Multimodal Topic Segmentation of Podcast Shows with Pre-trained Neural Encoders

Iacopo Ghinassi
i.ghinassi@qmul.ac.uk
Queen Mary University of London
London, UK

Lin Wang
lin.wang@qmul.ac.uk
Queen Mary University of London
London, UK

Chris Newell
BBC R&D
London, UK

Matthew Purver
Queen Mary University of London, London, UK
Jožef Stefan Institute, Ljubljana, Slovenia

ABSTRACT

We present two multimodal models for topic segmentation of podcasts built on pre-trained neural text and audio embeddings. We show that results can be improved by combining different modalities; but also by combining different encoders from the same modality, especially general-purpose sentence embeddings with specifically fine-tuned ones. We also show that audio embeddings can be substituted with two simple features related to sentence duration and inter-sentential pauses with comparable results. Finally, we publicly release our two datasets, the first in our knowledge publicly and freely available multimodal datasets for topic segmentation.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

topic segmentation, multi-modal

ACM Reference Format:
Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023. Multimodal Topic Segmentation of Podcast Shows with Pre-trained Neural Encoders. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3591106.3592270

1 INTRODUCTION AND RELATED WORK

Topic segmentation of broadcast material is a useful and well-known task [26]. Topic segmentation have been traditionally performed on text, but early works combining text and audio also exist [5, 13, 32]. Early attempts used unsupervised approaches [6, 16] that have been successfully applied with text neural embeddings (i.e. sentence embeddings) [15, 30]. Neural audio embeddings have also been used as input for a Bidirectional Long-Short Term Memory (BiLSTM) neural network [17] with success [3]. Performing topic segmentation with neural networks such as BiLSTM [19, 33] and Transformers [23, 24] have recently gained momentum thanks to bigger text datasets [2, 19].

However, while effective models have been developed for audio-visual content such as TV news broadcasts, often using simple classifiers on top of visual features [9], no such features are available for podcast programmes while a renewed interest in automatically segmenting this type of multimedia content has recently emerged [1].

In this context, multimodality has been investigated by fusing word-level text and low-level acoustic features into BiLSTM networks [28, 31]: no attempt to combine neural text and audio embeddings exists. Multimodal experiments are limited by the fact that no freely available, recent dataset for multimodal topic segmentation exists, as the existing ones are either private or not free such as the TDT datasets [20, 34] which is also more than 20 years old.

In our experiments, we use two datasets that we created for multimodal topic segmentation and we release them to fille the existing gap. We use two neural sentence encoders for the text domain and two neural audio encoders for the audio domain to extract features. We try two common approaches in literature on multimodality. Given the relative success of text embeddings alone, we experiment with concatenating simple duration of sentences and pauses in between them to the text-only embeddings to simplify the model.

We show that multimodality can for specific domains lead to significant improvements. Finally, we assess whether the use of neural audio embeddings is really justified or if lower-level acoustic features lead to similar results.

2 METHODOLOGY

2.1 Neural Embedding Models

2.1.1 Sentence Embeddings. We experiment with two different versions of RoBERTa [22], a popular language model that has shown good results for topic segmentation [23]. **RoB** (*RoBERTa Average Pooling*): the base version of RoBERTa [22], which is a 12-layer transformer encoder optimized based on BERT [8]. **Top** (*RoBERTa Topic Segmentation*): RoBERTa fine-tuned on the training dataset with the objective of making sentences from the same topic segment closer in the embedding space. The loss function for fine-tuning is expressed as $\mathcal{L} = ||label_{(i;i+1)} - \frac{e_i \cdot e_{i+1}}{||e_i||_2 \cdot ||e_{i+1}||_2}||_2$, where e_i and e_{i+1} are the embeddings for sentences i and $i+1$; $label_{(i;i+1)}$ equals 0.5 if the two sentences belong to the same segment, and equals -1 otherwise.

