Towards News Aggregation in Russian: a BERT-based Approach to News Article Similarity Detection

Glazkova A. V.
University of Tyumen
Russia, Tyumen
a.v.glazkova@utmn.ru

Abstract

This paper presents neural models developed for the Dialogue Evaluation 2021: Russian News Clustering shared task. The participants were offered the news clustering task, but solutions based on the classification were also accepted. We applied a classification approach and implemented a hard-voting ensemble of pretrained language models fine-tuned on the official competition dataset. Our solution ranked third among 17 teams at the Russian News Clustering shared task in both private and public leaderboards and obtained 95.73% of F1-score (private LB) and 96.5% (public LB).

Keywords: news similarity, news clustering, Russian, BERT, ensembling learning
DOI: 10.28995/2075-7182-2021-20-1050-1054

К вопросу об агрегации русскоязычных новостей: подход к обнаружению похожих новостных статей с помощью BERT

Глазкова А. В.
Тюменский государственный университет
Тюмень, Россия
a.v.glazkova@utmn.ru

Аннотация

В статье описаны нейросетевые модели, разработанные для соревнования по кластеризации русскоязычных новостей Dialogue Evaluation 2021. Участникам соревнования было предложена задача кластеризации, однако решения, построенные на применении классификации, также принимались оганизаторами. Данная статья описывает подход, построенный на применении бинарной классификации и реализованный с помощью ансамбля предобученных лингвистических моделей, настроенных (fine-tuned) на официальном наборе данных. Описанное решение заняло третье место среди результатов 17 команд. Показатель качества в соответствии с F-мерой составил 96.5% на открытом лидерборде и 95.73% на приватном лидерборде.

Ключевые слова: схожесть новостных текстов, кластеризация новостей, русский язык, BERT, ансамблевое обучение.

1 Introduction

News sites and communities produce a large amount of textual information on various subjects. In recent years, there has been an increasing interest in the creation of news aggregators that consolidate thousands of articles from different publishers and websites into a single page showing the latest news. In this regard, detecting news stories related to one plot to combine them into news clusters seems to be an urgent research task. Here we focus on the problem of news similarity detection in Russian as a part of the Russian News Clustering, Headline Selection, and Headline Generation (Dialogue Evaluation 2021) shared task [5].

This work is based on the participation of our team in the news clustering task. Organizers of the shared task provided participants with train and test sets of Russian news item pairs with a binary annotation

indicating the relation or no relation to one news plot. We experiment with several text classification models using ensemble learning and additional text preprocessing. The code is available on Github[1].

Inspired by the recent success of state-of-the-art transformer architectures and language models, we explore an application of Bidirectional Encoder Representations from Transformers (BERT) [2] model extension for Russian language (RuBERT) [6] to news clustering. We consider this task as a text classification problem. To date, a number of studies have begun to examine BERT-based models for news text classification, such as [1, 3, 10, 11]. We demonstrate that the text classification approach for news similarity detection shows sufficiently high quality. In our experiments, the word representation of numbers and the use of both headline and news text improve the model quality.

The rest of the paper is organized as follows. The definition of the task and the data description have been summarized in Section 2. The proposed methods, experimental settings, results, and discussion have been presented in Section 3. Section 4 is a conclusion.

2 Shared Task Overview

2.1 Task Definition

The purpose of the shared task is to determine whether a pair of news items are related to the same plot. A plot is a collection of news stories about the same event. The notion of the «event» is defined narrow. It means something that happened at a certain time, with certain people, and in a certain way. Different sources can write about the same event. Descriptions of these events in different sources are combined into one plot.

According to the organizers of the shared task, a pair of news items refers to the same event when these news items have the same:

- time of the event (and close to the time of writing);
- numbers, such as the stock price of a company or the number of victims;
- locations, for example, the location of an event or the location of an accident.

A pair of news items is NOT related to the same event when the news items contain:

- inconsistent facts: the time or place of the event, significantly distinguishing the number of victims, etc.;
- a description of an event in one of the news and a commentary on this event by some person in another.

Originally, the task is formulated as a clustering problem. However, participants could use both clustering-based approaches and approaches using text classification techniques. Formally, the classification task can be described as follows.

- **Input.** Given a pair of news items.
- **Output.** One of two different labels, such as "OK" (related to the same plot) or "BAD" (not related to the same plot).

The results were evaluated using a standard F1-score.

