Query-based dialogue summarization using BART

Lingxiao Du

United International College, Dalian Maritime University, Dalian-Liaoning, China, 116000

851448523@qq.com

Abstract. Conversation summarisation is the transformation of long conversational texts into concise and accurate summaries, the importance of which lies in improving the user experience and information filtering. As an important natural language processing task, conversation summarisation can provide concise and accurate information and avoid repetition and redundancy. In the dialogue summarisation task, pre-trained language models can be used to summarise long conversations and generate concise and accurate summaries. The aim of this paper is to investigate the possibility of using bidirectional and auto-regressive transformer models for dialogue summarisation tasks. In our experiments, we analysed the characteristics of the Query-based Multi-domain Meeting Summarization (QMsum) dialogue summarisation dataset, proposed a dialogue summarisation model based on the Bidirectional and Auto-Regressive Transformer model, and designed evaluation experiments to compare its performance with other methods in the dialogue summarisation task. The experimental results show that the results of this thesis are important for facilitating the development of dialogue summarisation tasks and the application of the Bidirectional and Auto-Regressive Transformer model.

Keywords: conversation summary, PTML model, BART model, QMsum dataset.

1. Introduction

Conversation summarisation is the process of extracting important information from long conversational texts and generating concise and accurate summaries. Firstly, the algorithm needs to be able to model the complexity of multi-round conversations. Secondly, it needs to be able to adapt to different scenarios and language styles. Thirdly, the difficulties of supervised data collection need to be addressed. Finally, the dialogue summarisation task requires data modelling and processing of long sequences, which requires a more efficient and array-based processing approach. Pre-trained language models have performed a variety of natural language processing tasks on large datasets, often with fine-tuning or feature extraction to give them greater generalisation power. The powerful application of pre-trained language models has accelerated the development of the field of natural language processing..Bidirectional and autoregressive Transformer is a pre-trained language model based on the Transformer architecture and supports bidirectional and autoregressive generation. It is capable of fine-tuning on both supervised and unsupervised datasets.

This thesis will present research related to pre-trained language models and dialogue summarisation tasks, analyse existing dialogue summarisation datasets and their features, and propose a dialogue summarisation model based on the Bidirectional and Auto-Regressive Transformers (BART) model. In addition, the paper will design experiments to evaluate the model's performance in a dialogue summarisation task.

2. Related work

2.1. Natural language generation and conversation summarization

Nature Language Processing (NLP) is a computer technology that converts structured data into natural language. It can be used in a variety of applications and domains and can enable communication and interaction between natural language and computers.

Conversation summarisation is a natural language processing technique that extracts key information from the large amount of conversation data generated in different scenarios and generates concise summaries. However, there are several challenges with this task, including the diversity of conversations for different scenarios and the difficulty of supervised data collection.

2.2. Pre-trained language models

BART is a pre-trained language model that uses a combination of an encoder-decoder model and an autoregressive mechanism. It uses an autocoder to learn unlabelled linguistic data during training to produce an understanding of natural language. In BART, both the encoder and decoder use the latest Transformer architecture. The encoder converts the input corpus into a semantic representation, while the decoder converts this representation back into the desired output.

2.3. Current status of research on the dialogue summarization task

The aim of dialogue summarisation is to compress the original dialogue into a short version containing important information [1]. The presentation of datasets has facilitated research on deep neural models oriented towards news summarisation. By empirically analysing them using state-of-the-art neural summarisers, the results show that the challenges of dialogue summarisation all require specific representation learning techniques to be better handled [2].

To create this dataset, Zhu et al. collected interview transcripts from National Public Radio (NPR) and Cable News Network (CNN) and used overviews and topic descriptions as summaries [3]. Zhang et al. conducted a comprehensive study of long dialogue summarization, examining three strategies for handling long input questions and locating relevant information: (1) an extended transformer model; (2) a retrieval-then-summarization pipeline model using several dialogue corpus retrieval methods; (3) a hierarchical dialogue coding model. By comparing experimental results on three long dialogue datasets it is shown that the retrieval-then-summarization pipeline model yields the best performance [4].