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592270

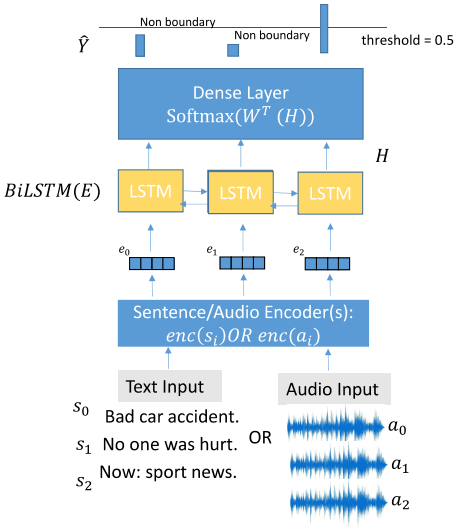

Figure 1: Unimodal model architecture

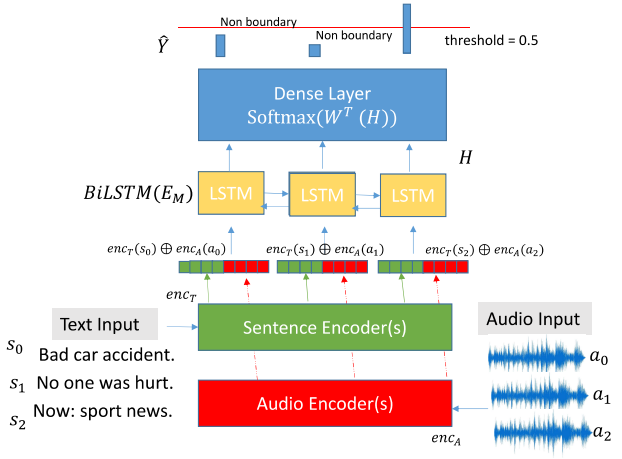

Figure 2: Early Fusion model architecture

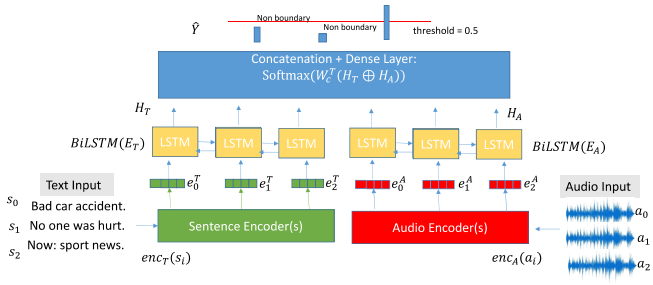

Figure 3: Late Fusion model architecture

For each sentence i, the neural model returns multiple sub-word embeddings, which are then averaged as a single embedding [27], which is represented as e_i^T.

2.1.2 Audio Embeddings. We use two pre-trained audio models for the task of topic segmentation. **XVEC** (*X-Vectors*) was proposed by [29] for speaker diarization. **OP** (*Openl3*) was proposed by [7] for audio classification.

For each sentence i, the neural model extracts multiple sub-embeddings, whose mean and standard deviation are contatenated as a single embedding, represented as e_i^A.

2.2 Topic Segmentation Models

2.2.1 Unimodal. We use BiLSTM, one of the most popular model [17], for unimodal topic segmentation. In the simplest case, defining $BiLSTM$ as a stack of n BiLSTM layers yielding vectors of dimension h, $Softmax$ as the softmax function and $W \in \mathbb{R}^{h \times 1}$ being the weights of the final classification layer we compute the posterior probabilities of each input $\hat{Y} = Softmax(W^T BiLSTM(E))$ where $E := \{e_0, e_1, ..., e_n\}$ is the sequence of embeddings, each corresponding to a sentence, as extracted by the current (sentence or audio) encoder, and the probabilities $\hat{Y} := \{\hat{y}_0, \hat{y}_1, ..., \hat{y}_n\}$ represent the probabilities the model attributed to sentences 0 to n respectively of that sentence being the end of a topical coherent segment.

We also combine different encoders from the same modality. For each sentence i, the embedding is represented as a concatenation of output of two encoders $enc1$ and $enc2$, i.e. $e_i = enc1(i) \oplus enc2(i)$, where $\oplus$ denotes concatenation. Figure 1 shows this visually.

