BERT for Russian news clustering

Khaustov S. V.
MTS AI
Moscow, Russia
haustovsv@gmail.com

Gorlova N. E.
MTS AI
Moscow, Russia
n.gorlova@yandex.ru

Kalmykov A. V.
MTS AI
Moscow, Russia
takiholadi@gmail.com

Kabaev A. S.
MTS AI
Moscow, Russia
askabay3@mts.ru

Abstract

This paper provides results of participation in the Russian News Clustering task within Dialogue Evaluation 2021. News clustering is a common task in the industry, and its purpose is to group news by events. We propose two methods based on BERT for news clustering, one of them shows competitive results in Dialogue 2021 evaluation. The first method uses supervised representation learning. The second one reduces the problem to binary classification.

Keywords: BERT, text clustering, news clustering, text classification, Russian text clustering, RuBERT
DOI: 10.28995/2075-7182-2021-20-385-390

BERT для кластеризации русскоязычных новостей

Хаустов С. В.
МТС ИИ
Москва, Россия
haustovsv@gmail.com

Кабаев А. С.
МТС ИИ
Москва, Россия
askabay3@mts.ru

Горлова Н. Е.
МТС ИИ
Москва, Россия
n.gorlova@yandex.ru

Калмыков А. В.
МТС ИИ
Москва, Россия
takiholadi@gmail.com

Аннотация

Статья описывает результаты участия в соревновании по кластеризации русскоязычных новостей Dialogue Evaluation 2021. Кластеризация новостей часто встречается в индустрии и основной целью является группировка новостей по событиям. Мы предложили два метода основанных на модели BERT, один из них показал конкурентный результат в соревновании. Первый метод использует обучение с учителем для получения оптимальных векторных представлений для кластеризации. Второй метод сводит задачу к бинарной классификации.

Ключевые слова: BERT, кластеризация текстов, кластеризация новостей, классификация текстов, кластеризация русских текстов, RuBERT

1 Introduction

This paper describes models used in Dialogue Evaluation 2021 competition in Russian news clustering [7]. Our team was called naergvae and took second place among thirteen participants. The competition aims to collect and compare approaches in news clustering task and the task of selecting the best headline for the resulting clusters. This paper focuses only on the first part of the competition - clustering. News clustering often occurs as a practical problem in news aggregators. The current problem aims to group news by one event into one cluster. These groups can be used for event importance estimation, news picture of the day visualization, and other tasks. The other important task is to filter similar content. In MTS AI we use this approach when developing a news service which is then used by a smart assistant. The smart assistant can read news from several sources, and it is better if the assistant does not deliver similar consecutive news about one event.

2 Related work

A natural approach to the competition task leads to review related work from the two following perspectives: choosing good news document representation(text embeddings) and choosing a good clusterization/classification scheme.

Some of the algorithms were based on unsupervised learning algorithms to cluster news articles, followed by supervised learning algorithms to classify recent articles, such as the K-Means Clustering algorithm. In [3] it is described several algorithms for news clustering, such as similarity measures. It was also found that k-means with knowledge from WordNet give better aggregate results when it comes to efficiency and the WordNet-enabled W-k means clustering algorithm significantly improves standard k-means generating. In [11] a novel clustering method was announced for an incoming stream of multilingual documents into monolingual and cross-lingual story clusters which consider a small and known number of labels.

Recent EMNLP work [14] evaluates the quality of news document embeddings and reports BERT outperforms Word2vec, GloVe, fastText, ELMo on Reuters and 20 Newsgroups datasets.

Using monolingual Russian pre-trained model RuBERT [9] is better than multilingual BERT for Russian language tasks. BERT [2] has set a new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity, which is very similar to the clustering task. However, this is computationally inefficient, and this problem is solved in Sentence-BERT [13] by deriving semantically meaningful sentence embeddings that can be compared using cosine-similarity.

Other papers using Sentence-BERT on news clustering: [6], [10]. Recently in the text clustering problem [15], Supporting Clustering with Contrastive Learning (SCCL) was proposed - a novel framework to leverage contrastive learning. Considering the landscape of the embedding clustering methods, there are no recent game-changers to our knowledge. Algorithms from a decade ago [1], [5] are still at the forefront of many near state-of-the-art results [12].

