Russian News Clustering and Headline Selection Shared Task

Ilya Gusev
Moscow Institute of Physics and Technology
Moscow, Russia
`ilya.gusev@phystech.edu`

Ivan Smurov
ABBYY,
Moscow, Russia
`ivan.smurov@abbyy.com`

Abstract

This paper presents the results of the Russian News Clustering and Headline Selection shared task. As a part of it, we propose the tasks of Russian news event detection, headline selection, and headline generation. These tasks are accompanied by datasets and baselines. The presented datasets for event detection and headline selection are the first public Russian datasets for their tasks. The headline generation dataset is based on clustering and provides multiple reference headlines for every cluster, unlike the previous datasets. Finally, the approaches proposed by the shared task participants are reported and analyzed.

Keywords: clustering, event detection, headline selection, headline generation, news, embeddings, Russian, dataset, NLP evaluation

Дорожка по кластеризации и выбору заголовков для русских новостей

Гусев И. О.
МФТИ
Москва, Россия
ilya.gusev@phystech.edu

Смуров И. М.
ABBYY
Москва, Россия
ivan.smurov@abbyy.com

Аннотация

В статье представлены результаты соревнования по кластеризации и выбору заголовков для новостей на русском. В соревновании предлагаются задачи по обнаружению новостных событий, отбору и написанию заголовков для новостей. Вместе с задачами предоставляются наборы данных и базовые решения. Представленные наборы данных для обнаружения новостных событий и выбора заголовков — первые общедоступные наборы данных на русском языке для своих задач. Набор данных для написания заголовков использует кластеризацию и содержит набор эталонных заголовков для каждого кластера, в отличие от предшественников. Представлены и проанализированы подходы, предложенные участниками соревнования.

Ключевые слова: кластеризация, определение событий, выбор заголовков, написание заголовков, новости, эмбеддинги, русский язык, набор данных

1 Introduction

Automatic news feeds and aggregators are a common way to read, search and analyze news. Event clustering is one of the core features for many news aggregators, including Google News and Yandex News. The news event clustering task is to collect news from different news agencies about the same event. It is essential to present all events from different perspectives to create a comprehensive picture. Moreover, selecting the most suitable headline and other entities is possible only within such clusters. The task can also be helpful for news monitoring systems. Correct clustering should help to accurately calculate any statistics about events, companies, or people's mentions in the news.

As for natural language research, this task is a good benchmark for different text clusterization or classification models. For example, the models should identify various types of paraphrasing and recognize named entities to distinguish clusters correctly.

After event clustering, it is necessary to choose the most relevant headline for every cluster. Thus, it is the headline selection task. One can define relevance in many ways. We provide a list of criteria by which a title can be considered suitable. The main points of this list are informativeness and the absence of clickbait.

Instead of choosing one of the existing headlines, one can generate a completely novel headline based on news texts. The main reason to do this is to evade the situations where all presented headlines are inadequate.

To solve these three problems, we composed a shared task as a part of the Dialogue 2021 conference[1]. These three problems are considered as independent tasks and evaluated separately. We provided datasets for each of them and hosted Kaggle-like competitions on the CodaLab platform to determine the best models.

The primary source of data for all tasks is the Telegram Data Clustering contest[2]. Telegram conducted it in 2020. The task was to build a news aggregator over a provided document collection. The subtasks were language detection, category detection, and event clustering. Only HTML documents without any additional annotations were given.

The contributions of our paper are as follows: we present datasets for the Russian news event clustering, headline selection, and headline generation tasks along with baselines. The first two datasets are the first of their kind for the Russian language. The last one is the first dataset for headline generation for Russian with multiple reference headlines. Furthermore, we present and analyze some of the works of the shared task participants.

2 Related work

Many works on event clustering exist. This task is also known as "documental event detection" or just "event detection"[37, 19]. We will use these names as synonyms in this paper.

The Topic Detection and Tracking (TDT) initiative [37] was the first significant effort for this task. The event detection task was a part of TDT. There were only 25 events in the dataset. The CMU approach used TF-IDF embeddings, hierarchical agglomerative clustering with average linking, and a time window with incremental IDF for online detection.

Azzopardi et al.[3] used TF-IDF document representations and incremental k-means clustering with cosine similarity between these representations. There were no human annotations involved, and the comparison was with Google News automatic clustering.

