Summarizing models in NLP

Ahmed Bahaa

March 13, 2024

Abstract

This report provides an overview of summarizing models in Natural Language Processing (NLP). The motivation behind this research is to understand the current state of the field and identify potential areas for improvement. The literature review section includes a summary of two recent works in the field, highlighting their contributions,and the results of their work based on performance evaluation metrics such as ROUGE. The report aims to provide a comprehensive understanding of the current state of summarizing models in NLP and suggest potential directions for future research.

1 Introduction

Natural Language Processing (NLP) stands at the forefront of artificial intelligence, revolutionizing how machines comprehend and interact with human language. By bridging the gap between human communication and computational analysis, NLP enables computers to understand, interpret, and generate human language in a way that mimics human intelligence. NLP has witnessed unprecedented growth and innovation in recent years, fueled by advancements in deep learning, neural networks, and transformer models (Khurana et al., 2023). These breakthroughs have propelled NLP applications to new heights, empowering machines to perform tasks such as machine translation, sentiment analysis, text summarization, and question answering with remarkable precision. The ability of NLP systems to extract insights from unstructured text data has revolutionized industries ranging from healthcare and finance to marketing and customer service.

Among the myriad applications within NLP, automatic text summarization stands out as a crucial task, aimed at distilling large volumes of text into concise and coherent summaries. This task holds significant potential across various domains, including information retrieval, document summarization, news aggregation, and content recommendation systems.

2 Motivation

The motivation behind this project stems from the growing volume of textual data available across various domains such as news articles, research papers, and social media posts. With the exponential increase in information overload, there is a pressing need for efficient methods to extract essential content swiftly and accurately. By developing a robust summarization model, we can enhance accessibility to vast amounts of text, enabling users to grasp the main points quickly and facilitating decision-making processes.

Furthermore, the advancement of deep learning techniques and transformer models has revolutionized the field of NLP, offering sophisticated tools for text processing tasks. Leveraging these innovations, we aim to explore state-of-the-art methodologies in text summarization and contribute to the development of cutting-edge solutions that can revolutionize how we interact with textual data. Through this project, we aspire to not only deepen our understanding of NLP but also make meaningful strides towards enhancing information extraction and comprehension in the digital age (Kang et al., 2020).

3 Literature Review

In the following section, we will delve into recent advancements in the field of summarization models. We will explore a spectrum of approaches, from traditional extractive methods to cutting-edge abstractive techniques, employed in recent research endeavors. Through an in-depth examination of recent works, we aim to elucidate the state-of-the-art methodologies, innovations, and challenges in automatic text summarization.

Recent work in summarization has made significant progress due to introducing large-scale datasets such as the CNN-DailyMail dataset (Nallapati et al., 2016) and the New York Times dataset (Sandhaus, 2008). However, less work has focused on summarizing online conversations. Early approaches to conversation summarization consisted of feature engineering, template selection methods, and statistical machine learning approaches (Oya et al., 2014). More recent modeling approaches for dialogue summarization have attempted to take advantage of conversation structures found within the data through dialogue act classification, discourse labeling, topic segmentation, and key-point analysis (Ganesh and Dingliwal, 2019). However, such approaches focus exclusively on dialogue summarization, and it is not trivial to extend such methods to longer conversations with many more participants. The authors thus introduce a method to model the structure of the discourse over the many-party conversation.(Barker and Gaizauskas., 2016b) identify three key components of conversational dialogue: issues (that individuals discuss), viewpoints (that they hold about these issues), and assertions (that they make to support their viewpoints). they build on this framework and advances in argument mining for end-to-end training for summarization. Work in argument mining has aimed to identify these argumentative units and classify them into claims, premises, and major claims, or claims describing the key concept in a text.

The authors of the first paper aim to address this research gap by crowd-sourcing a suite of four datasets, which they called ConvoSumm, that can evaluate a model's performance on a broad spectrum of conversation data (Fabbri et al.,2021). For the news comments subdomain, they used the NYT Comments dataset, which consists of 2 million comments made on 9,000 New York Times articles published between 2017 and 2018. For the discussion forums and debate subdomain, they selected Reddit data from CoarseDiscourse which contains annotations about the discourse structure of the thread. For the community question answering subdomain, they used StackExchange (Stack), which provides access to all forums and has been used in modeling for answer relevance and question deduplication. For the email threads subdomain, they used the publicly available W3C corpus. They used BART-large as their base abstractive text summarization model. BART is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. BART is particularly effective when fine-tuned for text generation (e.g. summarization, translation) but also works well for comprehension tasks (e.g. text classification, question answering). They finetuned BART using a polynomial decay learning rate scheduler with Adam optimizer. They used a learning rate of 3e-5 and warmup and total updates of 20 and 200 (Fabbri et al.,2021).

