBULLAYEVA, Aydın ÇETİN

Computer Engineering Department, Gazi University, Ankara, 06560, TURKEY
acetin@gazi.edu.tr

Received/Geliş: 16.06.2022 **Accepted/Kabul:** 07.09.2022

Abstract: Keywords provide a concise and precise description of the document's content. Due to the importance of the keyword and the difficulty of manual markup, automatic keyword extraction makes this process easy and fast. In this paper, Keyword Extraction from Kazakh News Dataset was presented. Model performance results were obtained by using the BERT base - uncased and BERT-base-multilingual-uncased pre-trained language model for the newly compiled Kazakh News Dataset-KND. Compiled Kazakh news data set consists of 7060 data. Data were collected from the web pages anatili.kazgazeta.kz, Bilimdinews.kz, and zhasalash.kz using the BeautifulSoap and Requests libraries. These web pages mostly contain news, history, and literary texts. The dataset includes the publication name or news title, the author of the publication or news subject, and the URL of the Kazakh news site. In the evaluation of the training results, it was observed that the BERT base-multilingual-uncased F-score performance was higher than the BERT model.

Keywords: Kazakh language, keyword extraction, natural language processing, BERT.

BERT ile Kazak Haber Veri Kümesinden Anahtar Kelime Çıkarımı

Öz: Anahtar kelimeler, belgenin içeriğinin kısa ve kesin bir tanımını sağlar. Anahtar kelimenin önemi ve manuel işaretlemenin zorluğu nedeniyle, otomatik anahtar kelime çıkarımı bu işlemi kolay ve hızlı hale getirmektedir. Bu makalede Kazak haber veri setinden anahtar kelime çıkarımı sunulmaktadır. Yeni derlenen Kazak Haber Veri seti için BERT ve BERT-Base-Multilingual-Uncased önceden eğitilmiş dil modeli kullanılarak model performans sonuçları elde edilmiştir. Derlenen Kazak haber veri seti 7060 veriden oluşmaktadır. Veriler BeautifulSoap ve Requests kütüphaneleri kullanılarak aikyn.kz, anatili.kazgazeta.kz, zhasalash.kz ve baq.kz web sayfalarından toplanmıştır. Bu web sayfaları çoğunlukla haber, tarih, edebiyat metinlerini içermektedir. Veri seti yayın adını veya haber başlığını, yayının veya haberin konusunu ve Kazak haber sitesindeki URL'yi içermektedir. Eğitim sonuçları değerlendirildiğinde, BERT base-multilingual-uncased F-score başarımının BERT base - uncased modeline oranla daha yüksek olduğu gözlenmiştir.

Anahtar Kelimeler: Kazak dili, anahtar kelime çıkarımı, doğal dil işleme, BERT.

1. Introduction

Scanning information on the Internet in daily life has become a common activity for computer users. With each passing day, it is getting harder and harder to reach the desired information in the crowd of information. Since thousands of internet news are published on the internet every day, it is difficult to obtain and summarize the relevant document effectively. Therefore, keyword or keyword extraction technique is used to provide the main content of a particular web page or document. Keywords are the most obvious words or phrases that may represent information. Keywords and phrases are often understood as structural units of the text that contain the most important information about the content of the text [1].

How to cite this article
Abibullayeva A., Çetin A. "Keyword extraction from Kazakh news dataset with BERT" ElCezerî Journal of Science and Engineering, 2022, 9(4); 1193-1200

Bu makaleye atıf yapmak için
Abibullayeva A., Çetin A. "BERT ile Kazak haber veri kümesinden anahtar kelime çıkarımı" El-Cezerî Fen ve Mühendislik Dergisi, 2022, 9(4); 1193-1200.
ORCID ID: 0000-0003-1800-7350, 0000-0002-8669-823X

Machine learning, deep learning methods, and natural language processing studies for the Kazakh language are limited. Keyword extraction for Kazakh documents is an unexplored topic and there are very few relevant publications. However, there are many scientific studies in the comparative-typological direction, in which the structural system of the Kazakh language is compared with other languages, showing personal similarities and differences.

The Kazakh language is one of the languages of low origin and belongs to the agglutinative language group. In the history of writing, the alphabet system of the Kazakh language has passed through various historical periods and reached the level of the national alphabet. It is known that Kazakh people have used the alphabet system based on Arabic graphics for centuries.