Miranda et al.[28] focused on cross-lingual clustering. They used TF-IDF document representation with separate vectors for different document sections and incremental centroid clustering. There were also two versions of these representations: monolingual and cross-lingual. They introduced a multilingual dataset adapted from Rupnik et al.[30] containing articles in English, Spanish and German. It was manually annotated with monolingual and cross-lingual event labels. Linger et al.[25] utilized the same multilingual dataset, but used the multilingual DistilBERT[13] and the Sentence-BERT[32] triplet network structure for multilingual document representation.

There are some works on news clustering that focus on thematic news clustering instead of event clustering (with categories like "sports" or "technologies") [34].

There are few papers on Russian news event clustering. Dobrov et al.[14] described event clustering on a ROMIP-2006 news collection. Three different document representations were utilized: TF-IDF for texts, TF-IDF for titles, and a conceptual index. Several clustering methods were used, including hierarchical clustering, DBSCAN[12], and centroid clusterings. They utilized custom software for creating a manual reference clustering.

Voropaev et al.[41] used the same document collection we do, Telegram Data Clustering contest document collection, but without a large-scale manual markup. They took the solution of one of the Telegram contest participants as a reference clustering. They utilized TF-IDF, BERT[6, 22], LASER[1] for document representation, and hierarchical clustering or DBSCAN[12].

---

[1]http://www.dialog-21.ru/evaluation/
[2]https://contest.com/docs/data_clustering2

As for headline selection, we did not find any papers with a similar task definition. However, there are several works on clickbait detection in headlines[11, 35] which is a part of our task.

The headline generation is a well-covered task in previous papers. Statistical models for news headline generation in Banko et al.[5] are among the first approaches to this task. Takase et al.[29] were the first to use encoder-decoder neural models for headline generation. In 2019, there was a Dialogue Evaluation shared task for Russian news headline generation[27]. The main drawback of the 2019 shared task was the fact that it was about single-document headline generation. It means that there was only one actual headline for every article, so automatic evaluation methods were not effective in this setup.

3 Clustering

3.1 Data and metrics

For the first two tasks, we took news documents from the Telegram Data Clustering contest covering three days of May 2020. After using all baseline clustering algorithms, we sampled a set of document pairs from their output and included a small number of random pairs. Next, we annotated every pair with Yandex Toloka, a Russian crowdsourcing platform. The task was to determine whether two documents describe the same event. The annotators were provided with a comprehensive guide and a simple heuristic: "if two titles of two documents are interchangeable, these documents are from the same cluster". Five people annotated each pair. Annotators were required to pass training, exam, and their work was continuously evaluated through the control pairs ("honeypots"). We included only pairs with an agreement of 4/5 or 5/5 in the final markup.

	May 25	May 27	May 29
#Documents	19380	20080	19096
#Pairs	14838	8493	8476
#Positives	7301	3918	3896
Fraction of positives	49.2%	46.1%	46.0%

Table 1: Clustering dataset statistics

The final statistics for every day are shown in Table 1. There are 661 unique hosts in the training document collection and at least 500 unique news agencies. The top 5 news agencies by the number of documents are shown in Table 2. The distribution of news agencies by the number of documents is in Figure 1.

Host	#Documents
tass.ru	645
lenta.ru	301
www.mk.ru	223
www.rosbalt.ru	199
ura.news	188

Table 2: Top 5 hosts by the number of documents, May 25

According to annotation guidelines, a pair of documents refer to the same cluster when they have the same: time of the event; numbers, such as the stock price of a company or the number of victims; locations. A pair of documents are not from the same cluster when they contain: inconsistent facts, such as the time or place of the event or significantly distinguished number of victims; description of an event in one of the documents, and a commentary on this event in another document. These criteria are similar to the definition of an event in TDT[37].

