Unsupervised Topic Segmentation of Meetings

Zhilkin Fedor

January 2023

Abstract

This project report describes the existing approaches to solving the problem of segmentation of meetings by topic and the implementation of the BERT Embeddings approach. The data and the solution's code are distributed through GitHub in the follow- ing repository: `https://github.com/Feodoros/Unsupervised-Topic-Segmentation-of-Meetings`.

1 Introduction

The average employee attends 62 company meetings per month, and many of these meetings are recorded but not used again. These recordings present an opportunity for increased productivity and transparency through topic segmentation, which is the process of dividing text into topically-coherent segments. However, topic segmentation of meetings is challenging due to the noisy nature of meeting transcripts and lack of ground truth data. In this report we implement approach of "Unsupervised Topic Segmentation of Meetings with BERT Embeddings" [Solbiati et al., 2021] and compare results on ICSI Meeting Corpus [Janin et al., 2003], AMI Meeting Corpus [Mccowan et al., 2005] and own datasets.

2 Related Work

Recent advancements in topic segmentation of written text mostly use bidirectional-LSTM embeddings, such as combining a BiLSTM with a pointer network [li et al., 2018], stacked BiLSTMs with attention [Badjatiya et al., 2018], and a custom LSTM architecture [Barrow et al., 2020]. Topic segmentation of spoken language is more challenging due to the added complexity of the underlying ASR system, which focuses on either monologue or dialogue data. Recent advancements in monologue data include neural-based architectures such as TCNs [Zhang and Zhou, 2019] and Bi-LSTMs [Sehikh et al., 2017], with large labeled training data mainly from broadcast news transcripts. Multi-party dialogue speech data, mainly from meeting transcripts, has not yet benefited from neural network advancements and mostly relies on measuring similarity/coherence between sentences to detect topic changes. Sentence embeddings have been used to

extract semantic similarity, with BERT [Devlin et al., 2018] and SentenceBERT [Reimers and Gurevych, 2019] being popular choices for deriving semantically meaningful sentence embeddings.

3 Model Description

We experiment with two different methods for representing sentences in order to extract semantic similarity.

3.1 BERT

We use RoBERTa [Liu et al., 2019], a pre-trained model, as the first approach to extract semantic similarity. RoBERTa is a configuration of BERT that is trained with the Masked Language Modelling objective on five large English language corpora, totaling over 160GB of text. We extract fixed features from the pre-trained model without additional fine-tuning by taking the max pooling of the second to last layer. RoBERTa is trained with L = 12 and H = 768, where L is the number of layers and H is the size of the hidden layer, resulting in an N*H embedding vector for a sentence of N words. We chose the second to last layer as it contains more semantic information.

3.2 Sentence-BERT

For our second approach, we use SentenceBERT [Reimers and Gurevych, 2019], the current state-of-the-art in sentence representation. SentenceBERT is pre-trained on the SNLI dataset [Bowman et al., 2015] and we extract fixed-size sentence embeddings by taking the mean of all the output vectors, similar to the method used for BERT.

3.3 Max pooling

Our extraction architecture is designed to handle noisy speech data [Shriberg, 2005], such as ASR miss-transcriptions, disfluencies, and turn-taking. To remove words that have limited semantic value, we use a max pooling operation multiple times to extract only the words that have high semantic value from an utterance.

3.4 Segmentation Scheme

After obtaining a valid sentence embedding, a common approach is to train a supervised classifier to perform sequence labelling, such as using TCNs [Zhang and Zhou, 2019]. However, we take a different approach by using an unsupervised method that doesn't require any labeled training data. Our method is a modified version of TextTiling [Hearst, 1997]. TextTiling detects topic changes by using a similarity score based on word frequencies, whereas we use a new similarity score based on BERT embeddings.

1. Compute BERT embeddings for each utterance S_i in the meeting transcript.

2. Divide the meeting corpus into blocks of utterances $S_i, , S_k$, and perform block-wise max pooling to extract embedding R_i for each block.

3. Compute the cosine similarity, sim_i, between adjacent blocks R_i and R_{i+1}, where sim_i represents the semantic similarity between two blocks separated at utterance S_i.

4. Identify the topic boundaries as pairs of blocks R_i and R_{i+1} with semantic similarity sim_i below a certain threshold. We obtain a sequence of topic changes $T = \{i \in [0, M] | sim_i < \mu_s - \sigma_s\}$, where μs and s are the mean and variance of the sequence of block similarities sim_i.

4 Dataset

To evaluate the effectiveness of our model, we conduct experiments using two major collections of meeting data that have been recently produced. These include the ICSI Meeting Corpus [Janin et al., 2003], which contains 75 recorded and transcribed meetings with topic segmentation annotations, and the AMI Meeting Corpus [Mccowan et al., 2005], which includes 100 hours of recorded and transcribed meetings also with topic segmentation annotations. Both datasets have a hierarchical structure for topic annotation, but for the purpose of this paper, we focus on the top-level meeting changes, using linear topic segmentation. Since in practical applications of meeting segmentation, labeled data may not be available due to the complexity of the annotation task, our unsupervised evaluation methodology is representative of real-world scenarios.

In addition, a dataset consisting of about 150 meeting recordings in Russian was collected and marked up manually. The records were collected by collecting records of meetings of different companies.