3 Method description

As for embeddings, we focus on the power of pre-trained Transformers encoders, specifically BERT, since it has been dominated on a wide range of NLP tasks.

In recent years, models based on transformers have become very popular in different NLP tasks. Our approach follows this trend. We introduce two models based on Bidirectional Encoder Representations from Transformer BERT. The first model is trying to learn good news text representation embeddings for subsequent clustering. The second one uses binary classification to classify if two news texts are from the same group or not. For both of these models, input is tokenized concatenation of headline and news text.

3.1 Embeddings using BERT triplet networks

The first approach was inspired by [13]. The idea is to train a model which will produce a fixed-size embedding representation vector for news text. These vector representations are then used for unsupervised clustering with cosine distance metric. The scheme of model inference is shown in Fig.1.

First, tokenized news text goes to the pre-trained BERT layers. Then, BERT output embeddings for each token are averaged by the pooling layer. These averaged vectors then proceed to the fully connected layer size of 768, followed by L2 normalization layer. For model training, hard triplet loss is used [8]. BERT weights do not freeze while training and initialized from pre-trained model RuBERT [9]. We trained the model on GPU Tesla V100 for ten epochs with a learning rate of 1e-6 and a batch size of 16. For the final result, the model is used to make representation embeddings for each news text, which is then used for agglomerative clustering with average linkage and cosine distance. Our model was compared on the public leaderboard with two models without fine-tuning and our model was better. To obtain vector representations of texts, the Universal Sentence Encoder (USE [4]) model and SBERT distilbert-multilingual-nli-stsb-quora-ranking model were applied, the comparison is shown in Table 1.

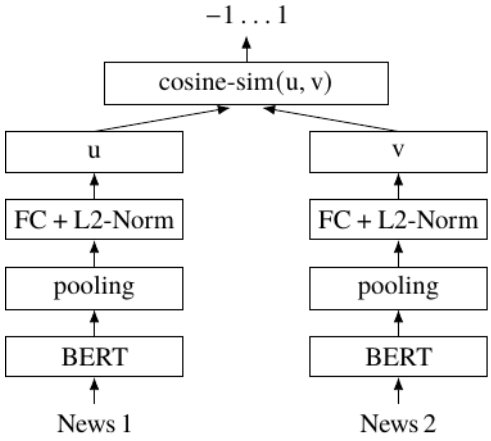

Figure 1: Model inference

Model	F1-score public LB
USE	89.4%
SBERT	88.1%
Our	91.7%

Table 1: Embeddings comparison on public leaderboard.

3.2 BERT classifier

The second approach is based on the transfer from clustering problem to news pair binary classification problem. We train a model which can classify whether the news texts pair is from the same cluster or not. To solve this problem, we used an approach similar to BERT Next Sentence Prediction problem in [2]. The general scheme of the approach is shown in Fig.2. The BERT inputs are a tokenized sequence of both news texts separated by a special token [SEP] preceded by a special token [CLS]. A token segments vector is needed for the model to understand which token belongs to which news text. The input passes through the BERT layers, which make the vector representations of each token. We use BERT pooled output which corresponds to [CLS] special token and follows this output with a softmax classification layer. The model is trained with cross-entropy loss. BERT weights do not freeze while training and are initialized from the pre-trained model RuBERT. Google bert-base-multilingual model and Sberbank AI sbert-large-nlu-ru model were also tried for weights initialization, but RuBERT performed better on the public leaderboard, the comparison is shown in Table 2. We trained a model on GPU Tesla V100 with the batch size of 8 for 6 epochs with a learning rate of 1e-5.

Model	F1-score public LB
RuBERT	96.7%
bert-base-multilingual	96.1%
sbert-large-nlu-ru	96.2%

Table 2: Initial weights comparison on the public leaderboard.