We do not publish documents themselves, only two URLs and an annotation result, to evade legal issues. However, Telegram shares archives with all documents, so it is easy to join documents and
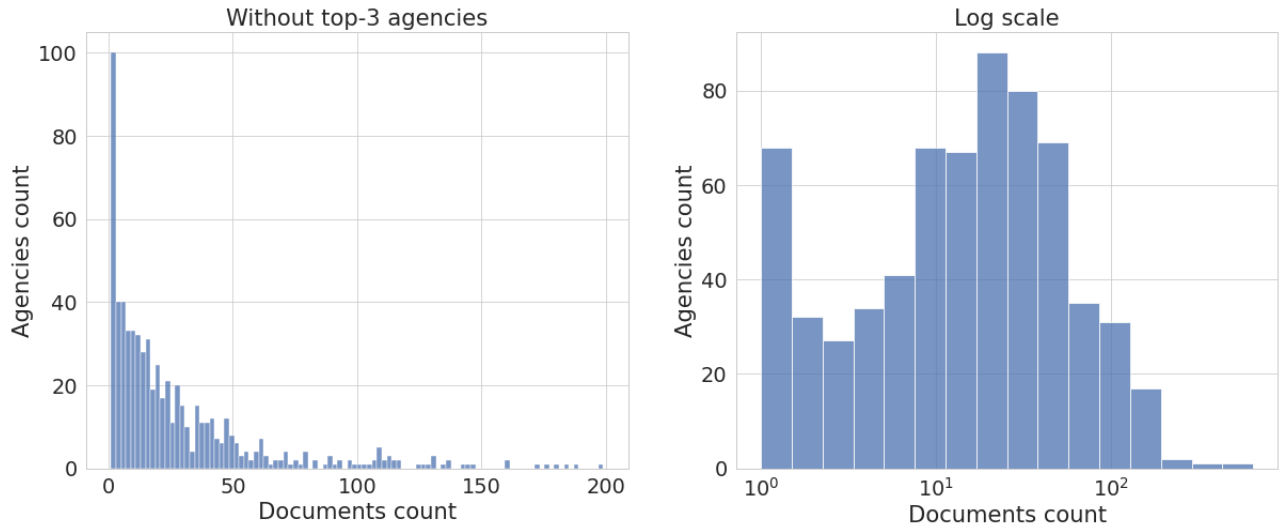
Figure 1: News agencies distribution by the number of documents, May 25

annotations. An example of the joining process can be found in the baseline script[3]. No additional scraping is needed.

In the shared task rules, we forbade the use of any news documents from the test set for training or pretraining, including pretraining word vectors on these documents. In the real world, texts are not available for the model in advance, so the tested models must not use the test-time texts in any way. However, it was permissible to use texts with earlier dates from the Telegram contest for pretraining.

It is possible to collect non-pairwise annotations for clustering with crowdsourcing methods, but it is hard to organize proper quality control of the annotation process. Non-pairwise annotations we know of for event detection were done by experts, not via crowdsourcing.

As the primary metric for this task, we chose F1-score (corresponds to Dice similarity coefficient for clustering). The proper clustering metrics like Adjusted Rand Index[31] are not directly applicable to our dataset as we have no reference clustering for all documents. Several dozen possible pair-counting based clustering metrics[18] are available, and it is an active research question what metric is better[17]. As for our choice, we decided that the metric should be interpretable from the classification perspective, and our classes are imbalanced, so we chose F1.

3.2 Baselines

As a baseline clustering algorithm, we utilized hierarchical agglomerative clustering with average linking.

We used different pretrained embeddings with cosine similarity to compute the distance matrix. The simplest ones are FastText[15] text embeddings, TF-IDF with truncated SVD for dimensionality reduction (latent semantic analysis, LSA[21]), and Universal Sentence Encoder (USE[38]). FastText text embeddings are computed as a concatenation of average, maximum, and minimum of FastText word embeddings. The examples of clusters based on USE embeddings are shown in Figure 2.

Text2Title embeddings are FastText embeddings with an additional linear matrix trained to determine whether a headline and a text are from the same document. HNSW[26] index on the original FastText embeddings was used to mine hard negatives for the triplet loss[20]. The model architecture is depicted on Figure 3. The main advantage of this model is its speed, as it consists of only one linear layer on top of the FastText embedding.

