Exploring the Power of BART and BERT: Context Similarity and Extraction in NLP

By

Hitakshi Chellani
Symbiosis University of Applied Sciences, Indore
E-mail: hitakshi.chellani@gmail.com
Saurav Singh
Symbiosis University of Applied Sciences, Indore
E-mail: sauravsingh9425@yahoo.com
Shakti Mourya
Symbiosis University of Applied Sciences, Indore
E-mail: mouryashakti@gmail.com
Devendra Chouhan
Symbiosis University of Applied Sciences, Indore
E-mail: devendra.chouhan@suas.ac.in

Abstract

In this paper, we propose a new pipeline for extracting context and analyzing semantic similarity from audio or video data. The pipeline consists of four main stages: first, video-to-audio conversion using the Moviepy library; second, speech-to-text transcription using a pre-trained whisper model; third, text summarization using the BART-based "facebook/bart-large-cnn" model from the Hugging Face Transformers library; and finally, similarity analysis using the BERT model from the same library. Our pipeline has numerous applications in fields such as natural language processing, information retrieval, and search engine algorithms. To validate the effectiveness of our pipeline, we conducted experiments on a dataset of audio and video files and demonstrated its high accuracy in context extraction and similarity analysis. Our research provides a useful framework for future studies in natural language processing and information retrieval and highlights the potential of pre-trained models for solving complex NLP tasks.

Keywords: Whisper model; Semantic similarity; Meeting summary; BART; BERT; Textual data extraction

1. Introduction

In today's fast-paced business environment, meetings are a common occurrence in organizations of all sizes. Meetings are essential for effective communication, decision-making, and problem-solving. However, keeping track of the discussions and decisions made during meetings can be challenging, especially when dealing with large volumes of information.

We provide a solution to this problem so that individuals won't have to listen to the entire tape or risk missing the important information while hearing it from others. This way, they can obtain the summary of the meeting they weren't able to attend. Our system involves a pipeline that employs video-to-audio conversion, audio-to-text transcription, text summarization, and text comparison to extract important information from multimedia data.

The initial step of the pipeline involves converting video to audio, which allows for easier processing of the multimedia data. Next, an accurate whisper model is utilized for audio-to-text transcription. The facebook/bart-large-cnn model of Hugging Face is then applied for text summarization in the third step, which can quickly and accurately generate a summary of the text. Finally, the fourth step involves comparing two texts using BERT, a state-of-the-art natural language processing model that can identify similarities and differences between the two texts.

The objective of this research paper is to evaluate the effectiveness of this pipeline in extracting crucial information from multimedia data. To test the accuracy of the pipeline, we will compare its output with manually generated summaries of the same data. Additionally, we will explore the potential applications of this pipeline in various fields such as journalism, social media analysis, and market research.

In summary, this pipeline provides an automated and efficient approach to process multimedia data. This research paper aims to contribute to the development of more accurate and efficient methods for processing multimedia data.

2. Basic Definitions, Preliminaries and Notations

2.1. Natural Language Processing (NLP) [15]: This field of study focuses on the interaction between human language and computers. It involves the development of algorithms and models for processing, understanding, and generating natural language data.

2.2. Information Retrieval (IR): Refers to the process of retrieving information from a set of documents or data based on user queries. It employs techniques such as indexing, searching, and ranking to identify the most relevant documents for a given query.

2.3. Video-to-Audio Conversion: The process of extracting audio content from video data, which can be achieved using libraries such as Moviepy or FFMPEG.

2.4. Speech-to-Text Transcription [14]: The process of converting spoken language into text. This involves utilizing models such as deep neural networks to recognize and transcribe speech.

2.5. Text Summarization: The process of generating a summary of a document or text. This can be done using techniques such as extractive summarization, where important sentences or phrases are selected from the original text, or abstractive summarization, where a new summary is generated using natural language generation techniques.

2.6. Semantic Similarity [10]: A measure of how similar two pieces of text are in terms of their meaning. This is often computed using models such as BERT or Word2Vec, which generate embeddings that capture the semantic meaning of words or phrases.

