Team NTR @ AutoMin 2023: Dolly LLM Improves Minuting Performance, Semantic Segmentation Doesn't

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

This paper documents the approach of Team NTR for the Second Shared Task on Automatic Minuting (AutoMin) at INLG 2023. 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 (Task A). We consider minuting as a supervised machine learning task on pairs of texts: the transcript of the meeting and its minutes. We use a two-staged minuting pipeline that consists of segmentation and summarization. We experiment with semantic segmentation and multi-language approaches and Large Language Model Dolly, and achieve Rouge1-F of 0.2455 and BERT-Score of 0.8063 on the English part of ELITR test set and Rouge1-F of 0.2430 and BERT-Score of 0.8332 on the EuroParl dev set with the submitted Naive Segmentation + Dolly7b pipeline.

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.

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 (AutoMin 2023 Task A, (Ghosal et al., 2022b, 2023)).

2 Related work

Meeting summarization as a scientific problem came to light in the early 2000s (Ghosal et al., 2022a). ISCI Meeting Project (Morgan et al., 2001; Janin et al., 2004) resulted, among other outcomes, in creating ICSI Meeting Corpus of audio recorded from informal, natural, and even sometimes impromptu meetings (Janin et al., 2003). About simultaneously, Klaus Zechner's work on summarization of meeting speech and dialogues (Zechner, 2002) helped to shape the investigations in this topic further. Augmented Multi-party Interaction (AMI) project followed soon, producing The AMI Meeting Corpus (McCowan et al., 2005).

It has been recognized early on that for a wide spectrum of applications identifying and including action items into minutes delivers the key business value (Purver et al., 2007). Detecting decisions in multi-party dialogues happened to be as important for the minuting (Fernández et al., 2008b,a; Bui et al., 2009). The CALO Meeting Assistant System (Riedhammer et al., 2010) that appeared soon after that was an important step, but the overall level of NLP at the time limited (with a few exceptions, for example, (Wang and Cardie, 2012; Liu et al., 2018)) proliferation of minuting research and applications

Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges, pages 132–137
September 11–15, 2023. ©2023 Association for Computational Linguistics

until recently.

Scientific interest to minuting reemerged about 2018, sparked both by an important review by Nedoluzhko and Bojar (Nedoluzhko and Bojar, 2019) and overall technology readiness. The interest came into a full swing in 2021 when many works have appeared (Koay et al., 2021; Shang, 2021; Fu et al., 2021; Chen and Yang, 2021; Fabbri et al., 2021; Zou et al., 2021; Cho et al., 2021; Chen et al., 2021; Zhong et al., 2021) and the first AutoMin competition was held at INTERSPEECH (Ghosal et al., 2022a).

The works that are the most close to ours are likely those by AutoMin 2021 winners (Shinde et al., 2021) and Borisov and Mikhaylovskiy (Borisov and Mikhaylovskiy, 2023). The authors of the former use a BART model and train it on the SAMSum dialogue summarization dataset. Their pipeline first splits the given transcript into blocks of smaller conversations, eliminates redundancies with a specially-crafted rule-based algorithm, summarizes the conversation blocks, retrieves the block-wise summaries, cleans, structures, and finally integrates the summaries to produce the meeting minutes. The authors of the latter introduce a Russian minuting dataset and use an approach similar to (Shinde et al., 2021). They also introduce semantic segmentation that improves ROUGE and BERTScore metrics of minutes on the above dataset by 1%-10% compared to naive segmentation.

3 Datasets

Two main datasets are considered in the AutoMin 2023 Task A:

- ELITR Minuting Corpus – a dataset of meeting transcripts and minutes (Nedoluzhko et al., 2022).
- EuroParlMin v1.0, introduced specifically for AutoMin 2023 (Ghosal et al., 2023)

In addition, we experiment with the following corpora:

- SamSum – a dataset of messenger dialogues with their summaries (Gliwa et al., 2019).

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 θ is to 100%.

4 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 Transformer-based models to the transcript text summarization, we, similarly to the winners of the AutoMin 2021 competition (Shinde et al., 2021) decompose the task of minuting into two subtasks:

- Text Segmentation – dividing the transcript text into segments of reasonable size.
- Segment Summarization – generating an abstract of the transcript segment.

In addition to the naive segmentation just fitting the chunk to the model's window size, we explore semantic segmentation in a hope to obtain higher quality reporting. The pipeline for the semantic segmentation is as follows:

- For utterances vectorization, the transformer all-MiniLM-L6-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

Name	Transcripts	Domain	Compression ratio, %
ELITR	179	project meetings	95.65
EuroParlMin Dev	187	corpus of European Parliament debates	53.08
SamSum	16369	dialogues from messengers	81.12

Table 1: Datasets

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 2: Segmentation methods performance metrics on the Engilsh part of ELITR test set

(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. We explore several models for abstractive summarization, as described below.

5 Experiments

5.1 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).

5.2 Comparing segmentation approaches on ELITR English test set

We compared naive and semantic segmentation approaches with and without UMAP dimensionality reduction on this dataset. In all the cases we have used open source version of MBART finetuned on SamSum dataset.[1]

Table 2 shows the performance of the approaches listed above on the English test part of the ELITR dataset. 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. The semantic segmentation works worse than the naive one. The effect of the UMAP dimensionality reduction is also mixed.