---

[3]https://github.com/dialogue-evaluation/Russian-News-Clustering-and-Headline-Generation/blob/main/baselines.ipynb

Figure 2: Dendrograms for two clusters based on USE embeddings, cosine similarity, and average linking. These clusters contain several errors.
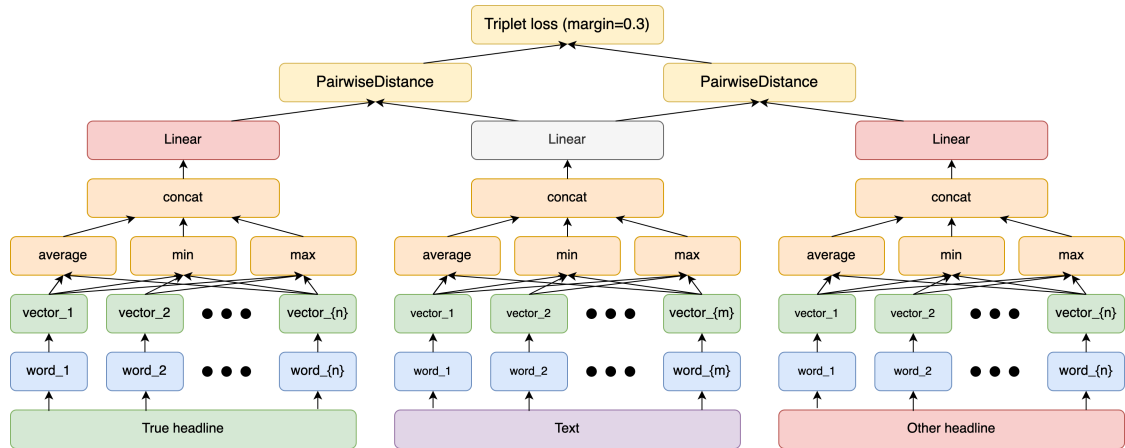

Figure 3: Text2Title model architecture

TTGenBottleneck is a BERT seq2seq model based on RuBERT[22] with practically disabled cross-attention between encoder and decoder. It was trained to predict a headline for a document. "Bottleneck" is the encoder embedding of the first token: the only embedding decoder can attend. We design it[4] to contain all information from the text needed to generate a headline.

We emphasize that none of the baseline models were fine-tuned to the news event detection task in any way.

We present the results of all baselines in Table 3. Text2Title has better scores than plain FastText, as additional pretraining helps to create more representative embeddings. TTGenBottleneck is the best baseline model.

Model	Validation	Public LB	Private LB
TTGenBottleneck	93.4	93.9	93.7
USE	89.3	89.4	87.8
Text2Title	86.5	86.4	84.8
LSA	83.1	82.7	80.5
FastText	80.9	81.9	80.1

Table 3: F1-scores for positive pairs in % for baseline models

3.3 Results

The task can be solved both as a classification task or as a clustering task. The baselines were for the clustering only, but many participants preferred to use classification models.

The final results are in Table 4. There are several participants without scores in the private leaderboard. We made a technical mistake that disabled automatic score evaluation of the private part, and the participants had to resubmit their answers. Not all of them managed to do that.

Rank	Codalab login	Public LB	Private LB
1	**maelstorm** [5]	**96.9**	**96.0**
2	naergvae [4]	96.7	96.0
3	g2tmn [16] [6]	96.5	95.7
4	Kouki [33]	95.5	95.5
5	alexey.artsukevich	95.8	95.3
6	smekur [33]	94.6	93.9
7	nikyudin	93.8	93.0
8	landges	91.6	90.6
9	kapant	90.7	89.9
10	bond005	90.2	89.2
11	anonym	90.6	89.1
12	mashkka_t [40]	85.3	71.5
13	vatolinalex [33]	95.2	47.6
-	blanchefort	94.1	
-	imroggen	90.3	
-	Abiks	89.4	
-	dinabpr [40]	84.4	

Table 4: Final results of the clustering track, F1-scores for positive pairs in %

[4]https://huggingface.co/IlyaGusev/gen_title_tg_bottleneck_encoder
[5]Leonid Pugachev and Alim Adelshin, DeepPavlov
[6]https://github.com/oldaandozerskaya/DE2021_news_similarity

The best solution was a classification model, an ensemble of 4 bert-base-multilingual models trained with stochastic weight averaging (SWA)[2]. The second-placed model was also classification-based and consisted of a single RuBERT[22] in a standard pair classification setting. Finally, the third-placed model was a hard-voting classification ensemble of RuBERT models.

The best clustering-based model (4th place overall) was based on SBERT[32]: a siamese BERT model over RuBERT. The authors[33] improved the model by using Global Multihead Pooling[10] and contrastive loss. It should be noted that this model scored only 0.005 F1 less than the best classification model on the private test. Additionally, this is the only model that did not experience score deduction on the private test set compared to the public set.