2.2.2 Multimodal with neural audio-embeddings. We experimented with two common modality fusion techniques:

Early Fusion [12] is summarised by the following equation: $\hat{Y} = Softmax(W^T BiLSTM(E_M))$ but in this case, $E_M := \{e_0^M, e_1^M, ..., e_n^M\}$ includes all the embeddings e_i^M corresponding to sentence s_i and audio chunk a_i being obtained as $e_i^M = enc_T(s_i) \oplus enc_A(a_i)$, where enc_T is one or the combination of the two text encoders and enc_A is one or two encoders from the audio modality. Figure 2 shows this visually.

Late Fusion [14] models the two modalities separately with two different BiLSTM networks. Having $E_T := \{e_0^T, e_1^T, ..., E_n^T\}$ as the collection of all the sentence embeddings $e_i^T \in (R)^{d_t}$ extracted from the programme transcript's sentences and $E_A := \{e_0^A, e_1^A, ..., E_n^A\}$ being the collection of all the corresponding audio embeddings $e_i^A \in (R)^{d_a}$, we compute $H_T = BiLSTM(E_T)$ and $H_A = BiLSTM(E_A)$ where $H_T \in (R)^{d_t \times n}$ and $H_A (R)^{d_a \times n}$ are the text and audio hidden representations respectively as encoded by two separate BiLSTM networks. Finally, we concatenate the output of the two networks: $\hat{Y} = Softmax(W_c^T (H_T \oplus H_A))$ with $W_c \in (R)^{d_t + d_a \times 1}$ being the classification layer. Figure 3 shows this visually.

2.2.3 Multimodal with low-level features. We also experiment with augmenting the text-only embeddings with the two low-level features of sentence durations and inter-sentential pauses. In this case, we employ the unimodal model described in the above section, but where each input embedding $e_i = enc_T(s_i) \oplus sd_i \oplus sp_i$, with enc_T being one of the text encoders described before, s_i is the current sentence i in our input transcript and sd_i and ps_i are the sentence duration and inter-sentential pause for sentence i.

Table 1: Details of the two datasets. Datasets details: total number of files (TF), total number of segments (TS), average number of segments per file (ASF), average number of sentences per file (ASpF) and average number of sentences per segment (ASpS).

Dataset	TF	TS	ASF	ASpF	ASpS
NonNewsSBBC	54	393	7.27	491.38	72.04
RadioNewsSBBC	48	561	11.69	346.79	28.93

Table 2: Unimodal results using Sentence (top) and Audio (middle) embeddings, and multimodal results (bottom) with low-level acoustic features (laf). '+' indicates combination of encoders.

Dataset	NonNewsSBBC			RadioNewsSBBC		
	B-F1	B-P	B-R	B-F1	B-P	B-R
RoB	62.56	61.24	72.70	59.86	49.63	81.97
Top	59.61	52.83	73.00	69.85	64.69	78.04
RoB+Top	**71.63**	70.92	**75.71**	**74.61**	68.94	**83.21**
XVEC	55.01	**80.70**	45.08	62.81	54.18	78.59
OP	64.00	72.35	67.86	61.92	**97.41**	45.82
XVEC+OP	62.09	80.62	54.68	62.61	63.57	64.38
RoB+laf	49.60	38.74	**83.30**	**80.01**	**76.94**	84.89
Top+laf	63.44	**71.84**	58.18	76.56	69.26	**88.17**
RoB+Top+laf	**65.52**	65.64	67.99	79.82	74.62	87.26

3 DATA & EXPERIMENTAL SETUP

Datasets. We introduce two new datasets for multimodal topic segmentation, each containing podcast episodes from different programmes available on the BBC Sounds platform. We obtained transcripts automatically by using the open-source Kaldi framework [25] and a private automatic recognition model internally trained at BBC. The model provided the initial sentence segmentation and the corresponding time labels, used to extract text sentences from the transcripts and audio chunks from the associated audio files. The datasets we used are presented below in more details. Given copyright limitations, we release just the extracted embeddings and the relative ground-truth labels from both datasets[1].