From 1929 to 1940, the Latin-based alphabet was included in the writing system, and since 1940 the Cyrillic alphabet has been used. In 2017, the new Latin alphabet of the Kazakh language was approved by the decree of the President of the Republic of Kazakhstan on October 26. It is planned to switch to a new alphabet in 2017-2025. Currently, the issue of transitioning from Cyrillic to Latin is widely discussed in society. The transition to the Latin alphabet, which has become the language of all advanced technologies, is a great spiritual phenomenon for our country, art, and culture. The introduction of the Latin alphabet is very important if we want the Kazakh language to rise from the world civilization to the international level. Developed countries in the world have now switched to the Latin alphabet. The transition of Kazakhstan to the Latin alphabet is of great importance both socio-economically and politically.

The most important problem of alphabet change is the forgetting of the old heritage. When the alphabet is changed, the texts written in that alphabet are deleted. For example, people today cannot read works in the old Turkish or yellow Uighur alphabet. In the case of switching from the Cyrillic alphabet to the Latin alphabet, keyword extraction will gain importance for document linking. When the keywords of the documents are removed, access to the document will be easier. Classification and clustering studies in Kazakh have just begun, and there is not a large enough corpus for the Kazakh language, so data were collected from 7000 texts from Kazakh news sites on the internet. Scanned from web pages written in Kazakh: aikyn.kz, zhasalash.kz, anatili.kazgazeta.kz and baq.kz. These web pages mostly contain official news, history, and literary texts. To date, there are many studies on extracting a large number of keywords in English and other languages. There are a limited number of studies that involves Kazakh language processing but none of them are focused on keyword extraction.

In this paper, Keyword Extraction from Kazakh News Dataset was presented. Model performance results were obtained by using the BERT base-uncased and BERT base-multilingual-uncased pre-trained language model for the newly compiled Kazakh News Dataset KND. The rest of the paper is organized as follows. In section 2, we present the related works and then present the materials and methods used in the study along with the followed process in Section 3. Concluding remarks are given in Section 4.

2. Related works

Keyword extraction is an important part of text mining. In the literature, there are many suggested methods for keyword extraction. These methods can be grouped under two main headings as supervised and unsupervised approaches. Sıddıqı and Sharan [2] grouped the keyword extraction approach according to style and method under four main headings as statistical, linguistic, machine learning, and domain-specific approaches. In the literature, models were created by using algorithms for keyword extraction, the prediction performances of the models obtained were

compared and it was examined which algorithm created more successful models in the data source used.

Most of the literature related to keyword extraction covers the studies that extract keyphrases from English texts. Moreover, it has been observed that natural language processing studies developed using deep learning methods focus on English. Although there are many supervised and unsupervised models have been proposed for keyword extraction, there is a limited number of studies available in other languages. According to the information available to us, there is no study on the extraction of keywords and expressions in the Kazakh language. Recently, some studies have been carried out in the Kazakh language due to the rapid progress of machine learning and deep learning. Bekbulatov and Kartbayev [3] focused on the internal morphological structure of Kazakh sentences and compared it with unsupervised rule-based language processing models.

Myrzakhmetov and Kozhirbayev [4] in their work, conducted language modeling experiments on a newspaper dataset using traditional n-gram and LSTM-based neural networks in their study and stated that neural-based models outperformed n-gram-based models.

Nugumanova and Mansurova examined statistical and graph-based methods in the monograph and showed the relationship between the term derivation task and the thematic modeling task. Also, general information about automatic recognition of terms is given and working methods of complex concepts such as terminology are discussed and examples are implemented in Python libraries and R ecosystem [5].

Rakhimova et al. [6] in their work, presented provide the research and development of a question-answer system based on the BERT model for the Kazakh language. Since a large enough corpus for the Kazakh language was not clear. In the study, a corpus of 60,000 sentences was collected by translating from English, the questions and answers into separate files.

While it used mostly machine learning-based classification algorithms before, it has now started to move towards the field of deep learning. The successful application of deep learning methods to various problems has been effective in the use of these methods for keyword extraction. Recently, they proposed the BiLSTM-CRF model to solve keyword extraction as a sequence tagging problem. Conditional Random Fields, which is an unsupervised method, is used in the Conditional Random Field (CRF) output layer [7]. In BiLSTM-CRF-based DAKE by Santosh et.al., [8] contextual information at the document level is included in the network with the document-level attention mechanism.