Unsurprisingly all top-placed models utilized pretrained masked language models. In particular, the top three classification-based models and the best clustering-based model (4th place overall) all used BERT[6]. The best solution not utilizing language model embeddings was able to score 0.930 F1 (7th place overall) and had the following architecture: CatBoost[9] over FastText[15] and USE[38] as well as some handcrafted features (hyperlinks and named entities intersections).

Another system of note is GPT-3-based[23] zero-shot model. The idea is as follows: the two headlines' perplexity is computed and compared to thresholds. While the final score is rather unimpressive, 0.7 F1, the completely unsupervised nature of training makes this result worth mentioning.

4 Headline selection

4.1 Data and metrics

The documents for this task are from the same collection as the documents for the clustering task. The task was, for each given pair of headlines, to predict which headline is better. There are four possible options: "left", "right", "draw", and "bad". The "bad" option is for the case if two headlines are from different news events. Annotation conditions are the same as for the clustering task. According to annotation guidelines, one headline is better than the other if some of the following conditions are met: it contains more information than the other; it does not hide any details; it does not contain undefined entities; it has no grammatical errors; it is not emotional; it is not too wordy. Some of these criteria are subjective, so we use only examples with high agreement to reduce subjectivity.

The final statistics for every day are shown in Table 5.

	May 25	May 27	May 29
#Pairs	5091	3147	3103
#Left won	2185	1254	1216
#Right won	2167	1269	1207
#Draw	362	184	161
#Bad	377	440	518

Table 5: Headline selection dataset statistics

The primary metric for this task is weighted accuracy. The dataset contains a lot of "bad" examples. They were added to the dataset to prevent data leaks for the clustering task. We ignore them in the denominator and numerator. The systems can predict any label for them. For the other three labels, scores in the numerator are calculated as stated in Table 6. We chose such weights because incorrect predictions for the "draw" pairs should not be penalized as hard as "left/right" errors. We do not provide F1-scores, as the dataset is balanced, and the only information these scores add to accuracy is how well "draw" pairs are detected.

4.2 Baseline

We used a ranking gradient boosting model (CatBoost[9]) with USE embeddings as features and Pair-Logit loss as a baseline for this task. This CatBoost mode is designed to take pairs as input, so it fits the

	Left	Right	Draw
Left	1.0	0.0	0.5
Right	0.0	1.0	0.5
Draw	0.5	0.5	1.0

Table 6: Predictions weights in the accuracy metric

task perfectly. We present the result accuracy of this model in Table 7.

$$PairLogitLoss(pairs, res) = -\sum_{p,n \in pairs} log(\frac{1}{1 + e^{-(res_p - res_n)}})$$

Model	Validation	Public LB	Private LB
USE + Catboost	81.0 ± 1.5	81.2 ± 0.5	81.1 ± 0.4

Table 7: Weighted accuracy for the baseline model, 5 runs, in %

4.3 Results

Most of the successful submissions are based on the same schema as the baseline. The results are presented in Table 8. There is only one submission that surpasses the baseline significantly. It utilizes an ensemble of ranking models trained on different embeddings, including USE, SBERT[32], RuBERT[22], XLM-R[39], and mT5[44]. The most effective single model was mT5.

Rank	Codalab login	Public LB	Private LB
1	**sopilnyak** [42]	**86.0**	**85.4**
2	landges	81.3	82.0
3	nikyudin	83.2	81.6
4	LOLKEK	80.8	81.4
5	maelstorm	81.8	79.8
6	a.korolev	65.8	66.2

Table 8: Final results of the headline selection track, weighted accuracy in %

5 Headline generation

5.1 Data and metrics

In this track, the task was to generate a headline for a news cluster. It should be similar to any of the headlines of cluster documents. Other formulations to this task are possible. For example, there is a much more complicated task formulation where one should generate a headline similar to the best headline of the cluster. However, we chose the weak formulation as it does not require any annotation. We were not able to use the documents and annotations of the first two tracks because of possible data leaks.

To make a dataset entirely unknown for participants, we scraped news documents from a test version of the Telegram news aggregator[7] from March 9 to March 12, 2021. We used the TTGenBottleneck clustering method with a distance threshold skewed into precision to the detriment of recall, as it forms very restricted clusters.