NonNewsSBBC: 54 magazine-style radio programmes from BBC Sounds covering different non-news topics. Topic boundaries were manually annotated by experts; audio file length ranges from 20 to 60 minutes approximately.

RadioNewsSBBC: 48 news bulletins from local, national and World Service radio channels by the BBC. Topic boundaries were manually annotated by experts and the length of each audio file range from 9 to 60 minutes approximately.

The statistics of the two datasets are presented in Table 1. For each dataset, we used pre-defined train, validation and test splits, where the validation and test folds are about 15% of the original datasets' sizes and the training set about 70%.

Experimental Setup. Our networks consisted of 2 BiLSTM layers of 256 hidden units per direction each. We used the Adam optimizer [18] with learning rate 0.001; for the Focal Loss function we used $\alpha = 0.9$ and $\gamma = 2$.

[1]https://zenodo.org/record/7825759; https://zenodo.org/record/7821475

Audio files were re-sampled at 16kHz before encoding. Sentences were extracted from the transcripts via the rule-based punkt tokenizer [4], punctuation having been automatically restored while generating the transcriptions.

Given the strong class imbalance, we use the focal loss function [21] in training; this combines class weighting α (to give more weight to the positive class, i.e. topic boundary) and an additional weight γ for samples with probability closer to the decision boundary (0.5).

The models were evaluated with a variant of precision, recall and F1 scores named boundary similarity [11]. For computing the metric we have used the standard *segeval* python library [10]; the resulting accuracy metrics are named B-F1, B-P and B-R and the parameter controlling when a prediction close to a real one is considered a "near miss" is set to $\frac{1}{7}$ of the average segment length of each document.

4 RESULTS

4.1 Unimodal and Multimodal Results

Table 2 shows unimodal results; text-based embeddings (top) usually outperform audio-based ones (middle), for all metrics but B-P, for which XVEC and OP are the best configuration for NonNews and RadioNews Datasets respectively. The two text encoders are not very strong when used by themselves; but combining them gives a noticeable gain, suggesting that they provide non-redundant information. Combining audio encoders does not improve results.

Table 3 shows the multimodal results. Early Fusion tends to outperform the Late Fusion approach in all metrics but recall for both datasets. The high recall and low precision of Late Fusion could be a consequence of this model overfitting on the limited training data due to the bigger size and, as such, being more confident in outputting topic boundaries.

OP+RoB+Top is consistently the best combination in terms of B-F1 and XVEC+RoB is usually the best when looking at B-R, while the ranking yielded by B-P seems more variable.

In terms of comparison with the unimodal settings, the multimodal results seem on average higher for RadioNews dataset, while for the NonNews one this difference is not evident. This is exemplified by the best performing combination from the multimodal experiments (i.e. OP+RoB+Top), which significantly outperforms the best unimodal setting (i.e. RoB+Top) in the RadioNews scenario, while the two settings clearly perform very similarly in the case of NonNews, leading to a statistically insignificant difference.

4.2 Multimodal Results: low-level acoustic features

In this section we explore adding two low-level acoustic features (see Methodology section) to the text-only embeddings.

Correlation Analysis. Figure 4 shows that pauses and durations correlate differently with topic boundaries in the different datasets. The pause correlation is positive in news podcasts, i.e. silences between sentences are a relatively good indicator of topic shifts: in this format an anchorperson often introduces the topics and employs longer intra-topic silences to better separate them. The correlation in magazine-style shows is much lower, indicating more

Table 3: Early Fusion and Late Fusion results for topic segmentation.