Wang et al. [9] in their work, presented The Multiple Article Keyword Extraction (M-GCKE) algorithm, which extracts keywords from multiple articles, as a method that expands the relationship between keywords from a single paper to multiple papers to extract keywords more accurately. To predict keywords more precisely, a graph convolutional network (GCN) is used to learn structural information and node attribute information in the network.

Mu et al. [10] in their work, presented Span Keyphrase Extraction (SKE) model that extracts a span-based attribute representation from all content for keywords. Architecture that models generator and extractor keyword generation as a single architecture divided into two subtasks and named BERT-Absent Keyphrase Generation (AKG) and BERT-Present Keyphrase Extraction (PKE) [11].

To capture the deep syntactic details, a parallel Seq2Seq model ParaNet which integrates deeply into the seq2seq model is proposed by Zhao et al. [12]. In another study, Vaswani et al. [13]

presented BERT model is designed to pre-train deep bidirectional representations from the unlabeled text by conditioning both left and right contexts together in all layers.

3. Method and Material

In the study, the Bidirectional Encoder Representations from Transformers (BERT) model, one of the most popular approaches among others in natural language processing (NLP), was chosen to investigate keyword extraction in the Kazakh language. Data set formation and the steps that include data collection, platform selection for model training, the test of the model, and evaluation of the model are seen in Figure1. In this section, we first overview the evaluated model and then explain the datasets generation and experimental setup, model training, and evaluation.

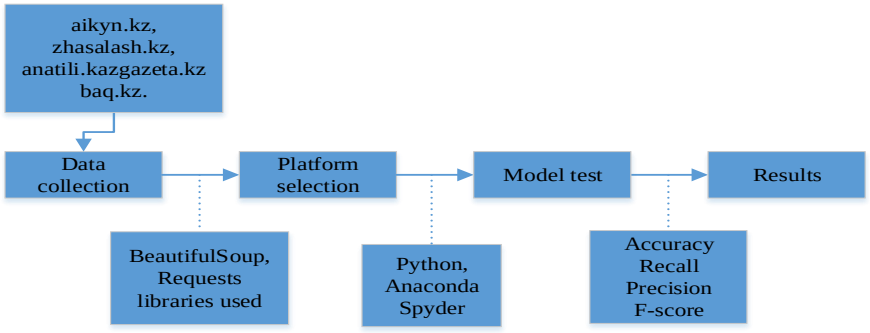

Figure 1. Process steps to apply the model.

3.1. Overview of the BERT model

In recent years, models based on transformers have become quite popular in different NLP tasks. BERT (Bidirectional Encoder Representations from Transformers) which is a machine learning model for natural language processing, is a converter model developed by Google, like many other algorithm updates, to better understand queries and provide more accurate results to its users. The transducers have a structure that works with the self-attention mechanism formed by the tail structure and the transformer - model architecture is given in Figure 2.

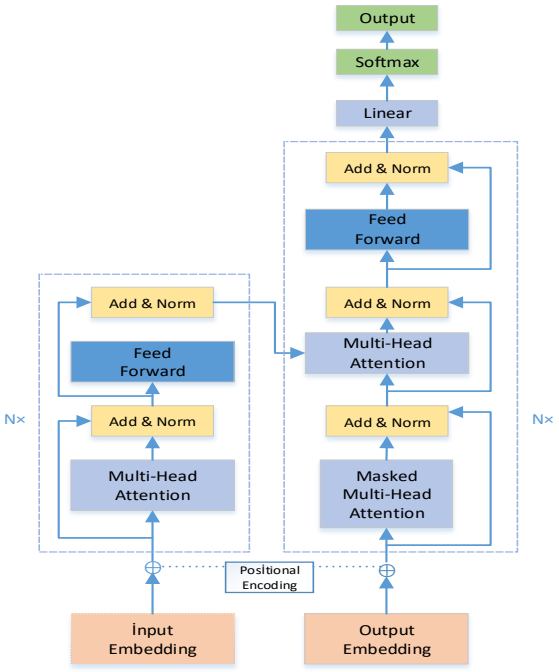

Figure 2. The transformer - model architecture [14].

BERT is designed to pre-train deep bidirectional representations from the unlabeled text by co-conditioning both left and right contexts across all layers. There are two approaches to applying a pre-trained model, feature-based and fine-tuning. BERT architecture is created by separating the encoder and decoder of the transformers seen in Figure 3 and connecting more than one encoder and decoder one after the other. In the training phase, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) techniques are used.