2.7. Paraphrase Identification: The task of determining whether two pieces of text convey the same meaning, even if the words used are different. This is often accomplished using models such as BERT or Siamese networks.

2.8. BERT: Bidirectional Encoder Representations from Transformers is a pre-trained language model that has achieved state-of-the-art performance on a wide range of natural language processing tasks.

2.9. Hugging Face Transformers [5]: A popular library for working with pre-trained models in natural language processing. The library provides a wide range of pre-trained models and tools for fine-tuning them on specific tasks.

2.10. Pipeline: A sequence of steps or operations performed in a specific order to achieve a particular goal. In this paper, we propose a pipeline for context extraction and similarity analysis from video or audio data.

2.11. Pre-trained Model: A model trained on a vast amount of data that can be used as a starting point for further training or fine-tuning on specific tasks

3. Methodology and Main Work

3.1. **Data Collection:** The first step in the proposed methodology is data collection. This involves collecting video recordings that will be used for further processing. The video recordings can be obtained from various sources, such as public archives or private collections.

3.2. **Video-to-audio conversion using the Moviepy library –** The first stage of our pipeline involves converting video to audio using the Moviepy library. It is a Python library that facilitates video editing, manipulation, and processing. It offers a user-friendly interface for working with video files and is capable of performing a variety of tasks, ranging from basic editing to intricate compositing and special effects. The library is built on top of

NumPy and the Python Imaging Library (PIL), which allows it to be seamlessly integrated with other scientific computing libraries.

3.3. Speech-to-text transcription using a pre-trained whisper model [17] – The second stage of our pipeline involves using a pre-trained Whisper model for speech-to-text transcription. Whisper is an automatic speech recognition (ASR) system that has been trained on a vast dataset of 680,000 hours of multilingual and multitask supervised data collected from the internet. It is capable of transcribing speech in multiple languages and translating them into English. The Whisper architecture is built on a simple end-to-end approach and implemented as an encoder-decoder Transformer. The audio input is divided into 30-second chunks, converted into a log-Mel spectrogram, and then fed into an encoder. A decoder is used to predict the corresponding text caption, interspersed with special tokens that guide the single model to perform various tasks, such as language identification, phrase-level timestamps, multilingual speech transcription, and speech-to-English translation.