	Early Fusion						Late Fusion					
Dataset	**NonNewsSBBC**			**RadioNewsSBBC**			**NonNewsSBBC**			**RadioNewsSBBC**		
	B-F1	B-P	B-R	B-F1	B-P	B-R	B-F1	B-P	B-R	B-F1	B-P	B-R
XVEC+RoB	65.83	85.03	58.81	74.68	67.25	85.99	52.21	44.59	**86.58**	73.48	65.64	**86.50**
OP+RoB	64.35	82.21	55.28	81.58	91.03	74.76	61.12	56.45	78.27	76.18	83.90	71.10
XVEC+OP+RoB	67.82	79.53	66.91	80.02	**96.35**	68.88	59.22	52.98	83.12	79.21	74.42	87.53
XVEC+Top	71.16	83.28	68.77	67.60	85.87	57.03	57.06	47.58	85.77	69.17	60.11	86.59
OP+Top	63.56	77.80	57.07	80.42	91.63	72.71	66.59	60.57	76.13	68.09	65.06	75.39
XVEC+OP+Top	67.83	**88.27**	57.96	72.49	87.34	62.89	61.81	52.86	84.88	71.96	65.38	83.92
XVEC+RoB+Top	60.88	87.65	48.97	68.60	84.59	59.40	54.33	44.22	81.68	76.11	69.94	86.43
OP+RoB+Top	**71.86**	77.80	70.84	**84.48**	91.38	78.79	71.43	67.64	80.55	73.46	71.05	78.22
XVEC+OP+RoB+Top	67.40	86.27	58.79	79.71	92.46	70.83	53.53	45.53	84.4	77.11	74.15	84.77

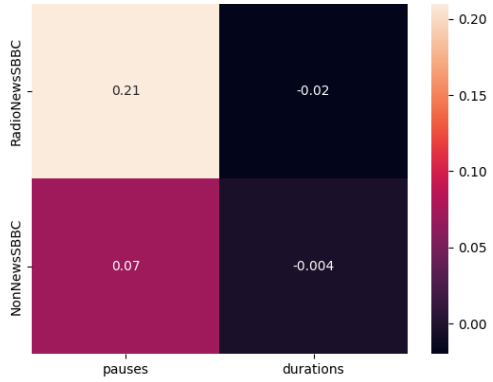

Figure 4: Correlation coefficients of pauses in-between sentences (pauses) and of sentence duration (durations) with topic boundaries.

variation in the shows' formats and how different topics are introduced. The correlation of sentence duration, on the other hand, is weakly negative for RadioNews, and practically null for NonNews.

Experiments. As Table 2 shows (bottom), results confirm the expectations from Figure 4: performance improves in RadioNewsBBC (where pause correlation is stronger), but not in NonNews. In this case, the base RoB model works the best and its B-F1 difference with the best multimodal setting is no longer significant even for RadioNewsSBBC.

The small sizes of the datasets might be the cause of such statistical insignificance and the multimodal results using audio embeddings are still the best even if not significantly.

5 CONCLUSION

Our experiments suggest these conclusions: (1) Multimodality can significantly boost performance but just in specific domains (here, news podcasts). (2) For both datasets and all models used, using OpenL3 and the combination of RoBERTa Average Pooling and RoBERTa fine-tuned for topic segmentation provides the best multimodal setting. (3) For news podcasts, simply concatenating text embeddings with raw sentence durations and inter-sentence pause durations is not significantly worse than adding full audio embeddings.

An open question remains about what audio embeddings seem to convey more than simple low-level acoustic features and if using more advanced fusion techniques they can yield significantly better results. By publicly releasing the two datasets, we hope to foster research towards future better approaches that could make full use of the various information encoded by both text and audio neural embeddings.

ACKNOWLEDGMENTS

We acknowledge financial support from several sources: the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103), and the UK EPSRC via the projects Sodestream (Streamlining Social Decision Making for Improved Internet Standards, EP/S033564/1) and ARCIDUCA (Annotating Reference and Coreference In Dialogue Using Conversational Agents in games, EP/W001632/1).

REFERENCES

[1] Abigail Alexander, Matthijs Mars, Josh C Tingey, Haoyue Yu, Chris Backhouse, Sravana Reddy, and Jussi Karlgren. 2021. Audio Features, Precomputed for Podcast Retrieval and Information Access Experiments. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*. Springer, 3–14.

