Automated Minuting on DumSum Dataset

Eugene Borisov
NTR Labs and Higher IT School
of Tomsk State University
Tomsk, Russia
eborisov@ntr.ai

Nikolay Mikhaylovskiy
NTR Labs and Higher IT School
of Tomsk State University
Moscow, Russia
nickm@ntr.ai

Abstract

Meeting minutes are short texts summarizing the most important outcomes of a meeting. The goal of this work is to develop a module for automatic generation of meeting minutes based on a meeting transcript text produced by an Automated Speech Recognition (ASR) system. We consider minuting as a supervised machine learning task on pairs of texts: the transcript of the meeting and its minutes. No Russian minuting dataset was previously available. To fill this gap we present DumSum - a dataset of meetings transcripts of the Russian State Duma and City Dumas, complete with minutes. We use a two-staged minuting pipeline, and introduce semantic segmentation that improves ROUGE and BERTScore metrics of minutes on City Dumas meetings by 1%-10% compared to naive segmentation.

Keywords: Minuting, Summarization

Автоматическое протоколирование на наборе данных думских заседаний DumSum

Аннотация

Протокол собрания представляет собой короткий текст, резюмирующий его наиболее важные итоги. Целью данной работы является разработка модуля автоматического протоколирования на основе текста стенограммы собрания, созданного системой автоматического распознавания речи (ASR). Протоколирование рассмотрено как задача машинного обучения с учителем на парах текстов: стенограмма встречи и ее протокол. Ранее не было русскоязычного набора данных протоколирования. Чтобы восполнить этот пробел, представлен DumSum - датасет стенограмм заседаний Государственной Думы и нескольких Городских Дум России с протоколами. Использован двухэтапный конвейер протоколирования и предложена семантическая кластеризация, которая улучшает показатели ROUGE и BERTScore автоматических протоколов заседаний Городских Дум на 1%-10% по сравнению с наивной кластеризацией.

Ключевые слова: Автоматическое протоколирование, Суммаризация

1 Introduction

Discussions and meetings are an integral part of any human activity that involves a group of people. On important meetings, an audio recording is often made, and specially appointed people create a brief summary of the most important things that happened at the meeting. This process is quite laborious.

The ability to produce high-quality documentation of business meetings decisions without allocating additional human resources can improve the productivity of the organizations. This way important points and decisions made will not be lost due to an information overflow. Thus, automated minuting of business meetings is becoming an increasingly desirable solution.

An automated minuting system can be useful not only for businesses but also for government agencies and educational institutions. Hundreds of meetings are held daily, and the ability to automatically generate a summary of the most important decisions made can significantly reduce the time and resources spent on documenting. Thanks to an automatic minuting system, meeting participants can focus on important points without spending time on note-taking.

Name	Transcripts	Domain	Compression ratio, %
ELITR	179	project meetings	95.65
SamSum	16369	dialogues from messengers	81.12
DialogSum	13460	conversations from real life	82.3
DumSum	22647	meetings of the State and regional Dumas	70.77

Table 1: Datasets

Duma	Avg. # tokens (transcript)	Avg. # tokens (minutes)	# meetings
Moscow City	2612	259	7113
State	2182	91	13092
Kirov	1184	90	435
Tomsk	779	562	1053
Tyumen	764	474	924
Samara	708	208	139

Table 2: Key stats of DumSum dataset

The goal of this work is to develop a module for automatic generation of meeting minutes based on a meeting transcript text produced by an Automated Speech Recognition (ASR) system.

2 Datasets

The source for our automated minuting module are ASR transcript texts. Each transcript text consists of a sequence of utterances of the meeting participants. We consider minuting as a supervised machine learning task on pairs of texts: the transcript of the meeting and its minutes. Such datasets are available in English:

- ELITR Minuting Corpus - a dataset of meeting transcripts and minutes (Nedoluzhko et al., 2022).
- SamSum - a dataset of messenger dialogues with their summaries (Gliwa et al., 2019).
- DialogSum - a dataset of dialogues with their summaries (Chen et al., 2021b).