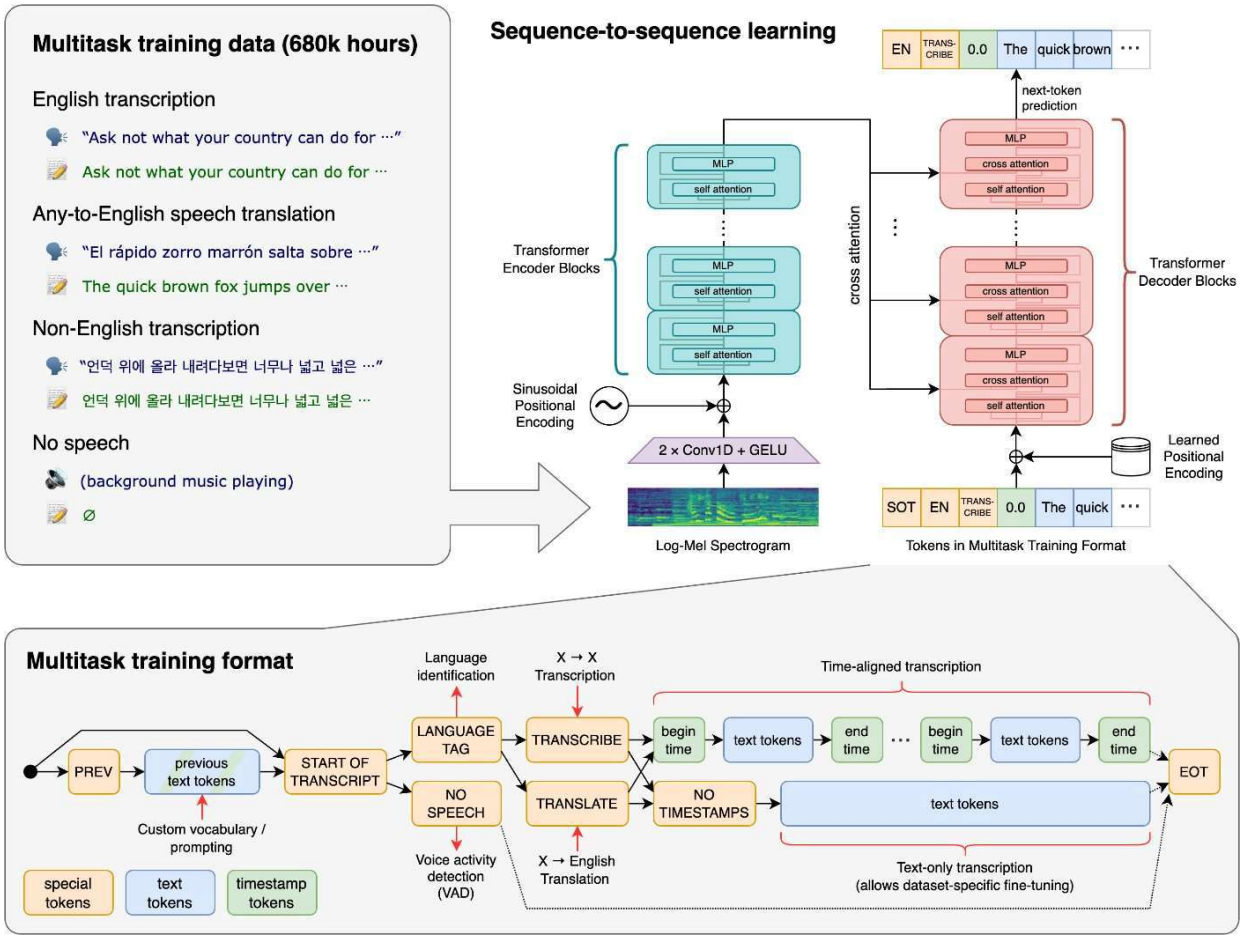


Figure 1. Illustrates an overview of the transformer approach.

3.4. Text summarization using the BART-based "facebook/bart-large-cnn" model – The third phase of out pipeline involves using the facebook/bart-large-cnn summarization model from Hugging Face Transformers library'. BART [1,7,11] (Bidirectional and Auto-Regressive Transformers) is a versatile sequence-to-sequence model introduced by Facebook AI Research. It combines Transformer architecture with techniques from autoregressive language modeling and denoising autoencoders. BART excels in various natural language processing tasks such as text generation, summarization, translation, and question answering. Its unique feature is the ability to handle both autoregressive and bidirectional tasks, making it adaptable and effective for specific applications.

The input text was first converted into tokens, and then the first chunk of tokens was selected. If the length of the chunk exceeded the maximum input length of the model, it was appended to a list of chunks that would be processed separately. To ensure that each chunk remained coherent and meaningful, we trimmed the chunk to a length of 1024 tokens and located the last full stop before this point. The index of the last full stop was stored in a variable, and the chunk appended in the list was replaced with the new chunk that ended at this point. This process was repeated until the end of the input text.

Once the input text was split into chunks, we converted each chunk back into a string and passed it to the model to generate a summary. By looping through each chunk in the list and processing them separately, we were able to generate summaries for longer texts that exceeded the maximum input length of the model. It's worth noting that in cases where the last full stop in a chunk occurred after the 1024th token, we adjusted our strategy for finding the last full stop or split the chunk at a different point to ensure that the input to the model was not truncated in the middle of a sentence.

Overall, our chunking strategy allowed us to generate summaries for longer texts that exceeded the maximum input length of the facebook/bart-large-cnn model, while maintaining coherence and relevance in the generated summaries.