We made available for participants only clustered document texts without any headlines or additional meta information. There are 6726 documents and 1035 clusters in the dataset.

---

[7] https://1398.topnews.com/ru/

This formulation can be considered more robust than the single document (and single news agency [27]) headline generation as it does not force the model to generate a headline in a particular style to get good automatic metrics.

We use traditional automatic metrics for headline generation, ROUGE[24], and BLEU[8]. We calculate metrics between a predicted headline and every actual headline of a cluster and use the maximum score as a prediction score.

5.2 Baselines

There are two baselines for this track. The first is Random Lead-1, where we choose the first sentence of a random document from the cluster as a baseline. The second baseline is a generative sequence to sequence[36] RuBERT[22] trained on news text-title pairs from the Telegram contest[8]. We generate a headline for every text in the cluster and choose randomly from them. We present the metrics for the baselines in Table 9.

Model	ROUGE	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Seq2seq RuBERT	44.9 ± 0.3	74.5 ± 0.5	52.5 ± 0.4	32.8 ± 0.4	49.5 ± 0.5
Random Lead-1	30.2 ± 0.5	47.8 ± 0.4	36.8 ± 0.6	20.4 ± 0.4	33.6 ± 0.6

Table 9: Results of headline generation baseline models, 5 runs, in %

5.3 Results

There were only two participants in this track, mainly because of hard time restrictions. Their scores can be seen in Table 10. The baseline remained unbeaten.

The solution by `Rybolos` was based on fine-tuning the ruGPT-3 Large model. We suppose their metrics are low compared to baselines because of possible technical error or only a tiny part of the dataset used for training.

Codalab login	ROUGE	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
LOLKEK	38.7	69.5	46.3	26.4	43.3
Rybolos	29.2	59.6	36.5	17.6	33.5

Table 10: Final results of the headline generation track, in %

6 Conclusion

6.1 Reproducibility

All materials of the shared task are available at the official repository[9], including all data, annotation guidelines, baselines, and links to the CodaLab competitions. The competitions themselves will be permanently open in order to make comparison easier for new researchers.

6.2 Organization notes

The timeline was as follows:
- February 8: Clustering task started on Codalab.
- February 26: Headline selection task started on Codalab.
- March 13: Headline generation task started on Codalab.
- March 22: Final deadline for all competitions.
- March 28: Final deadline for paper submission.

---

[8]https://huggingface.co/IlyaGusev/rubert_telegram_headlines
[9]https://github.com/dialogue-evaluation/Russian-News-Clustering-and-Headline-Generation

The main reason for a short time window for submissions for the last task was that it was unclear where to get previously unseen data from roughly the same domain. We should have solved this question even before the shared task was announced. Unfortunately, the news clustering task was the only one that we fully prepared before the start.

The other major problem was with the private leaderboard for the clustering task, as the initial private part of the test dataset was a copy of a public part by our technical mistake, so it was impossible to measure metrics without resubmitting answers.

6.3   General conclusions

In the event detection task, most of the successful models were classification-based BERT models. However, it turns out clustering embeddings can be almost as effective when trained with correct pooling and loss function. Moreover, they generalize better, they are easier to deploy, and they are more computationally effective. It is arguably the most important takeaway of the shared task.

Unsurprisingly, big multilingual models and ensembles showed the best metrics in the headline selection task. Nevertheless, the task participants did not diverge much from the baseline solution, so we hope more sophisticated schemas and models will be developed in the future.

There was only one week to develop a valid generative model in the headline generation task. Only two participants managed to present a complete model in these challenging time restrictions, so we consider this track results inconclusive. However, the dataset itself can be helpful in future research.