For Russian, there are summarization datasets for the news domain only. The nature of news and meeting transcripts differs too much. The news abstract is largely contained in the first few sentences of the news, while dialogue minuting requires information flowing from the beginning of the discussion to the end (Chen et al., 2021a). No Russian minuting dataset was previously available. To fill this gap we present DumSum - a dataset of meetings transcripts of the Russian State Duma and City Dumas, complete with minutes. The datasets are compared in Table 1. The summary compression ratio θ in the Table 1 is calculated using the following formula:

$$\theta = 1 - \frac{T_A}{T_T} * 100, \tag{1}$$

where T_A is the number of tokens in the abstract and T_T is the number of tokens in the transcript. Thus, the smaller the abstract compared to the original transcript text is, the closer the θ is to 100%.

DumSum dataset was collected by scraping the public websites of the respective Dumas and includes the proceedings of the Dumas listed with its key statistics in Table 2.

3 Methods

All Transformer (Vaswani et al., 2017) language models have a limit on the size of the input context window and do not work well with long texts, such as transcripts of long meetings. Thus, to make

it possible to apply neural networks to the transcript text summarization, we decompose the task of minuting into two subtasks:

- Text Segmentation - dividing the transcript text into segments containing information on a single topic.
- Segment Summarization - generating an abstract of the transcript segment.

As a baseline solution, the naive segmentation proposed by the winners of the AutoMin 2021 competition (Shinde et al., 2021) was adapted for Russian. The naive segmentation involves dividing the transcript text into segments according to the size of the input context window used by the summarization model. (Shinde et al., 2021) used the multilingual MBART, retrained on the English news summarization dataset XSum (Narayan et al., 2018) and SamSum.

Instead of the naive segmentation described above, we suggest to use clusterization of utterances, to obtain a higher quality reporting. The pipeline is as follows:

- For utterances vectorization, the transformer paraphrase-multilingual-MiniLM-L12-v2 from the sentence transformers library (Reimers and Gurevych, 2019) was used. Each utterance was vectorized sequentially using the Mean Pooling (Reimers and Gurevych, 2019): initially, each utterance is broken down into sentences, then, using Mean Pooling, a vector of sentences is obtained, finally, the average of the sentence vectors is taken as the utterance vector.
- For dimensionality reduction, the UMAP (Uniform Manifold Approximation and Projection) algorithm was used (McInnes et al., 2018). The resulting compressed vector representations retain the necessary information to create clusters of semantically similar utterances. Thus, in the clustering of utterances, the use of UMAP allows you to preserve the quality of the segments obtained by clustering, while generally increasing the speed of segmentation due to working with lower-dimensional vectors.
- For clustering the obtained utterance vectors, the density-based HDBSCAN algorithm (Campello et al., 2013) is used. It allows to detect clusters in data without knowing their exact number initially, and is also resistant to noise and outliers, which allows to filter out utterances that are not relevant to the topics of discussion at the segmentation level. The BERTopic library (Grootendorst, 2022) was used to implement the clustering algorithm in the semantic segmentation module.
- Transcript Segments Summarization. Similarly to (Shinde et al., 2021) we use MBART for abstractive summarization, but train it on more relevant corpora as described below.

4 Experiments

4.1 Metrics

The key indicators of the effectiveness of a text summarization algorithm we use are the ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020). To calculate the metrics, we used the ROUGE and BERTScore adaptation for the Russian language (Gusev, 2020).

4.2 Training

Deep neural network models were created using PyTorch. The weights of the pretrained models were loaded from the HuggingFace model hub. The razdel library from the Natasha project was used to split utterances into sentences.

The models were trained using a server using AMD EPYC 7313 16-Core @ 3.00GHz with two NVIDIA RTX 3090 GPUs with 24 GB of VRAM each. We created a training set from DialogSum and SamSum datasets automatically translated from English into Russian via Google Translate API. A validation set was created from the automatically translated ELITR dataset cleared of unnecessary tags and brought to the expected format.

We used MBART finetuned for summarization on Gazeta news summarization corpus (Gusev, 2020) as a base model. It was then finetuned for 4 epochs with a batch size of 2. AdamW (Loshchilov and Hutter, 2018) optimizer was used with the cross-entropy loss function. Figure 1 is a graph of the loss function during the training of the summarization model on SamSum + DialogSum translated datasets.