2.2 Data Description

News documents were taken from the Telegram Data Clustering Contest[2] and annotations were collected via Yandex.Toloka crowdsourcing platform. The dataset contains 31,811 news pairs split into training (14,838) and test (16,973) data. During the competition, it was forbidden to use any news documents from the test set for training or pretraining, including pretraining word vectors on these documents.

Below are examples of news pairs from the dataset:

- **Headline 1:** На мужчину, заразившего коронавирусом двух постояльцев хостела в Москве, завели дело *[A case was opened against a man who infected two guests of a hostel in Moscow with coronavirus]*.
  Headline 2: В Москве завели дело на мужчину с COVID-19, который поселился в хостеле, чтобы не заражать домочадцев *[In Moscow, a case was opened against a man with*

[1] https://github.com/oldaandozerskaya/DE2021_news_similarity
[2] https://contest.com/docs/data_clustering2

COVID-19, who settled in a hostel so as not to infect households].
Label: OK.

- **Headline 1:** Правительство выделило Свердловской области «антикризисную помощь» на 5 миллиардов рублей *[The government allocated the Sverdlovsk region "anti-crisis assistance" for 5 billion rubles].*
  Headline 2: Муниципальные образования Республики Чувашии получат дополнительную финансовую помощь *[Municipalities of the Republic of Chuvashia will receive additional financial assistance].*
  Label: BAD.

3 Our Solution

Our final solution is a classifier based on the BERT [2] architecture. All BERT-based models had similar preprocessing and all models were trained or fine-tuned on the official training set. For each model, the training data were randomly divided into a training and validation subset in a ratio of 90:10. Depending on their average performance the best three models were ensembled and submitted.

3.1 Input Representation

A good headline represents a summary of an article. However, headlines should encourage users to open full articles and should not contain complete information [4]. Therefore, we represented the news item as a concatenation of its headline and text in our final solution. Since the BERT-based models can process sequences of no more than 512 tokens, we truncated each text to 200 tokens. Further, the texts of the two news items were concatenated into one sequence and supplemented with special tokens «[CLS]» and «[SEP]»:

- «[CLS]» + headline1 + text1[:200] + «[SEP]» + headline2 + text2[:200] + «[SEP]»

Further, news articles contain a lot of numbers. Moreover, the coincidence or proximity of numbers is one of the key criteria for attributing the news to one plot. We experimented with two forms of numbers: original digits and their word form. For word forms, we replaced every number in the text with its word form (e.g. «23» → «двадцать три»[«twenty-three»]). The replacement is performed with num2words[3] Python package. Our model exhibits a tiny increase in performance for word-form numbers (see subsection 3.3).

3.2 Model

Our classifier is based on RuBERT [6], a pretrained language model for Russian. RuBERT was initialized using a multilingual version of BERT-base and trained on the Russian part of Wikipedia and news data. During the development phase, we compared several types of input representation. Further, we trained three models using the most efficient input representation with random training and validation data splitting and combined these models into a hard-voting ensemble.

We conducted our experiments on Google Colab Pro (CPU: Intel(R) Xeon(R) CPU @ 2.20GHz; RAM: 25.51 GB; GPU: Tesla P100-PCIE-16GB with CUDA 10.1). Each model was fine-tuned for 3 epochs. The models are optimized using AdamW [7] with a learning rate of 2e-5 and epsilon of 1e-8, a max sequence length of 512 tokens, and a batch size of 4. We implemented our models using Pytorch [8] and Huggingface's Transformers [12] libraries.

Moreover, we used a Linear Support Vector Classifier (Linear SVC) baseline, which was implemented with Scikit-learn [9]. The baseline was trained on news headlines represented as a bag-of-words matrix for 10,000 words.

3.3 Results

Table 1 shows our experimental results obtained during the development phase of the competition.

As can be seen from the table above, adding news texts to input data greatly improves the quality of the classification (more than 5% relative to a headline-only trained model). Moreover, during our exper-

[3] https://github.com/savoirfairelinux/num2words

Table 1: Experimental results

Model	F1-score (%, public LB)
Linear SVC baseline (headlines)	88.3
RuBERT (headlines)	90.5
RuBERT (+ news text)	96
RuBERT (+ converting numbers to words)	90.7
RuBERT (+ ensembling)	91.1
RuBERT (+ news text, converting numbers to words, and ensembling)	96.5

Table 2: Results on the test set

Rank	Team	F1-score (%, public LB)	F1-score (%, private LB)
1	maelstorm	96.9	96.04
2	naergvae	96.7	95.98
3	**g2tmn**	**96.5**	**95.73**
	Avg result	92.4	87.94

iments, models with word-form number representation showed slightly better performance (+0.2%). As expected, the use of ensemble learning also contributes to an increase in quality (+0.6%). The model aggregated adding news text, converting numbers to words, and ensembling outperforms the baseline by 8.2%.

