Research Proposal

The Abstract Group

Marc de Fluiter 5928087 | Idan Grady 7304447 | Maximilian Fehrentz 5016983 | Alexandros Constantinou 2126974 | Michael Konstantinou 2784602 | Yue Shi 6904818

I. Introduction

We live in an age where people have access to a wealth of information that can be found in the form of a document, article, paper, journal, blog, and many other forms of text. Over 7 million new publications are estimated to be published each year in academic publications alone [1]. The goal of automatic text summarization is to shorten text while preserving its information content [2]. As a modern technology, text summarization has become an important and timely tool for users to make sense of large volumes of information in a short amount of time. These days, summarization techniques can be categorized into two types, extractive and abstractive [3]. In extractive summarization, the text's most important sentences are identified by considering statistical features. The abstract summarization technique, on the other hand, creates novel sentences by either rephrasing or using the new words, rather than simply extracting the important sentences [4]. In response to advances in neural methods, the focus of research in automatic text summarization has gradually shifted from extractive to abstractive methods during the last few years [5]. A plethora of techniques and approaches have been tried to solve this task effectively. In this project, our interest is focused on automated text summarization and how this problem can be solved using deep neural networks.

II. Literature Review

The first attempt at this topic was made by Luhn et al [6] in 1958. This research introduced a sentence significance measurement. By analyzing the frequency of words within a sentence, it was possible to estimate the significance of each sentence and generate a summary based on the most significant sentences. According to this research, due to the lack of human orientation, the generated summary has a high degree of reliability, consistency, and stability. On the other hand, the summary provided by this approach is highly influenced by the author's style which can lead to inadequate results. Edmundson et al. [7] elaborated further on this research by introducing the following new techniques :

1) Position: Sentences in specific sections (e.g. after Introduction) contain the topic
2) Cue method: Pragmatic words such as "hardly", "significant", "impossible" should affect the relevance of a sentence
3) Location: First and last sentence of a paragraph should contain topic information

Although the research managed to solve some of the issues observed in the previous approaches, yet the core problem of this technique, which is the reliance on the author's style, is still not solved.

A. Current State

In the past few years, neural abstractive text summarization with sequence-to-sequence(seq2seq) models has gained a lot of popularity. The most common architecture used to build Seq2Seq models is the Encoder-Decoder architecture.

In 2015, Rush et al. published a paper which is the pioneering work of using neural networks to solve the text summarization problem.[8]. This paper first introduced a neural attention seq2seq model with an attention-based encoder and a Neural Network Language Model (NNLM) decoder to the abstractive sentence summarization task, which has achieved a significant performance improvement over conventional methods. In 2016, Chopra et al. extended the model by using a seq2seq structure of CNN-RNN[9].

In the same year, Ramesh Nallapati and his team made many improvements to the above baseline[10] by adding several novel elements and proposed a new data set CNN/Daily Mail. Before this dataset was introduced, many abstractive text summarization models had concentrated on compressing short documents into single-sentence summaries.

In recent years, there have been endless technological improvements based on the idea of seq2seq models. For example, Abigail See et al. proposed a pointer-generator network that implicitly combines the abstraction with the extraction[11], using a pointing/copying mechanism for accurately reproducing actual information[12] etc.

B. Evaluation

There are many open research problems in this area that are still to be solved. One of them is the question of how to measure the quality of a summary. Evaluation techniques are classified as quantitative and qualitative according to whether the evaluation is carried out by comparing results obtained by using various automatic methods or by comparing results obtained by people assessment. This challenge is an active research field these days. For instance, Steinberger and Jezek [13] proposed to use the Latent Semantic Analysis (LSA) to identify the main topics of a document and used it as a "baseline". The ranking is done based on the similarity of the main topics of the summaries and the reference documents. A widely adopted approach of measurement as of today,

however, is the ROUGE method [14]. This evaluation metric measures lexical overlap between the generated and target summaries and compares them to other (ideal) summaries created by humans [15]. Hardy, Narayan, and Vlachos [16] brought to light, however, that current studies differ greatly in their evaluation protocols. Existing studies often refer to only a few baselines and offer human assessments that are inconsistent with prior studies. Additionally, many of the metrics currently used were developed and assessed by using shared-task datasets such as those from the Document Understanding Conference (DUC)[1] and Text Analysis Conference (TAC)[2]. Yet, it has been shown that the mentioned datasets contain human judgments for model outputs scoring on a lower scale than current summarization systems, raising the question of the true value of those metrics [17].

III. Research Plan

A. Overview and Project Goals

This project will consist of three parts. First, we want to build a deep neural network ourselves for generative text summarization. This part of the project serves as an entry point to the topic to get familiar with the current state-of-the-art architecture (seq2seq with Encoder-Decoder) and gain a deeper understanding, hence, the performance of the developed network will not be the main concern of this first stage. The details of the architecture of the network to be developed are yet to be determined, however, it is intended to use a Sequence-to-Sequence Recurrent Neural Network which will also utilize the Attention mechanism[18].

The second part will be the exploration of transfer learning on the pretrained PEGASUS model[3] by Google. Traditional transformer models (such as BERT) use token-level masking and predicting, along with next sentence predictions, to pre-train and learn the context of words. They can then be fine-tuned to different downstream tasks, including abstractive text summarization. The developers of PEGASUS have claimed that their self-supervised pre-training method, called gap-sentence generation, is tailored specifically for the aforementioned task, achieving state-of-the-art performance on the CNN/ DailyMail dataset, after evaluating both using the ROUGE metric, as well as human evaluation. Our aim is to manage to fine-tune both models, using the dataset at hand (Section B), and evaluate them to draw our own conclusion.