6.4   Future research

There are several possible directions for future research:
1. Finding models with good accuracy/speed trade-offs is a very promising research path. Almost all of the models used by participants were extremely parameter-heavy and slow. Distillation of these models into lightweight ones is needed in order to enable their use in production systems.
2. Different clustering methods should be inspected. Solutions for the clustering task were agglomerative-centered — almost no one used BIRCH clustering[43], and only a few people used DBSCAN[12] and its modifications.
3. The Telegram Data Clustering news documents are multilingual, so datasets and models should be multilingual too. We created only a Russian dataset because we used a Russian crowdsourcing platform and had a constrained money budget. There are no principal reasons why these datasets should not be multilingual.
4. The existing clustering dataset is focused on 24-hour time windows. Clustering on more extensive periods can be significantly harder. Moreover, online event detection (as it originally was in TDT) is also possible and requires windowed or incremental clustering.
5. Synthetic checklists[7] can be used to evaluate the quality of event detection and headline selection. It is easy to create pairs of documents differing only in numbers, entities, or time of an event.
6. As for the headline generation task, the clustering enables proper conditioning. It is possible to train a model to write headlines in the style of the particular news agency without a thematic bias, as agencies will be equally presented in clusters.
7. It is possible to revisit other TDT tasks in light of recent advances in building document representations with big pretrained models like BERT.

Acknowledgements

References

[1] Artetxe Mikel, Schwenk Holger. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond // Trans. Assoc. Comput. Linguistics. — 2019. — Vol. 7. — P. 597–610. — Access mode: https://transacl.org/ojs/index.php/tacl/article/view/1742.

[2] Averaging weights leads to wider optima and better generalization / Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov et al. // arXiv preprint arXiv:1803.05407. — 2018.

[3] Azzopardi Joel, Staff Christopher. Incremental Clustering of News Reports // Algorithms. — 2012. — Vol. 5, no. 3. — P. 364–378. — Access mode: https://doi.org/10.3390/a5030364.

[4] BERT for Russian news clustering / Khaustov, Gorlova, Kalmykov, Kabaev // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[5] Banko Michele, Mittal Vibhu O, Witbrock Michael J. Headline generation based on statistical translation // Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. — 2000. — P. 318–325.

[6] 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.

[7] Beyond Accuracy: Behavioral Testing of NLP Models with CheckList / Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 4902–4912. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.442.

[8] Bleu: a Method for Automatic Evaluation of Machine Translation / Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. — Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, 2002. — Jul. — P. 311–318. — Access mode: https://www.aclweb.org/anthology/P02-1040.

[9] CatBoost: unbiased boosting with categorical features / Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev et al. // arXiv preprint arXiv:1706.09516. — 2017.

[10] Chen Qian, Ling Zhen-Hua, Zhu Xiaodan. Enhancing Sentence Embedding with Generalized Pooling // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA : Association for Computational Linguistics, 2018. — Aug. — P. 1815–1826. — Access mode: https://www.aclweb.org/anthology/C18-1154.

[11] Clickbait detection / Martin Potthast, Sebastian Köpsel, Benno Stein, Matthias Hagen // European Conference on Information Retrieval / Springer. — 2016. — P. 810–817.

[12] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). — AAAI Press, 1996. — P. 226–231.

[13] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf // arXiv preprint arXiv:1910.01108. — 2019.

[14] Dobrov Boris, Pavlov Andrey. Basic line for news clusterization methods evaluation // Proceedings of the 5-th Russian Conference RCDL-2010. — 2010.

[15] Enriching word vectors with subword information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

[16] Glazkova Anna. Towards News Aggregation in Russian: a BERT-based Approach to News Article Similarity Detection // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[17] Gösgens Martijn, Tikhonov Alexey, Prokhorenkova Liudmila. Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures // arXiv preprint arXiv:1911.04773. — 2019.

[18] Ground Truth Bias in External Cluster Validity Indices / Yang Lei, James C. Bezdek, Simone Romano et al. // Pattern Recogn. — 2017. — May. — Vol. 65, no. C. — P. 58–70. — Access mode:

`https://doi.org/10.1016/j.patcog.2016.12.003`.

[19] A History and Theory of Textual Event Detection and Recognition / Yanping Chen, Zehua Ding, Qinghua Zheng et al. // IEEE Access. — 2020. — Vol. 8. — P. 201371–201392.

[20] Hoffer Elad, Ailon Nir. Deep metric learning using Triplet network. // ICLR (Workshop) / Ed. by Yoshua Bengio, Yann LeCun. — 2015. — Access mode: `http://dblp.uni-trier.de/db/conf/iclr/iclr2015w.html#HofferA14`.

[21] Indexing by latent semantic analysis. / S. Deerwester, S.T. Dumais, G.W. Furnas et al. // Journal of the American Society for Information Science 41. — 1990. — P. 391–407.