3.5. Similarity analysis using the BERT [2,3,8] model – The last stage of the pipeline involves using the BERT model from the same library to determine the semantic similarity between two given sentences. The code utilizes the BERT Tokenizer to encode text, and we use the base-base-uncased pre-trained model. The Bert Semantic Data Generator class generates batches of data with an option to include or exclude labels based on whether they are used for training/validation or testing. After encoding the sentence pairs with the tokenizer, the model predicts the similarity between the two sentences by generating probabilities for each of the three labels: contradiction, entailment, and neutral. The predict() function produces the predicted probability for each label, which can then be used to determine the semantic similarity between the two sentences. This stage of our pipeline is critical in determining the similarity between two given sentences, and its applications can extend to various natural language processing domains, such as information retrieval and question-answering systems.

4. Numerical Examples and Results

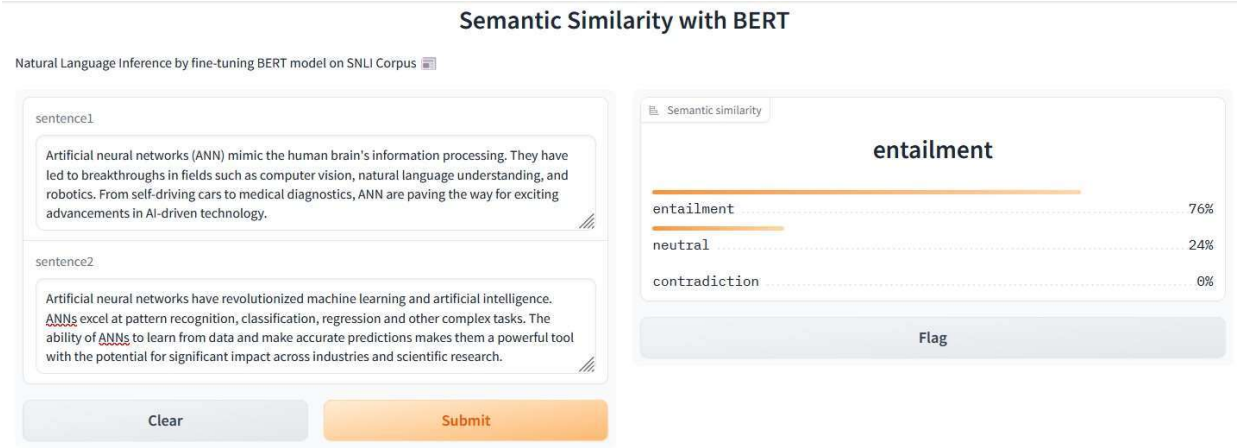
4.1.Audio-to-text rouge[16] score table- The table below shows how well the whisper model performed while converting audio to text. The output generated by the whisper model is compared with the ground truth of the meeting and the wellness of the model is determined by rouge score which is a value between [0,1]. We get an average of 0.8-0.9 using the whisper model, which is considered quiet good.