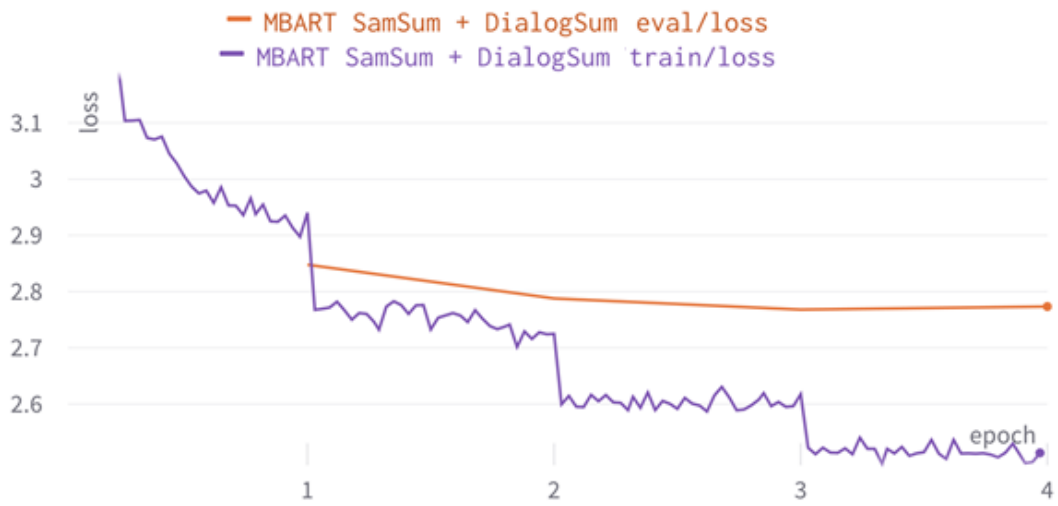

Figure 1: Loss graph

Model	Rouge1-F	Rouge2-F	RougeL-F	BERT-Score
first-line-baseline	0.2037	0.0450	0.1606	0.6408
rut5-base-absum	0.2095	0.0545	0.1641	0.7108
mbart ru sum gazeta	0.2446	0.0597	0.1879	0.7251
mbart ruDialogSum	**0.4135**	0.1803	**0.3560**	0.7847
mbart samdialogsum	0.4110	**0.1844**	0.3509	**0.7873**

Table 3: DialogSum validation results

4.3 Results on translated SamSum and DialogSum

After training, the resulting dialogue summarization model was tested on withheld validation samples from automatically translated SamSum and DialogSum datasets. For a comparison, metrics of other models on the same datasets were also calculated: a first line baseline often used in news summarization, summarization based on ruT5 pretrained model based on (Raffel et al., 2020), MBART finetuned for summarization on Gazeta news summarization corpus (Gusev, 2020), and, for the sake of ablation, MBART finetuned only on DialogSum for the same number of epochs (without addition of SamSum translated dataset).

The results on SamSum and Dialogsum validation sets are presented in Table 3 and Table 4. It is clearly seen that adding SamSum to DialogSum slightly improves performance on DialogSum, but significantly - on SamSum. Both are significantly better in terms of both ROUGE and BERTSore than baselines and alternative solutions not trained on dialog datasets. We can conclude these datasets comprise different domains of dialogues and minutes.

4.4 Comparing segmentation approaches on translated ELITR

We preprocessed the English and Czech parts of the ELITR dataset and translated it using the Google Translate API. All PERSON tags were replaced with example names so that the texts of the transcripts felt more like real dialogues, rather than synthetic ones. We compared naive and semantic segmentation approaches with and without UMAP dimensionality reduction on this dataset. In all the cases we have used MBART finetuned on SamSum + DialogSum translated datasets.

Table 5 shows the performance of the approaches listed above on the English part of the ELITR dataset, while Table 6 – on the Czech part. One can see that in the domain of meetings of distributed teams most similar to day-to-day work discussions, semantic segmentation did not provide significant improvement.