5 Experiments

5.1 Metrics

We use two commonly used metrics, Pk [Beeferman et al., 1999] and $WinDiff$ [Pevzner and Hearst, 2002], to measure the performance of our model. Both of these metrics involve sliding a fixed window over the document, and comparing the predicted segmentation with the reference segmentation from the annotations to compute the probability of segmentation error

5.2 Experiment Setup

First, we need to implement paper "Unsupervised Topic Segmentation of Meetings with BERT Embeddings" [Solbiati et al., 2021]. After we try different back-

bones and configuration of this model:

- BERT (RoBERTa base)
- BERT (multilingual-uncased)
- BERT (multilingual-cased)
- XLM (with language embeddings)
- XLM (RoBERTa base)
- SBERT (all-mpnet-base-v2)
- SBERT (paraphrase-multilingual-mpnet-base-v2)

5.3 Baselines

We compare our implementations against other approaches:

- Random
- Even
- TextTiling

TextTiling is a technique for segmenting text into coherent "tiles" or chunks. It is based on the idea that coherent text will have a similar distribution of words, and uses techniques from computational linguistics and information theory to identify boundaries between segments. The method has been used in various natural language processing tasks such as text summarization, information retrieval, and text understanding.

The Random method for topic segmentation is a technique that randomly assigns a segmentation to a text. It is mainly used as a baseline or a control method in experimental evaluations of other topic segmentation algorithms. The idea is that if a segmentation algorithm performs better than the random method, it suggests that the algorithm is identifying meaningful segments in the text, rather than simply splitting the text in arbitrary places. In practice, the random method for topic segmentation is usually implemented by randomly choosing a set of cut points in the text and dividing the text into segments at those points.

The Even method for topic segmentation is a technique that evenly divides a text into a pre-specified number of segments. It is mainly used as a baseline or a control method in experimental evaluations of other topic segmentation algorithms. The idea is that if a segmentation algorithm performs better than the even method, it suggests that the algorithm is identifying meaningful segments in the text, rather than simply splitting the text in a uniform way. In practice, the even method for topic segmentation is usually implemented by dividing the text into equal-length segments based on the number of segments desired. This method is not as useful as it doesn't consider the context or coherence of the text.

6 Results

We compare baseline solution with our implementation. The table shows that our implementation of the paper shows the best results for all metrics. The results are presented in Tab. 1.

Model	AMI PK	AMI WD	ICSI PK	ICSI WD	Own PK	Own WD
BERT (RoBERTa base)	0.462	0.474	0.482	0.511	0.49	0.53
BERT (multilingual-uncased)	0.449	0.464	0.43	0.456	0.47	0.47
BERT (multilingual-cased)	0.451	0.469	0.423	0.453	0.45	0.48
XLM (with language embeddings)	0.44	0.456	0.443	0.478	0.48	0.49
XLM (RoBERTa base)	0.452	0.47	0.474	0.502	0.5	0.52
SBERT (all-mpnet-base-v2)	0.457	0.48	0.468	0.519	0.49	0.53
SBERT (paraphrase-multilingual)	0.457	0.48	0.467	0.521	0.49	0.5
Random	0.609	0.762	0.645	0.844	0.79	0.83
Even	0.523	0.557	0.614	0.671	0.67	0.74
TextTiling	0.394	0.41	0.384	0.406	0.44	0.45
Paper implementation	**0.339**	**0.334**	**0.336**	**0.349**	**0.35**	**0.4**

Table 1: Results

The Tab. 2 shows example of splitting meeting transcript by topic. Each topic was given a name according to the content of the chapter.

Parts of transcript	Topic (chapter name)
0:00:06 **A**: Я начну запись. Ставлю демонстрацию экрана... ... 0:01:15 **B**: Получается, что обсуждаем новую структуру, далее идем по демо... ... 0:01:46 **A**: Сегодня показываю новый анализ...	Приветствие
0:01:59 **A**: На сайте. Фронтенд поменяем. В случае чего. В первую очередь... ... 0:05:27 **A**: Я получается. Делаю вот этот баг, смещение. Цвет ... 0:14:47 **C**: Если успею. Ок.	Исправление сайта
0:14:56 **C**: Правильно. Давайте попробуем сделать через пупитер... ... 0:17:05 **B**: Точно вряд ли. Можно использовать стрим...	Исправление проблем сервера
0:25:18 **C**: Я заведу багу. Посмотрим. ... 0:27:48 **A**: Всем пока.	Завершение встречи

Table 2: Sample

7 Conclusion

During the study, a test dataset was collected and marked up manually, paper [Solbiati et al., 2021] was implemented and various tests were conducted. The best solution was built into the project and started to be used for meeting processing.

References

[Badjatiya et al., 2018] Badjatiya, P., Kurisinkel, L., Gupta, M., and Varma, V. (2018). Attention-based neural text segmentation.

[Barrow et al., 2020] Barrow, J., Jain, R., Morariu, V., Manjunatha, V., Oard, D., and Resnik, P. (2020). A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, Online. Association for Computational Linguistics.

[Beeferman et al., 1999] Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Mach. Learn.*, 34(1–3):177–210.

[Bowman et al., 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Hearst, 1997] Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

[Janin et al., 2003] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The icsi meeting corpus. pages I–364.

[li et al., 2018] li, J., Sun, A., and Joty, S. (2018). Segbot: A generic neural text segmentation model with pointer network.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Mccowan et al., 2005] Mccowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska Masson, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research*.

[Pevzner and Hearst, 2002] Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

[Sehikh et al., 2017] Sehikh, I., Fohr, D., and Illina, I. (2017). Topic segmentation in asr transcripts using bidirectional rnns for change detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 512–518.

[Shriberg, 2005] Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. pages 1781–1784.

[Solbiati et al., 2021] Solbiati, A., Heffernan, K., Damaskinos, G., Poddar, S., Modi, S., and Calì, J. (2021). Unsupervised topic segmentation of meetings with BERT embeddings. *CoRR*, abs/2106.12978.

[Zhang and Zhou, 2019] Zhang, L. and Zhou, Q. (2019). Topic segmentation for dialogue stream. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1036–1043.