4 Results

We use the dataset with 15K training news document pairs provided by the competition organizers. This dataset is collected from the same data sources as Telegram Data Clustering Contest (2021) but specified with additional human annotations crowdsourced via Yandex.Toloka. Annotators were asked to conduct

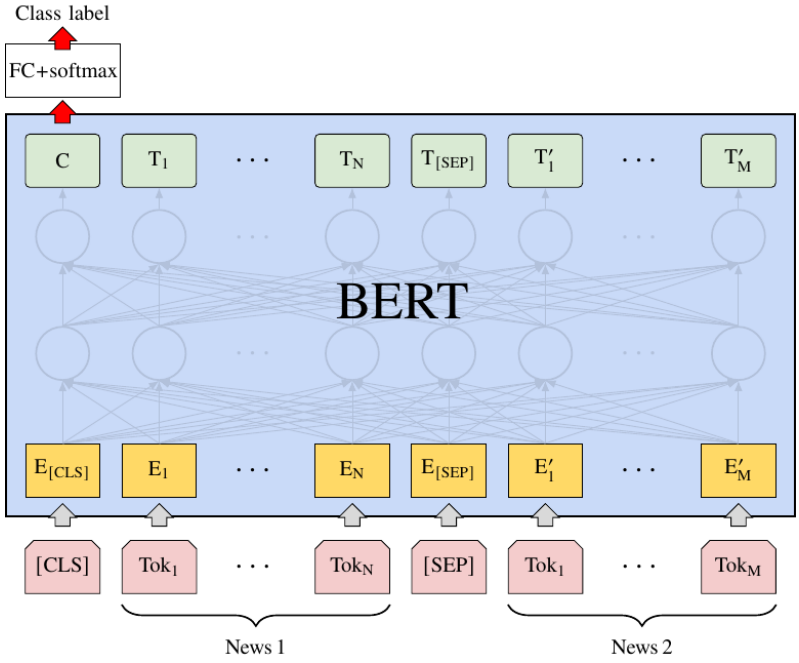

Figure 2: BERT NSP task architecture, depicted from the original paper.

a pairing task, judging pairwise news clustering relatedness. Competition leaderboard (8.5K public, 8.5K private) evaluates in terms of F1-score for positive examples (calculated only on truly positive sentence pairs, i.e., representing the same news documents).

We considered two models: BERT embeddings with agglomerative clustering and BERT classifier in Next Sentence Prediction setting. The results are reported in Table 3. As the table shows, our second approach clearly outperforms the first one. At the end of the competition, this approach took the second place. However, the advantage of our first approach is comparative computational efficiency. It represents a better practical suite for the production inference. While NSP-like Bert classifier requires at inference pairwise news documents comparison (quadratic from the expectedly huge number of news documents), BERT embeddings with clustering require only a single inference and a single neighbor search (allows efficient approximate nearest neighbor search with such tools as Facebook's Faiss, Spotify's Annoy).

We conclude that using a common trick from recommender systems practice, such as generating candidates shortlist with a less complicated model and later reranking with a more advanced one, may yield the solution that will perform well enough in general.

Model	F1-score public LB	F1-score private LB
BERT Embeddings + clusterization	91.7%	91.27%
BERT Classifier	96.7%	95.98%
Competition winner	96.9%	96.04%

Table 3: F1 score on public/private leaderboards and comparison with the competition leader.

5 Error analysis

We analyzed the results of our best-performing model (BERT classifier) on the public test data of 8.5K samples to identify the potential common causes of errors. The confusion matrix shows there were 263 misclassifications: 148 false positive and 115 false negative. Detailed investigation of a random subsample of the errors can be summarized as follows. Table 4 provides some examples.

error type	first news fragment	second news fragment
false positive, doubtful labels	Столичные отделения загс зарегистрировали за апрель около 4 тыс. браков, что на 24 % меньше, чем за аналогичный период прошлого года. Количество разводов в апреле этого года сократилось на 65 % по сравнению с апрелем 2019 года - говорится в сообщении.	В апреле нынешнего года количество разводов в российской столице уменьшилось на 65 % по сравнению с апрелем прошлого года ... В апреле в отделах загс и дворцах бракосочетания москвы поженились около четырех тысяч пар, на 24 % меньше, чем в апреле прошлого года - сообщили агентству в пресс-службе.
false positive, addition of continuation	Житель вологды дмитрий губин подал в суд на кадырова из-за комендантского часа в чечне.	Верховный суд чечни не стал рассматривать иск вологжанина дмитрия губина, в котором он пытался оспорить спецмеры из-за пандемии коронавируса, введенные рамзаном кадыровым.
false positive, same topic but different place/time	В туле покупатели устроили давку в очереди за дешевыми кастрюлями	В башкирии устроили давку из-за кастрюль за 99 рублей
false negative, usage of hypernym/hyponym relations	Существует несколько способов борьбы с сонливостью , но самые частые методы взбодриться — это кофе и физическая нагрузка . команда ученых из канады решила проверить , какой из способов самый действенный.	Ученые из лаборатории университета западного онтарио изучают, как физические упражнения могут улучшить различные показатели здоровья, один из которых — когнитивные способности.