[22] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // CoRR. — 2019. — Vol. abs/1905.07213. — 1905.07213.

[23] Language Models are Few-Shot Learners / Tom Brown, Benjamin Mann, Nick Ryder et al. // Advances in Neural Information Processing Systems. — Vol. 33. — Curran Associates, Inc., 2020. — P. 1877–1901. — Access mode: `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[24] Lin Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. — Barcelona, Spain : Association for Computational Linguistics, 2004. — Jul. — P. 74–81. — Access mode: `https://www.aclweb.org/anthology/W04-1013`.

[25] Linger Mathis, Hajaiej Mhamed. Batch Clustering for Multilingual News Streaming // Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only]. — Vol. 2593 of CEUR Workshop Proceedings. — CEUR-WS.org, 2020. — P. 55–61. — Access mode: `http://ceur-ws.org/Vol-2593/paper7.pdf`.

[26] Malkov Yu A., Yashunin D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2020. — Vol. 42, no. 4. — P. 824–836.

[27] Malykh V.A. Kalaidin P.S. Headline Generation Shared Task on Dialogue'2019 // Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue". — 2019.

[28] Multilingual Clustering of Streaming News / Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, Guntis Barzdins // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium : Association for Computational Linguistics, 2018. — Oct.-Nov. — P. 4535–4544. — Access mode: `https://www.aclweb.org/anthology/D18-1483`.

[29] Neural headline generation on abstract meaning representation / Sho Takase, Jun Suzuki, Naoaki Okazaki et al. // Proceedings of the 2016 conference on empirical methods in natural language processing. — 2016. — P. 1054–1059.

[30] News across languages-cross-lingual document similarity and event tracking / Jan Rupnik, Andrej Muhic, Gregor Leban et al. // Journal of Artificial Intelligence Research. — 2016. — Vol. 55. — P. 283–316.

[31] Rand William M. Objective criteria for the evaluation of clustering methods // Journal of the American Statistical association. — 1971. — Vol. 66, no. 336. — P. 846–850.

[32] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. — Association for Computational Linguistics, 2019. — P. 3980–3990. — Access mode: `https://doi.org/10.18653/v1/D19-1410`.

[33] Smirnova, Vatolin, Shkarin. Russian News Similarity Detection with SBERT: pre-training and fine-tuning // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[34] Stankevicius Lukas, Lukosevicius Mantas. Testing Pre-trained Transformer Models for Lithuanian News Clustering // Proceedings of the Information Society and University Studies 2020, Kaunas, Lithuania, April 23, 2020. — Vol. 2698 of CEUR Workshop Proceedings. — CEUR-WS.org, 2020. — P. 46–53. — Access mode: `http://ceur-ws.org/Vol-2698/p08.pdf`.

[35] Stop clickbait: Detecting and preventing clickbaits in online news media / Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, Niloy Ganguly // 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) / IEEE. — 2016. — P. 9–16.

[36] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems. — Vol. 27. — Curran Associates, Inc., 2014. — Access mode: `https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf`.

[37] Topic Detection and Tracking Pilot Study: Final Report / J. Allan, J. Carbonell, G. Doddington et al. // Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. — Lansdowne, VA, USA, 1998. — Feb. — P. 194–218. — 007.

[38] Universal sentence encoder / Daniel Cer, Yinfei Yang, Sheng-yi Kong et al. // arXiv preprint arXiv:1803.11175. — 2018.

[39] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // CoRR. — 2019. — Vol. abs/1911.02116. — 1911.02116.

[40] Using Generative Pretrained Transformer-3 Models for Russian News Clustering and Title Generation tasks / Tikhonova, Pisarevskaya, Shliazhko, Shavrina // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[41] Voropaev Pavel, Sopilnyak Olga. Comparison of news clustring methods. — 2020.

[42] Voropaev Pavel, Sopilnyak Olga. Transformer-based Embeddings for Russian News Clustering and Headline Selection // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[43] Zhang Tian, Ramakrishnan Raghu, Livny Miron. BIRCH: an efficient data clustering method for very large databases // ACM sigmod record. — 1996. — Vol. 25, no. 2. — P. 103–114.

[44] mT5: A massively multilingual pre-trained text-to-text transformer / Linting Xue, Noah Constant, Adam Roberts et al. // CoRR. — 2020. — Vol. abs/2010.11934. — 2010.11934.