Existing approaches typically use sequence-based models to process conversations. zhao et al. propose a topic-word-guided attention network for dialogue graphs. The dialogue is better understood by integrating cross-sentence information flow through a masked graph self-attention mechanism. Conversation word features are also introduced to aid the decoding process [5]. Gliwa et al. evaluate the model on a new dataset, the SAMSum corpus [6]. Feng et al. demonstrate that dialogue summarisation systems typically encode text into a number of generic semantic features for more powerful dialogue modelling capabilities [7].

3. Methodology

3.1. Data

Query-based Multi-domain Meeting Summarization (QMsum) is a very large conference abstract dataset, containing conferences spanning multiple domains. The cross-domain QMSum dataset was designed to improve the generalisation performance of the model and to provide a venue for evaluating the performance of the model across different domain conferences [8].

To validate the performance of the model, about 15% of the conferences were randomly selected as the validation and test sets in this paper. The shorter abstract length of QMSum compared to other conference abstract dtasets is due to the fact that our dataset focuses not only on the abstract but also on the specific content of the conference.

Product Meetings: The dataset AMI1 contains minutes from industrial-scale product design meetings.

Academic Meetings: The ICSI2 dataset contains abstracts from the ICSI panel sessions. Unlike AMI, the ICSI meetings focus on student research discussions

Committee Meetings: Meetings of parliamentary committees are also significant. These gatherings emphasize formal talks of a variety of subjects [8].

3.2. Data pre-processing

The pre-processing of the data is partially divided into three stages: topic segmentation, query generation and query-based summarisation.

Opic segmentation: Meeting minutes frequently include discussions on a variety of subjects. The meeting's structure is made clearer by asking annotators to note the main subjects discussed in the meeting and their related text ranges [8].

Query generation: Processing the data requires annotators to construct queries based on patterns, focusing on the assessment of query-based summarisation skills. The query pattern lists contain important aspects that are likely to be of interest, covering the most common issues in meetings with multiple people discussing multiple topics. For multi-granularity meeting content queries, the query pattern list is further divided into a generic query pattern list and a specific query pattern list, with annotators designing meeting content queries for generic and specific queries respectively.Annotators were asked to select two to four main topics and the range of texts associated with each topic, and then design approximately three specific queries based on a predetermined list of patterns for each key topic. Annotators were asked to create queries with a range of relevant content greater than 10 circles or 200 words to verify that the piece was a summary, rather than a question-answer.[8].

Query-based summarization: Query-based summaries require annotators to produce faithful summaries based on the designed queries and minutes. Consistency with the minutes and questions is the most important criterion. Annotated summaries should also be short and fluent. Word limits are set for general queries and answers to specific queries to maintain brevity.

3.3. Theoretical basis for the experiment

BART. BART's pre-trained model can be used to generate and modify text. The structure of this summarisation model consists of an encoder and a decoder (Figure 1). The encoder is a Transformer encoder that converts the input corpus into a vector representation. The decoder is a similar transformer[9].

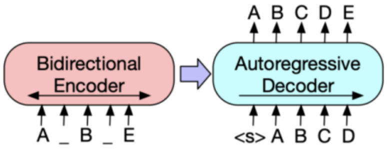

Figure 1. Bart structure diagram.

In the encoder,Each input token is mapped to its embedding representation and passed to a multi-layer Transformer encoding layer. Each encoding layer contains a multi-headed self-attentive mechanism that learns the dependencies and importance between the tokens in the input sequence and generates a contextually relevant vector representation for each token.

The task of the decoder is to generate a summary of the target. It takes the output vector of the encoder as input and locally autoregresses to predict the next marker in the output sequence. The decoder contains a token classification header which is used to predict the probability of the next token in the sequence. [10].

Model architecture. Transform architecture: The Transformer architecture is one of the current state-of-the-art models in the field of NLG and conversation summarisation. The two most important components of this architecture are the self-attentive mechanism and the non-linear transformation layer.