5.3 Experiments with a LLM

In addition to MBART, we have tested a large pretrained language model – Dolly v2 7b (Conover et al., 2023), chosen for its permissive license and competitive performance. We have tried the prompts: ”Summarize” and ”Briefly extract Key Points from the meeting transcript”, and settled for the second as it provided somewhat better performance and more interesting texts. Table 3 shows the difference between two prompts on EuroParl dev set.

Table 4 provides a comparison of Dolly with MBART on English part of ELITR test set. We have also tested Dolly v2 7b on EuroParl dev set, the results provided in the Table 5

Given the above results, we have chosen to submit Naive Segmentation + Dolly results.

6 Conclusion and Future Work

In this work, we describe our system run for the second AutoMin shared Task A on automatic minuting. Our proposed system leverages a pretrained Large Language Model Dolly to generate readable minutes from multi-party meeting proceedings. In the future, we plan to implement similar pipelines for different languages, including low-resource ones.

[1] https://huggingface.co/philschmid/distilbart-cnn-12-6-samsum

Dolly v2 7b Promt	**Rouge1-F**	**Rouge2-F**	**RougeL-F**	BERT-Score
“Summarize”	**0.2465**	**0.0751**	**0.1927**	0.8251
“Briefly extract Key Points from the meeting transcript”	0.2430	0.0694	0.1843	**0.8332**

Table 3: Performance metrics on the EuroParl dev set with different promts.

Model	**Rouge1-F**	**Rouge2-F**	**RougeL-F**	**BERT-Score**
BERTopic + MBART	0.244	**0.0309**	**0.1756**	0.7999
Naive + MBART	0.2022	0.0171	0.132	0.8019
Naive + Dolly7b	**0.2455**	0.0294	0.1656	**0.8063**

Table 4: Performance metrics on the English part of ELITR test set

Model	**Rouge1-F**	**Rouge2-F**	**RougeL-F**	**BERT-Score**
Naive + MBART	0.1539	0.0522	0.0843	0.8392
Naive + Dolly7b	**0.2430**	**0.0694**	**0.1843**	0.8332

Table 5: Performance metrics on the EuroParl dev set

References

Eugene Borisov and Nikolay Mikhaylovskiy. 2023. Automated Minuting on DumSum Dataset. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2023”*, Online.

Trung H. Bui, Matthew Frampton, John Dowding, and Stanley Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009 Conference: 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, September, pages 235–243.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Wang Chen, Piji Li, Hou Pong Chan, and Irwin King. 2021. Dialogue summarization with supporting utterance flow modelling and fact regularization. *Knowledge-Based Systems*, 229:1–29.

Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. StreamHover: Livestream transcript summarization and annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM.

Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 6866–6880.

Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. 2008a. Identifying relevant phrases to summarize decisions in spoken meetings. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, September, pages 78–81.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008b. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, SIGDIAL 2008*, June, pages 156–163.

Xiyan Fu, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Changlong Sun, and Zhenglu Yang. 2021. RepSum: Unsupervised dialogue summarization based on replacement strategy. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pages 6042–6051.

Tirthankar Ghosal, Ondřej Bojar, Marie Hledíková, Tom Kocmi, and Anja Nedoluzhko. 2023. Overview of the Second Shared Task on Automatic Minuting (AutoMin) at INLG 2023. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2022a. Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021. In *Proceedings of Interspeech 2021*, September, pages 1–25.

Tirthankar Ghosal, Marie Hledíková, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022b. The second automatic minuting (AutoMin) challenge: Generating and evaluating minutes from multi-party meetings. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 1–11, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

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

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Mac, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. THE ICSI MEETING PROJECT : RESOURCES AND RESEARCH. *Cognitive Science*, 2004(September).

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A sliding-window approach to automatic creation of meeting minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 68–75, Online. Association for Computational Linguistics.

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

Hui Liu, Xin Wang, Yuheng Wei, Wei Shao, Jonathan Liono, Flora D. Salim, Bo Deng, and Junzhao Du. 2018. ProMETheus: An intelligent mobile voice meeting minutes system. In *ACM International Conference Proceeding Series*, November, pages 392–401.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, Dennis Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140. Noldus Information Technology.

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

Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. 2001. The meeting project at ICSI. In *Proceedings of the First International Conference on Human Language Technology Research*.

Anna Nedoluzhko and Ondřej Bojar. 2019. Towards automatic minuting of meetings. In *CEUR Workshop Proceedings*, volume 2473, pages 112–119.

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. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France. European Language Resources Association (ELRA). In print.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 18–25, Antwerp, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Korbinian Riedhammer, G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tür, J. Dowding, and F. Yang. 2010. The CALO Meeting Assistant System. *IEEE Transactions on Audio, Speech and Language Processing*, 2010(18(6)).

Guokan Shang. 2021. *Spoken Language Understanding for Abstractive Meeting Summarization*. Ph.D. thesis, l'École Doctorale de l'Institut Polytechnique de Paris.

Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team ABC @ AutoMin 2021: Generating Readable Minutes with a BART-based Automatic Minuting Approach. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 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. In *Advances in Neural Information Processing Systems*, pages 5999–6009.

Lu Wang and Claire Cardie. 2012. Focused meeting summarization via unsupervised relation extraction. In *SIGDIAL 2012 - 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, July, pages 304–313.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

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

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.