Concerning the results, they trained BART on 200 examples from their validation set for abstractive models, using the remaining 50 as validation and test on the final test set of 250 examples. They evaluated the results using the scores of ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 Measures overlap of unigrams (single words) between the generated summary and reference summary. ROUGE-2 Measures overlap of bigrams (pairs of adjacent words) between the generated summary and reference summary. ROUGE-L Measures longest common subsequence between the generated summary and reference summary. For the first dataset (NYT), the ROUGE-1/2/L scores were 35.91/9.22/31.28 respectively. For the second dataset (Reddit), the scores were 35.50/10.64/32.57 respectively. For the third dataset (Stack), the scores were 39.61/10.98/35.35 respectively. Finally, for the last dataset (Email), the scores were 41.46/13.76/37.70. After that, the authors of the paper fine-tuned the Vanilla Bart model by using BART-arg, which is trained on argument-mining input. For the first dataset, The ROUGE-1/2/L scores after fine tuning were 36.60/9.83/32.61 respectively. For the second dataset, the scores were 36.39/11.38/33.57 respectively. For the third dataset, the scores were 39.73/11.17/35.52 respectively. Finally, for the last dataset, the scores were 40.32/12.97/36.90.

In summary, the results suggest that using argument-mining input improves the performance of the BART model in generating summaries. BART-arg achieves higher ROUGE scores for the first 3 datasets. For email data, however, this did not improve upon the BART baseline, likely due to the dataset's characteristics; email data is shorter and more linear, not benefiting from modeling the argument structure or removing non-argumentative units.

In the second work, the authors apply the attentional encoder-decoder for the task of abstractive summarization. They propose several novel models that address critical problems in summarization that are not adequately modeled by the basic architecture, such as modeling key-words, capturing the hierarchy of sentence-to word structure, and emitting words that are rare or unseen at training time (Nallapati et al.,2016). The model that they used is the Encoder-Decoder RNN. The encoder consists of a bidirectional GRU-RNN (Chung et al., 2014), while the decoder consists of a uni-directional GRU-RNN with the same hidden-state size as that of the encoder, and an attention mechanism over the source-hidden states and a soft-max layer over target vocabulary to generate words. In addition to the basic model, They also adapted to the summarization problem, the large vocabulary 'trick' (LVT) described in (Jean et al., 2014). In their approach, the decoder-vocabulary of each mini-batch is restricted to words in the source documents of that batch. In addition, the most frequent words in the target dictionary are added until the vocabulary reaches a fixed size. The aim of this technique is to reduce the size of the soft-max layer of the decoder which is the main computational bottleneck. Moreover, this technique also speeds up convergence by focusing the modeling effort only on the words that are essential to a given example. This technique is particularly well suited to summarization since a large proportion of the words in the summary come from the source document in any case. As discussed earlier, one of the key challenges is to identify the key concepts and key entities in the document, around which the story revolves. In order to accomplish this goal, the authors need to go beyond the word-embeddings-based representation of the input document and capture additional linguistic features such as parts-of-speech tags, named-entity tags, and TF and IDF statistics of the words. Therefore, they create additional look-up based embedding matrices for the vocabulary of each tag-type, similar to the embeddings for words. For continuous features such as TF and IDF, they convert them into categorical values by discretizing them into a fixed number of bins, and use one-hot representations to indicate the bin number they fall into. This allows us to map them into an embeddings matrix like any other tag-type. Finally, for each word in the source document, they simply look-up its embeddings from all of its associated tags and concatenate them into a single long vector.

The authors used the CNN/Daily Mail Corpus as their dataset. the authors used the human generated abstractive summary bullets from new-stories in CNN and Daily Mail websites as questions (with one of the entities hidden), and stories as the corresponding passages from which the system is expected to answer the fill-in-the-blank question. The authors used three main models in the experiments. The first one is words-lvt2k-1: This is the baseline attentional encoder-decoder model with the large vocabulary trick. This model is trained only on the first sentence from the source document. The second model is words-lvt2k-2-ptr. This model is identical to the model above except for the fact that it is trained on the first two sentences from the source. The last model is words-lvt2k-hieratt (Nallapati et al.,2016).

The performance of the three models on the CNN/Daily Mail test set is as follows: For the first model (words-lvt2k-1) the Rouge-1/ Rouge-2 / Rouge-L scores are 32.49 11.84 29.47 respectively. For the second model (words-lvt2k-ptr), the scores are 32.12 11.72 29.16 respectively. Finally, for the last model (words-lvt2k-hieratt), the scores are 31.78 11.56 28.73 respectively. These scores indicate the quality of the generated summaries compared to reference summaries. The words-lvt2k model achieved the highest scores in all three Rouge metrics.

To conclude, in this work the authors applied the off-the-shelf attentional encoder-decoder RNN that was originally developed for machine translation to summarization and show that it already outperforms state of-the-art systems on the CNN/Daily Mail dataset.

References

[1] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, C¸ aglar Gulc¸ehre, and Bing Xiang. 2016. ˘ Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290, Berlin, Germany. Association for Computational Linguistics

[2] Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12):e26752

[3] Fabbri, A. R., Rahman, F., Rizvi, I., Wang, B., Li, H., Mehdad, Y., & Radev, D. (2021). ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. arXiv preprint arXiv:2106.00829.

[4] Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023.

[5] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications, 82(3), 3713-3744.

[6] Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172.

[7] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In Proceedings of the 8th International Natural Language Generation Conference (INLG), pages 45–53, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

[8] Prakhar Ganesh and Saket Dingliwal. 2019. Abstractive summarization of spoken and written conversation. CoRR, abs/1902.01615.

[9] Emma Barker and Robert Gaizauskas. 2016b. Summarizing multi-party argumentative conversations in reader comment on news. In Proceedings of the Third Workshop on Argument Mining (ArgMining2016), pages 12–20, Berlin, Germany. Association for Computational Linguistics