The self-attentive mechanism compares each word in the input sequence with other words to find relevant contexts and is used for tasks such as generating summaries or answering questions. The non-linear transformation layer, on the other hand, is primarily used to learn the mapping relationships between inputs and outputs. This part uses a number of fully connected layers with non-linear activation functions.[11].

Attention mechanism. In BART's dialogue summary model, the attention mechanism is one of the keys. In this model, there are two types of attention mechanisms: self-attention and attention mechanisms (Figure 2).

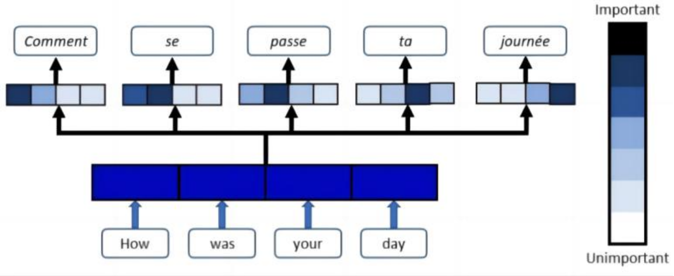

Figure 2. Architecture diagram of attention mechanism.

The self-attention mechanism means that given a fragment of the input sequence, the model uses the same embedding vector to represent all the words and then uses multiple different heads to calculate the attention weights between words. Each head maps the input sequence into a triplet of k, q and v and calculates the mutual weights of all words in the input sequence and the surrounding words. Ultimately, these weights are combined to obtain the output vector (Figure 3).

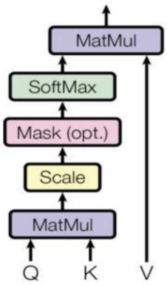

Figure 3. Architecture diagram of attention mechanism.

On the other hand, in the BART model, both self-attention and attention mechanisms are used to create an efficient model of conversation summarisation. This allows the model to process long sequences quickly and to maintain an efficient link between the inputs and outputs of the model.

3.4. Experimental design and flow

This experiment was conducted on a Colaboratory experimental platform with a GPU of 2080 using the BART conversation summary model for experimental comparison, using the QMSsum dataset as the test base. The model was trained using three pre-trained models: test_bart_cnn, test_bart_base and test_bart_large.

The model evaluation metric Range Score was then used to evaluate the similarity metric between the generated natural language text and the target text, which was calculated by:

Split the generated text and the target text into word-based token sequences.

Calculate the intersection between the token sequence of the target text and the token sequence of the generated text, as well as the concatenation between them.

Divide the intersection by the result of the concatenation to obtain the Range Score value.

The range score is the ratio of the number of identical words in the generated text to the number of words in the target text. As the purpose of generating abstracts is to make general statements about the target text, there should be some overlap between the generated abstracts and the original text. The range score measures the degree of overlap between the predicted summary and the reference summary, and thus the performance of the model.

3.5. Experimental results

This experiment found that the three pre-trained models, test_bart_cnn, tast_bart_base and tast_bart_large, all exhibited a 'f' score of Rouge-1 essentially greater than 0.35 on our corresponding conversation dataset summaries. The Rouge metric used in this paper is a common evaluation metric in areas such as automated summarisation. A Rouge1 metric greater than 0.35 indicates that more than 35 per cent of the words are repeated between the generated summary and the reference summary; the more words are repeated between the generated summary and the reference summary, the closer the generated summary is to the reference summary, and the better the model used in this paper represents.

However, overall our experiments using test_bart_cnn on the dataset showed an improvement of 0.4 and 0.6 respectively over the other two pre-trained models on the 'f' scour of rouge-1. This is because testbartcnn, in addition to using tast_bart_base and tast_bart_cnn base and tast_bart_large pre-trained datasets in addition to the tast_bart_cnn dataset. The dataset used is larger than the separate datasets used by the other two pre-trained models, so the results are likely to be more accurate.