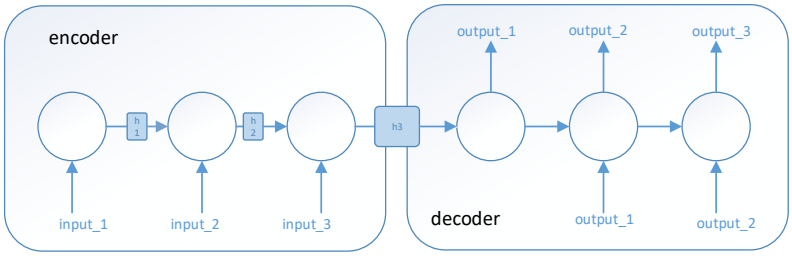

Figure 3. Transmission of layers from the decoder to encoder.

There are two steps for introducing BERT in detail: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is initialized with pre-trained parameters and all parameters are fine-tuned using labeled data from downstream tasks. Each downstream task has individual fine-tuned patterns, even if they are started with the same pre-trained parameters. Structures built with the BERT model require a pre-trained model. For this reason, in our study, results were obtained with the BERT base-multilingual-uncased model, which is a pre-trained BERT model with datasets with 7060 news obtained from Kazakh news websites.

3.2. Datasets and experimental setup

Dataset was collected from Aikyn.kz, Anatili.kazgazeta.kz, zhasalash.kz and Baq.kz published in the Kazakh language. Four data sets include 7060 news published online and it is 10 MB in size. BeautifulSoap and Requests libraries were used to load data from the news site published in the Kazakh language. There are several web pages whose texts were written in both Russian and Kazakh. In the study, sites with originally written in only Kazakh were selected. These web pages mostly contain news, history, and literary texts. In the available data set, the headline of the Kazakh news site contains the news text. The statistics of the datasets are given in Table 1.

Table 1. Statistics of the Kazakh News Data Set (KND)

Source	**# document**
Aikyn.kz	2170
Anatili.kazgazeta.kz	2110
Zhasalash.kz	1000
Baq.kz	1780
Sum	7060

After generating the datasets, the datasets were preprocessed. Data preprocessing requires two different types of processing. The first of these operations is the selection and merging of the dataset, and the second is the processing of the data to make the data more useful for data mining. In the first stage, parts such as URL and author name were removed from the text, fields, and attributes to be used in the data set were determined, and integration processes were carried out. In the second stage, the data set is cleared of repetitive data, punctuation marks, and symbols and is ready to work on the cleaned data.

Experimental studies were carried out on a 2-core computer with Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz processor and 8 GB RAM. Python was used as the programming language. The application was carried out in the Anaconda Spyder environment.

3.3. Evaluation metrics

The performance of the model was evaluated by using accuracy, precision, sensitivity, and f-measurement (F-Score) evaluation metrics. True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values are used to calculate the F_1 score. TP means that the model yields positive results for both the estimated and actual values; TN means that the model yields negative results on both the estimated and actual values; FP is when the model prediction is positive, but the true value is negative; FN is explained as the true value giving positive results while the prediction of the model is negative. In this case, TP and TN are considered correct results, FP and FN are incorrect results.

The accuracy is calculated by the ratio of the TP and TN values that the model predicts correctly to all the predicted values of TP, TN, FP, FN:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

The precision is the ratio of the number of TP values predicted by the model to the number of TP and FP values, which are all positive outcomes the model produces:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The sensitivity, on the other hand, can be found by the ratio of the number of TP values predicted by the model to the numbers of TP and FN, which are all true positive results that the model should produce:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1 score, on the other hand, can be defined as the harmonic mean of precision and sensitivity values:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

F-Score criterion is used in the evaluation of keyword extraction algorithms. In the calculation of this score, the complexity matrix is created by looking at the actual /predicted values that the trained model has produced. Table 2 contains the confusion matrix for the BERT-Base-multilingual-uncased pre-trained (BERT Multilingual) language model and BERT. True Positive, which is a keyword in the matrix and is predicted as a keyword, False Positive, which is not a keyword but is predicted as a keyword, contains False Negative, which is not a keyword but is marked as a keyword, and finally True Negative, which is not a keyword and is marked as not a keyword.

Table 2. Confusion matrix of language models

		Actual			
		BERT Multilingual		*BERT*	
Predicted		*Positive*	*Negative*	*Positive*	*Negative*
	Positive	356	1237	302	1354
	Negative	1075	67232	1146	67098

The results obtained with the BERT-base-uncased and BERT-base-multilingual-uncased language model for the Kazakh news data set (KND) are shown in Table 3.