	Generated_transcript	Ground_truth	rouge1	rouge2	rougeL	rougeLsum
0	Hello. Hello. Hi, Akshay. How are you? I'm good. So today we're going to talk. Yes, yes, continue. Today we're going to talk about artificial neural networks. That is ANN. ANN has become the cornerstone of the world. The ability to mimic the human brain's information processing has lead to breakthroughs in fields such as computer vision, natural language understanding, and robotics. ANN consists of interconnected layers of neurons that connect the brain to the brain. The brain is a complex structure that is connected to the brain. The brain is a complex structure that is connected to the brain. ANN consists of interconnected layers of neurons that perform calculations on input data, gradually extracting meaningful features, and generating output predictions. The strength of ANN lies in their ability to learn and adapt through a process called training, where they adjust their internal parameters based on labeled data set. This enables ANN to generalize from examples and make accurate predictions on unseen data. From self-driving cars to medical diagnostics, ANN are transforming industries and paving the way for exciting advancements in AI-driven technology. That was a wonderful session. Thank you so much. you	Hello, Hello. Hi, Takshi, how are you? Hi, Saurav. I'm good. So today we're going to talk, Yes, yes, continue. Today we're going to talk about artificial neural networks. That is Ann. Ann have become the cornerstone of modern AI applications. Their ability to mimic the human brains information processing has lead to breakthroughs in fields such as computer vision, natural language understanding, and robotics. Ann consists of interconnected layers of neurons that perform calculations on input data, gradually extracting meaningful features and generating output predictions. The strength of Ann lies in their ability to learn and adapt through a process called training, where they adjust their internal parameters based on labeled data set. This enables Ann to generalize from examples and make accurate predictions on unseen data. From self driving cars to medical diagnostics, Ann are transforming industries and paving the way for, Exciting advancements in AI driven technology. That was a wonderful session. Thank you so much.	0.847645	0.818942	0.847645	0.847645
1	Hello, Takshya. Good morning. Hello. Good morning, Sourav. So, as last time we were talking about artificial neural networks, we're going to continue with that discussion. So, artificial neural networks have revolutionized machine learning and artificial intelligence. ANNs excel at pattern recognition, classification, regression and other complex tasks. During the training process, ANNs learn to recognize patterns and make predictions by adjusting the weights and biases associated with connections between neurons. This process, known as backpropagation, involves iteratively feeding the network with input data and comparing its output with desired output. With advances in deep learning, ANNs with numerous layers known as deep neural networks have achieved remarkable success in various domains such as image and speech recognition, natural language processing and autonomous vehicles. The ability of ANNs to learn from data and make accurate predictions makes them a powerful tool with the potential for significant impact across industries and scientific research. Okay. So, thank you so much for this session. Bye.	Hello, Takshay. Good morning. Hello, Good morning, Sourav. So as last time we were talking about artificial neural networks, we're going to continue with that discussion. OK. So artificial neural networks have revolutionized machine learning and artificial intelligence. Ann's excel at pattern recognition, classification, regression and other complex tasks. During the training process, Ann learn to recognize patterns and make predictions by adjusting the weights and biases associated with connections between neuron. This process, known as backpropagation, involves iteratively feeding the network with input data and comparing its output with desired output. With advances in deep learning, Ann with numerous layers known as deep neural networks, have achieved remarkable success in various domains such as image and speech recognition, natural language processing and autonomous vehicle. The ability of Anns to learn from data and make accurate predictions makes them powerful tool with the potential for significant impact across industries and scientific research. Okay so thank you so much for this session. Bye.	0.954128	0.904615	0.954128	0.954128

4.2.Text to summary rouge score table - Again, we use rouge score to determine the summary generated by our fine-tuned BART model, presented in the table below.

	Generated_summary	Ground_truth	rouge1	rouge2	rougeL	rougeLsum
0	Artificial neural networks (ANN) mimic the human brain's information processing. They have led to breakthroughs in fields such as computer vision, natural language understanding, and robotics. From self-driving cars to medical diagnostics, ANN are paving the way for exciting advancements in AI-driven technology.	Artificial Neural Networks (ANN) have revolutionized AI applications by mimicking human brain processes. ANN's ability to learn, adapt, and make accurate predictions through training has led to breakthroughs in computer vision, natural language understanding, and robotics. They are transforming industries such as self-driving cars and medical diagnostics, paving the way for exciting advancements in AI technology.	0.750000	0.470588	0.615385	0.615385
1	Artificial neural networks have revolutionized machine learning and artificial intelligence. ANNs excel at pattern recognition, classification, regression and other complex tasks. The ability of ANNs to learn from data and make accurate predictions makes them a powerful tool with the potential for significant impact across industries and scientific research.	Artificial neural networks (ANNs) have revolutionized machine learning by excelling in pattern recognition and prediction tasks through backpropagation. Deep neural networks (DNNs) with multiple layers have achieved impressive success in various domains. ANNs have the potential for significant impact across industries and scientific research.	0.516129	0.329670	0.473118	0.473118

4.3.Comparison of summaries - Here, we present the results of the comparison of the summaries generated by the fine-tuned BART model, that is done using BERT. The model takes in input two paragraphs or sentences and tells us the extent to which they are related to each other. It rates the inputs using 3 parameters, i.e. entailment, neutral and contradiction. Here's an example of the two meetings we recorded, converted them to text, generated their summaries, and finally compared them. Both the meetings were about discussing the topic 'ANN' and hence we find similarity between the two generated summaries.