Model	Rouge1-F	Rouge2-F	RougeL-F	BERT-Score
first-line-baseline	0.2896	0.1895	0.2439	0.7602
rut5-base-absum	0.146	0.0445	0.1091	0.6871
mbart ru sum gazeta	0.1664	0.0400	0.1354	0.6736
mbart ruDialogSum	0.2260	0.0786	0.1940	0.7230
mbart samdialogsum	**0.4687**	**0.3312**	**0.4120**	**0.7776**

Table 4: SamSum validation results

Model	Rouge1-F	Rouge2-F	RougeL-F	BERT-Score
Naive segmentation	**0.1977**	**0.0375**	**0.1624**	**0.6806**
Semantic Segmentation	0.1791	0.0339	0.1370	0.6768
Semantic Segmentation with UMAP	0.1771	0.0341	0.1431	0.6304

Table 5: Performance metrics on the Engilsh part of ELITR test set

On the English translated part of ELITR, the semantic segmentation works worse than the naive one, on the Czech side, on the contrary, it is better. The effect of the UMAP dimensionality reduction is also mixed.

4.5 Comparing segmentation approaches on DumSum

Similarly to the above we have compared the segmentation approaches on DumSum dataset. Table 7 shows the performance of the approaches on different parts of DumSum. Thus, in the domain of City Duma meetings, the improvement semantic segmentation provides over the naive segmentation is 1–3%, although the metric improvement on the meetings of the Samara Duma is over 10%. On the other hand, there is no improvement in metrics on the State Duma meetings.

5 Conclusion

The described approach to autominuting is limited to generating only a summary of the meeting transcript. Full-fledged minutes of the meeting should include highlighting the problems discussed during the negotiations and the decisions made. We see this path as moving from meeting summarization task to question answering over the whole meeting transcript. This is the future work.

References

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. // Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, *Advances in Knowledge Discovery and Data Mining*, P 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wang Chen, Piji Li, Hou Pong Chan, and Irwin King. 2021a. Dialogue summarization with supporting utterance flow modelling and fact regularization. *Knowledge-Based Systems*, 229:107328.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. DialogSum: A real-life scenario dialogue summarization dataset. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 5062–5074, Online, August. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. // *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, P 70–79, Hong Kong, China, November. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Model	Rouge1-F	Rouge2-F	RougeL-F	BERT-Score
Naive Segmentation	0.1683	0.0285	**0.1261**	0.6480
Semantic Segmentation	**0.1693**	**0.0306**	0.1211	0.6525
Semantic Segmentation with UMAP	0.1611	0.0206	0.1236	**0.6657**

Table 6: Performance metrics on the Czech part of ELITR test set

Duma	Rouge1-F	Rouge2-F	RougeL-F	BERT-Score
	Naive segmentation			
State	**0.2221**	**0.1063**	**0.1858**	0.6861
Moscow City	0.2856	**0.1258**	0.2199	**0.6842**
Tomsk	0.2851	0.1406	0.2149	0.6668
Tyumen	0.1447	0.0418	0.1116	**0.585**
Kirov	0.2506	0.1311	0.208	**0.639**
Samara	0.3242	0.1823	0.2412	0.6114
mean	0.25205	0.1213	0.1969	0.6454
	Semantic segmentation			
State	0.21	0.0956	0.1783	**0.6914**
Moscow City	**0.2938**	0.1169	**0.2387**	0.6816
Tomsk	**0.2996**	**0.1745**	**0.2321**	**0.6958**
Tyumen	**0.1545**	**0.0462**	**0.1157**	**0.585**
Kirov	**0.2718**	**0.1666**	**0.2227**	0.6185
Samara	**0.4318**	**0.2586**	**0.3667**	**0.7128**
mean	**0.2769**	**0.1725**	**0.2257**	**0.6641**

Table 7: Comparing segmentation approaches on DumSum.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. // Andrey Filchenkov, Janne Kauttonen, and Lidia Pivovarova, *Artificial Intelligence and Natural Language*, P 122–134, Cham. Springer International Publishing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. // *Text Summarization Branches Out*, P 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. // *International Conference on Learning Representations*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. // *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, P 1797–1807, Brussels, Belgium, October-November. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR Minuting Corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. // *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France, June. European Language Resources Association (ELRA). In print.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan.

Nils Reimers and Iryna Gurevych. 2019. 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), P 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team ABC @ AutoMin 2021: Generating Readable Minutes with a BART-based Automatic Minuting Approach. // *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, P 26–33.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // *Advances in Neural Information Processing Systems*, P 5999–6009.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. // *International Conference on Learning Representations*.