Consequently, the third part of the project focuses on the exploration of a promising area of research as proposed in [5], which is concerned with experimenting with different summary evaluation types, as discussed in section C. This will allow for a more concrete comparison of the developed/ fine-tuned models, as well as the completion of the underlying goal of the project: Enhance our understanding and skills on abstractive summarization model development and evaluation.

[1]https://duc.nist.gov/l

[2]https://tac.nist.gov/

[3]https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html

B. Dataset

For this project, the CNN/Daily Mail[4] data will be used. The CNN/Daily Mail dataset was first introduced by Nallapati et al. [19] as part of their research on abstractive text summarization. The dataset consists of 286817 training pairs, 13368 validation pairs, and 11487 test pairs. Since then, it has been cited by 277 papers although each research used the dataset for different experiments or problems (e.g. question answering). The dataset contains human-generated summary bullets that were written from online CNN or DailyMail news stories.

C. Evaluation

As discussed in the second part of the literature review, the ROUGE metric is typically used to compare generated and target summaries. As this method has been standardized, it will also be used to compare the developed/fine-tuned models in this project. As an extension, in the third part of our project, we will attempt to utilize the SummEval toolkit [17], which contains 14 automatic evaluation metrics. Some of the metrics focus on the similarity of the predicted and target outputs, by measuring n-gram overlapping or cosine similarity, whilst others compare the produced summaries with the input documents. This is done either by using the raw texts and similarity measures or by applying summarization techniques themselves to the input documents and comparing the results. Possibly the most sophisticated metric, SummaQA, uses the generated summaries and a BERT-based question answering model to answer cloze-style questions which are produced based on the evaluated summaries.

D. Responsibilities

Overall, all members will share equal responsibilities and the work is going to be divided evenly. Maximilian Fehrentz and Idan Grady will work on the first task of the project, the implementation of a base model in Python, Michael Konstantinou and Alexandros Constantinou will dive into the second task of the project regarding PEGASUS and transfer learning, and Marc de Fluiter and Yue Shi will look into above mentioned open areas of research regarding the evaluation of the model. Each group will be responsible to document the process of their task, however all three groups will work interchangeably in order for the progress of the implementation and documentation of the project to be coherent.

[4]https://paperswithcode.com/dataset/cnn-daily-mail-1

References

[1] M. Fire and C. Guestrin, "Over-optimization of academic publishing metrics: Observing goodhart's law in action," *GigaScience*, vol. 8, Jun. 2019. DOI: 10.1093/gigascience/giz053.

[2] S. Yeasmin, P. Tumpa, A. Nitu, M. P. Uddin, E. Ali, and M. I. Afjal, "Study of abstractive text summarization techniques," Aug. 24, 2017.

[3] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Systems with Applications*, vol. 121, pp. 49–65, May 1, 2019, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.12.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418307735 (visited on 12/13/2021).

[4] A. Pai, "Summarizer using abstractive and extractive method," *International Journal of Engineering Research*, vol. 3, no. 5, p. 5, 2014.

[5] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 9815–9822, Jul. 17, 2019, Number: 01, ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33019815. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5056 (visited on 12/13/2021).

[6] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[7] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[8] A. M. Rush, S. Chopra, and J. Weston, *A neural attention model for abstractive sentence summarization*, 2015. arXiv: 1509.00685 [cs.CL].

[9] S. Chopra, M. Auli, and C. M. Rush, *Abstractive sentence summarization with attentive recurrent neural networks*, 2016. arXiv: 93.98 [cs.CL].

[10] R. Nallapati, B. Zhou, C. N. dos santos, C. Gulcehre, and B. Xiang, *Abstractive text summarization using sequence-to-sequence rnns and beyond*, 2016. arXiv: 1602.06023 [cs.CL].

[11] A. See, P. J. Liu, and C. D. Manning, *Get to the point: Summarization with pointer-generator networks*, 2017. arXiv: 1704.04368 [cs.CL].

[12] K. Song, L. Zhao, and F. Liu, *Structure-infused copy mechanisms for abstractive summarization*, 2018. arXiv: 1806.05658 [cs.CL].

[13] J. Steinberger and K. Jezek, "Evaluation measures for text summarization.," *Computing and Informatics*, vol. 28, pp. 251–275, Jan. 1, 2009.

[14] P. N. Jun. "(PDF) better summarization evaluation with word embeddings for ROUGE — jun ping ng - academia.edu." (2019), [Online]. Available: https://www.academia.edu/15075384/Better_Summarization_Evaluation_with_Word_Embeddings_for_ROUGE (visited on 12/13/2021).

[15] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013 (visited on 12/13/2021).

[16] H. Hardy, S. Narayan, and A. Vlachos, "HighRES: Highlight-based reference-less evaluation of summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3381–3392. DOI: 10.18653/v1/P19-1330. [Online]. Available: https://aclanthology.org/P19-1330 (visited on 12/13/2021).

[17] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, Apr. 26, 2021, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00373. [Online]. Available: https://doi.org/10.1162/tacl_a_00373 (visited on 12/13/2021).

[18] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2017. arXiv: 1706.03762 [cs.CL].

[19] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. DOI: 10.18653/v1/K16-1028. [Online]. Available: https://aclanthology.org/K16-1028.