Table 1. Rouge-1 Score on Different Models.

rouge-1	test_bart_cnn	test_bart_base	test_bart_large
‘r’	4.04E-01	3.80E-01	3.58E-01
‘p’	4.11E-01	3.85E-01	3.68E-01
‘f’	3.93E-01	3.57E-01	3.35E-01

Table 2. Rouge-2 Score on Different Models.

rouge-2	test_bart_cnn	test_bart_base	test_bart_large
‘r’	1.59E-01	1.50E-01	3.47E-01
‘p’	1.45E-01	1.32E-01	3.53E-01
‘f’	1.45E-01	1.28E-01	3.26E-01

Table 3. Rouge-3 Score on Different Models.

rouge-l	test_bart_cnn	test_bart_base	test_bart_large
‘r’	3.65E-01	1.36E-01	3.23E-01
‘p’	3.71E-01	1.24E-01	3.34E-01
‘f’	3.55E-01	1.17E-01	3.03E-01

4. Discussion and analysis

4.1. Advantages of the BART model

The BART model has several advantages. Firstly, the BART model uses the structure of an autoregressive generative model to generate text based on the Transformer structure, so its inherent structure is very simple. In addition, the attention mechanism in the BART model helps the model to assign attention and relate words to each other. Secondly, the BART model performs particularly well in the task of summary generation, where it can generate text similar to or even better than humanwritten text through a carefully designed training and optimisation method. Finally, BART is a generic pre-trained model that can be used for a wide range of natural language processing tasks.

4.2. Limitations of the BART model

The BART model was trained using an English dataset, so there may be performance issues in other languages or cultural contexts. In addition, BART models are sensitive to the size and quality of the dataset. BART models are time and computational resource intensive to train and test and can take days or even weeks of computational resources to train.

5. Conclusion

To investigate the effectiveness and applicability of dialogue summarisation techniques, this paper uses the BART model for the dialogue summarisation task, analyses the characteristics of the QMsum dialogue summarisation dataset, and proposes a dialogue summarisation model based on the BART model. Evaluation experiments are also designed to compare the performance of the model with other methods in the dialogue summarisation task. It is confirmed that the BART model not only helps the model to perform attention assignment and associate relationships between words, but also performs well on several NLP tasks.

However, there are still limitations in the dataset and language environment, and also the BART model is sensitive to the size and quality of the dataset. In addition, as the field of conversation summarisation is constantly and rapidly changing and evolving, the experiments allow us to keep abreast of the latest developments faster and better, and also provide the basis and ideas for further research and practice of conversation summarisation techniques.

Reference

[1] Feng X, Feng X, Qin B. A survey on dialogue summarization: Recent advances and new frontiers. arXiv preprint arXiv:2107.03175. 2021 Jul 7.

[2] Chen Y, Liu Y, Chen L, Zhang Y. DialogSum: A real-life scenario dialogue summarization dataset. arXiv preprint arXiv:2105.06762. 2021 May 14.

[3] Zhu C, Liu Y, Mei J, Zeng M. MediaSum: A large-scale media interview dataset for dialogue summarization. arXiv preprint arXiv:2103.06410. 2021 Mar 11.

[4] Zhang Y, Ni A, Yu T, Zhang R, Zhu C, Deb B, Celikyilmaz A, Awadallah AH, Radev D. An exploratory study on long dialogue summarization: What works and what's next. arXiv preprint arXiv:2109.04609. 2021 Sep 10.

[5] Zhao, L., Xu, W., & Guo, J. (2020, December). Improving abstractive dialogue summarization with graph structures and topic words. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 437-449).

[6] Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. arXiv preprint arXiv:1911.12237.

[7] Feng, X., Feng, X., Qin, L., Qin, B., & Liu, T. (2021). Language model as an annotator: Exploring DialoGPT for dialogue summarization. arXiv preprint arXiv:2105.12544.

[8] Zhong, M., Da Y., Tao Y., Zaidi A., Mutuma M., Jha R., Awadallah A.H., Celikyilmaz A., Liu Y., Qiu X. Radev D. "QMSum: A new benchmark for query-based multi-domain meeting summarization." arXiv preprint arXiv:2104.05938 (2021).

[9] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. 2019 Oct 29.

[10] Lundberg, C., Viñuela, L. S., & Biales, S. (2022, July). Dialogue Summarization using BART. In Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges (pp. 121-125).

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