[2] Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics* 7 (2019), 169–184. https://doi.org/10.1162/tacl_a_00261

[3] Oberon Berlage, Klaus-Michael Lux, and David Graus. 2020. Improving Automated Segmentation of Radio Shows with Audio Embeddings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 751–755. https://doi.org/10.1109/ICASSP40776.2020.9054315

[4] Steven Bird and Ewan Klein. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media Inc.

[5] Shi Sian Cheng and Hsin Min Wang. 2003. A sequential metric-based audio segmentation method via the Bayesian information criterion. In *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH)*.

[6] F Choi. 2000. Linear text segmentation : approaches, advances and applications. In *Proceedings of CLUK 3*.

[7] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3852–3856. https://doi.org/10.1109/ICASSP.2019.8682475

[8] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1.

[9] Bailan Feng, Zhineng Chen, Rong Zheng, and Bo Xu. 2014. Multiple style exploration for story unit segmentation of broadcast news video. *Multimedia Systems* 20 (2014). Issue 4. https://doi.org/10.1007/s00530-013-0350-0

[10] Chris Fournier. 2013. *Evaluating Text Segmentation.* Master's thesis. University of Ottawa.

[11] Chris Fournier. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics*. 1702–1712.

[12] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. 1–6. https://doi.org/10.23919/FUSION45008.2020.9190246

[13] Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse Segmentation of Multi-Party Conversation. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics*. 562–569. https://doi.org/10.3115/1075096.1075167

[14] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* 91 (2023), 424–444. https://doi.org/10.1016/j.inffus.2022.09.025

[15] Iacopo Ghinassi. 2021. Unsupervised Text Segmentation via Deep Sentence Encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content. (5 2021). https://doi.org/10.5281/ZENODO.4744399

[16] Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23 (1997). Issue 1.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[18] Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations (ICLR)*.

[19] Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 2. https://doi.org/10.18653/v1/n18-2075

[20] Wessel Kraaij, Alan F Smeaton, and Paul Over. 2004. TRECVID 2004 - An Overview. *TRECVID 2004 Text Retrieval Conference TRECVID Workshop*. https://doi.org/10.1145/1027527.1027678

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* (2019).

[23] Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence. In *EMNLP*.

[24] Michael Lukasik, Boris Dadachev, Gonçalo Simões, and Kishore Papineni. 2020. Text segmentation by cross segment attention. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4707–4716.

[25] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Hilton Waikoloa Village, Big Island, Hawaii, US). IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[26] Matthew Purver. 2011. Topic Segmentation. In *Spoken Language Understanding*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119992691.ch11

[27] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv* (2020). https://doi.org/10.18653/v1/2020.emnlp-main.365

[28] Imran Sehikh, Dominique Fohr, and Irina Illina. 2018. Topic segmentation in ASR transcripts using bidirectional RNNS for change detection. *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017 - Proceedings* 2018-January. https://doi.org/10.1109/ASRU.2017.8268979

[29] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2018-April. https://doi.org/10.1109/ICASSP.2018.8461375

[30] Alessandro Solbiati, Kevin Hefferman, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised Topic Segmentation of Meetings with BERT Embeddings. *arXiv* (2021).

[31] Emiru Tsunoo, Peter Bell, and Steve Renals. 2017. Hierarchical recurrent neural network for story segmentation. In *Proc. Interspeech*, Vol. 2017-August. https://doi.org/10.21437/Interspeech.2017-392

[32] Gökhan Tür, Andreas Stolcke, Dilek Hakkani-Tür, and Elizabeth Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics* 27 (2001). Issue 1. https://doi.org/10.1162/089120101300346796

[33] Linzi Xing and Giuseppe Carenini. 2021. Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Singapore and Online, 167–177. https://aclanthology.org/2021.sigdial-1.18

[34] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. Van Mulbregt. 1998. A hidden Markov model approach to text segmentation and event tracking. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 1. https://doi.org/10.1109/ICASSP.1998.674435