5. Conclusion

In this paper, we have developed a meeting summarization system that is able to automatically generate summaries of meetings. The system collects meeting data from various sources, preprocesses the data, and applies a summarization model to create concise summaries of the most important points discussed during the meeting. The summaries are then compared to identify common themes or patterns, and the quality of the summarization model is evaluated.

Through the development of this system, we have shown that meeting summarization can be an effective way to help stakeholders stay informed about the content of meetings, without having to spend the time and resources necessary to attend every meeting in person. By generating summaries automatically, the system is able to improve the efficiency and effectiveness of communication among stakeholders.

However, it is important to note that the effectiveness of the meeting summarization system depends on the quality of the summarization model. As such, it is important to continue to refine and optimize the model to ensure that it is able to accurately capture the most important points discussed during the meeting.

Overall, the meeting summarization system developed in this paper represents an important step towards improving communication and collaboration among stakeholders. By providing concise

summaries of meetings, the system makes it easier for stakeholders to stay informed and make more informed decisions based on the content of the meetings.

References

[1] M. Bewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, Ves Stoyanov, Luke Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880.

[2] J. Devlin, M. Wei, C. Kenton, L. Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of NAACL-HLT 2019, pages 4171–4186 Minneapolis, Minnesota, June 2 - June 7, 2019. c 2019 Association for Computational Linguistics.

[3] K. Fleming.. “BERT: A Review of Applications in Natural Language Processing and Understanding”. 10.48550/arXiv.2103.11943 (2021).

[4] A. Gadford, J. Wook Kim, T. Xu, G. Brockman, Christine McLeavey, Ilya Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision”.

[5] L. Jones, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, A. N. Gomez, Ł. Kaiser, I. Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[6] R. Kils, I. Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).

[7] Y. Lang, Y. Sun, J. Dai, X. Hu, Q. Guo, X. Qiu, X. Huang. "BART-Reader: Predicting Relations Between Entities via Reading Their Document-Level Context Information." In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pp. 159-171. Cham: Springer International Publishing, (2022).

[8] E. Maufiq, L. Zeratul, I. M. Yusoh, B. M. Aboobaider. "Enhancing the Takhrij Al-Hadith based on Contextual Similarity using BERT Embeddings." *International Journal of Advanced Computer Science and Applications* 12.11 (2021).

[9] S. Michel, L. Chi-kiu, "Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data." *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL).* (2019).

[10] D. Mila. “Semantic Similarity- A Review of Approaches and Metrics”. International Journal of Applied Engineering Research. (2019).

[11] A. Mundara , C. Mankar, S. Nagrale, P. Malviya, A. Sangle, M. Navrange, “BART Model for Text Summarization : An Analytical Survey and Review”, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 2, Issue 5, May 2022

[12] W. Niajie., “Literature review on vulnerability detection using NLP technology” (2021)..

[13] G. Raima, R. Stefan, S. Yücel. “Context-based Extraction of Concepts from Unstructured Textual Documents”. Information Sciences. 588. 10.1016/j.ins.2021.12.056 (2021)..

[14] V. Ronit , M. Saraswathi , S. S. Pranav, “Implementation of Video and Audio to Text Converter”, International Journal of Research Publication and Reviews, Vol 4, no 5, pp 1204-1208 May 2023

[15] E. Sambria, B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014, doi: 10.1109/MCI.2014.2307227.

[16] L. C. Yenk, “ROUGE: A Package for Automatic Evaluation of Summaries”, Appl. Math. Comput, (2022).

[17] R. A. Yuhan, "Robust speech recognition via large-scale weak supervision." arXiv preprint arXiv:2212.04356 (2022).