The final results of the competition are summarized in Table 2. Our system achieved 96.5% of F1-score on the public leaderboard and 95.73% of F1-score on the private leaderboard of this shared task that attracted 17 submitted teams in total. Some error examples are presented below:

- **Headline 1:** Совет ЕС утвердил соглашение об упрощении визового режима с Белоруссией*[EU Council approves visa facilitation agreement with Belarus]*.
  Headline 2: Евросоюз упростил получение визы для Беларуси, уменьшив список документов и стоимость краткосрочной визы*[The European Union simplifies obtaining a visa for Belarus by reducing the list of documents and the cost of a short-term visa]*.
  True label: OK.
  Predicted label: BAD.
- **Headline 1:** В мэрии Оренбурга сообщили о выемке документов по делу о халатности*[The mayor's office of Orenburg reported the seizure of documents in a negligence case]*.
  Headline 2: Следователи пришли с обыском в мэрию Оренбурга *[Investigators came to search the Orenburg mayor's office]*.
  True label: OK.
  Predicted label: BAD.
- **Headline 1:** Від пацанки до панянки-4: дворецкий Школы леди вспомнил трогательные моменты выпускного бала*[From tomboy to girly girl-4*[4]*: the butler of the Lady's School remembered the touching moments of the prom]*.
  Headline 2:Опубликованы видео со съёмок дорамы «Самый прекрасный момент» *[Filming Videos for 'The Most Beautiful Moment' Released]*.
  True label: BAD.
  Predicted label: OK.
- **Headline 1:** Минкульт назвал этапы «разморозки» работы кинотеатров *[The Ministry of Culture called the stages of "defrosting" of the work of cinemas]*.

[4]In Ukrainian

Headline 2:Владелец кинотеатра рассказал, от чего зависит возврат фильмов на большие экраны *[The owner of the cinema told what determines the return of films to the big screens]*.
True label: BAD.
Predicted label: OK.

As you can see from the examples, false-negative errors are usually related to the fact that one of the news items describes the event in more detail. The third example shows an interesting false positive error case probably explained by the spelling of the TV show's name in Ukrainian. The first news in the last example describes the event and the other contains a commentary on this event. This case is stipulated by the organizers of the competition as a criterion for attributing the news to different plots.

4 Conclusion

In this work, we have explored an application of Bidirectional Encoder Representations from Transformers (BERT) to the task of news article similarity detection in Russian. We have compared several approaches to input text representation and noted that our BERT-based model benefits from adding article fragments to input data. A hardvoting ensembling technique obtained our best result on the test data.

References

[1] A BERT-based Ensemble Model for Chinese News Topic Prediction / Jingang Liu, Chunhe Xia, Xiaojian Li et al. // Proceedings of the 2020 2nd International Conference on Big Data Engineering. — 2020. — P. 18–23.

[2] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[3] Glazkova A., Glazkov M., Trifinov T. g2tmn at CONSTRAINT@ AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection // Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT). — 2021. — P. 116–127.

[4] Gusev I. Importance of copying mechanism for news headline generation // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2019. — P. 229–236.

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

[6] Kuratov Y, Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2019. — P. 333–339.

[7] Loshchilov I., Hutter F. Fixing weight decay regularization in adam // arXiv preprint arXiv:1711.05101. — 2018.

[8] PyTorch: An Imperative Style, High-Performance Deep Learning Library / Adam Paszke, Sam Gross, Francisco Massa et al. // Advances in Neural Information Processing Systems. — 2019. — Vol. 32. — P. 8026–8037.

[9] Scikit-learn: Machine learning in Python / Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort et al. // the Journal of machine Learning research. — 2011. — Vol. 12. — P. 2825–2830.

[10] Singh Roshan, Chun Soon Ae, Atluri Vijay. Developing Machine Learning Models to Automate News Classification // The 21st Annual International Conference on Digital Government Research. — 2020. — P. 354–355.

[11] Tan Zhixiong, Chen Bihuan, Fang Wei. Analysis and Application of Financial News Text in Chinese Based on Bert Model // Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference. — 2020. — P. 35–39.

[12] Transformers: State-of-the-art natural language processing / Thomas Wolf, Julien Chaumond, Lysandre Debut et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — 2020. — P. 38–45.