Таблица 4: Error analysis.

We identify 45% of false positive samples to have questionable labels: even though they are labeled as a different news stories, they are strongly the same in our view. We observe another 15% of false positive also relates to the same news main event, but with the addition of continuation or person's comments; such cases are indeed errors due to the competition terms. The rest 40% of false positive samples definitely belong to different news stories but shares similar topics or context, often with different locations and dates: weather forecasts, news about financial exchange rates, announcements of football matches or their results, etc.

Using the previous errors classification terms, we find that false negative samples follow the next distribution: 33% have questionable labels, 48% are supported with additional recap/continuation/person's comment, and 19% are surely the model's errors. Interestingly we noticed several cases of the same news that classifier confused due to the usage of different hypernym/hyponym relations, specifically geographical toponyms. Another pattern that leads to errors is probably the abundance of quoted speech in the texts.

6 Conclusion

In this paper, we presented and compared two approaches for news clustering based on BERT, one of them showing competitive results in Dialogue 2021 evaluation. The first approach is supervised representation learning followed by clustering. This approach is computationally efficient and can be easily applied in real life to a large set of documents. We showed that representation learning was able to outperform unsupervised approaches from baselines. The second method with a binary classifier shows the superiority of supervised learning over unsupervised methods. This

method has shown promising results, but it can hardly be applied without modifications in real life due to performance.

References

[1] Recent Developments in Document Clustering : Rep. : TR-07-35 / Computer Science, Virginia Tech ; Executor: Nicholas O. Andrews, Edward A. Fox : 2007. — 11.

[2] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. — 2018. — 1810.04805.

[3] Bouras Christos, Tsogkas Vassilis. A clustering technique for news articles using WordNet // Knowledge-Based Systems. — 2012. — Vol. 36. — P. 115–128. — URL: https://www.sciencedirect.com/science/article/pii/S0950705112001864.

[4] Daniel Cer Yinfei Yang Sheng-yi Kong Nan Hua Nicole Limtiaco Rhomni St John NoahConstant Mario Guajardo-C espedes Steve Yuan Chris Tar et al. Universal Sentence Encoder. — 2018. — 1803.11175.

[5] Minaee Shervin, Kalchbrenner Nal, Cambria Erik et al. Deep Learning Based Text Classification: A Comprehensive Review. — 2021. — 2004.03705.

[6] Saravanakumar Kailash Karthik, Ballesteros Miguel, Chandrasekaran Muthu Kumar, McKeown Kathleen. Event-Driven News Stream Clustering using Entity-Aware Contextual Embeddings. — 2021. — 2101.11059.

[7] Gusev Ilya; Smurov Ivan. Russian News Clustering and Headline Selection Shared Task // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. — 2021.

[8] Hermans Alexander, Beyer Lucas, Leibe Bastian. In Defense of the Triplet Loss for Person Re-Identification // CoRR. — 2017. — Vol. abs/1703.07737. — 1703.07737.

[9] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — 1905.07213.

[10] Linger Mathis, Hajaiej Mhamed. Batch Clustering for Multilingual News Streaming. — 2020. — 2004.08123.

[11] Miranda Sebastiao, Znotins Arturs, Cohen Shay B., Barzdins Guntis. Multilingual Clustering of Streaming News. — 2018. — 1809.00540.

[12] Pugachev Leonid, Burtsev Mikhail. Short Text Clustering with Transformers. — 2021. — 2102.00541.

[13] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. — 2019. — 1908.10084.

[14] Sia Suzanna, Dalmia Ayush, Mielke Sabrina J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! — 2020. — 2004.14914.

[15] Zhang Dejiao, Nan Feng, Wei Xiaokai et al. Supporting Clustering with Contrastive Learning. — 2021. — 2103.12953.