Table 3. Comparison of the results

Model	Data sets	Accuracy	Recall	Precision	F1-score
BERT	KND	0,96	0.18	0.21	0.20
BERTMultilingual	KND	0.97	0.25	0.23	0.24

In this study, 0.24 F_1-score performance was obtained with the BERT base-multilingual-uncased language model while BERT base-uncased model achieved 0.20 F_1-score for the Kazakh news data set KND. As seen from Table 3, both methods have similar and high accuracies. However, looking for only accuracies may be misleading since there is quite difference between actual keywords and correctly predicted ones as seen in confusion matrix given in Table 2.

4. Conclusions

In this paper, Keyword Extraction from Kazakh News Dataset was presented. Model performance results were obtained by using the BERT and BERT base-multilingual-uncased pre-trained language model for the newly compiled Kazakh News Dataset KND. KND consists of 7060 news that were collected from the web pages anatili.kazgazeta.kz, Bilimdinews.kz, and zhasalash.kz using the BeautifulSoap and Requests libraries. As expected, BERT base-multilingual-uncased model has achieved higher BERT base-uncased model has achieved higher performance since BERT is a pre-trained model on English language using a masked language modeling (MLM) objective. However, results reveals that BERT is not as successful in keyword extraction in Kazakh language as in sequence classification, token classification or question answering tasks that use the whole sentence.

Authors' contributions

Conceptualization, A.A and A.C.; methodology, A.A and A.C.; software, A.A.; validation, A.C.; formal analysis, A.A and A.C.; investigation, A.A.; resources, A.A.; data curation, A.A.; writing original draft preparation, A.A.; writing review and editing, A.C.; visualization, FAA.; supervision, A.C.

Both authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1]. Birdevrim, S. A., Boyacı, A., Al Thani, D. A. S., "İyileştirilmiş otomatik anahtar kelime çıkarımı (BRAKE)." İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi. 2018, 1(1): 11-19.

[2]. Siddiqi, S., Sharan, A., "Keyword and keyphrase extraction techniques: a literature review". International Journal of Computer Applications, 2015, 109 (2).

[3]. Bekbulatov, E., Kartbayev, A., "A study of certain morphological structures of Kazakh and their impact on the machine translation quality". In: 2014 IEEE 8th International Conference

on Application of Information and Communication Technologies (AICT), 2014, 1-5.
[4]. Myrzakhmetov, B., Kozhirbayev, Zh., “Extended language modeling experiments for kazakh.” the proceedings of 2018 International Workshop on Computational Models in Language and Speech, 2018.
[5]. Nugumanova, A., Mansurova, M., “Tabigi til matinderindegi terminderdi avtomatti turde tanu” Monografiya, Oskemen, ShQMU, 2019.
[6]. Raximova, D.R,. Qasimova, D.T, İsabaeva D.N., “Qazaq tiline arnalgan BERT modeli negizinde suraq-jauap juyesin zertteu jane azirleu.” Abay atındagı QazUPU-nin XABARSHISI, «Fizika-matematika gılımdarı» seriyası, 2021, 4 (76).
[7]. Alzaidy, R., Caragea, C., Giles, C., “Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents.” In: The world wide web conference, 2019, 2551-255.
[8]. Santosh, T.Y., Sanyal, D.K., Bhowmick, P.K., Das, P.P., “Dake: Document-level attention for keyphrase extraction.” In Proceedings of the European Conference on Information Retrieval, 2020, 392–401.
[9]. Wang, J., Peng, H., Hu, J. S., “Automatic Keyphrases Extraction from Document Using Neural Network.” In Advances in Machine Learning and Cybernetics, 4th International Conference, 2006, 633-641.
[10]. Mu, F., et.al., “Keyphrase extraction with span-based feature representations.” arXiv preprint arXiv: 2002. 05407.
[11]. Liu, R., Lin, Z.,Wang, W. “Addressing Extraction and Generation Separately: Keyphrase Prediction With Pre-Trained Language Models.” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3180-3191.
[12]. Zhao, J., Zhang, Y., “Incorporating Linguistic Constraints into Keyphrase Generation.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 5224-5233.
[13]. Vaswani, A., et.al., “Attention is all you need.” In Advances in neural information processing systems. 2017, 5998-6008.
[14]. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J. “Attention is all you need in speech separation.” In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, 21-25.