Project Proposal: Text Summarization Using NLP

Mark Trovinger
Purdue Fort Wayne
tromv01@pfw.edu

Atharva Atre
Purdue Fort Wayne
atreaa01@pfw.edu

Navyaprabha Rajappa
Purdue Fort Wayne
rajan02@pfw.edu

Abstract

This technical proposal presents a project to develop a tool for text summarization using natural language processing (NLP) techniques. The motivation for this project stems from the need to summarize lengthy texts, such as meeting minutes and news articles, for which taking notes can be overwhelming. The proposed solution will explore both extractive and abstractive summarization methods, with a focus on algorithms such as LexRank, LSA, Luhn, KL Sum, and Transformer-based models. The CNN/Daily Mail dataset might be chosen for training and evaluation of the models. ROUGE and BLEU metrics are used for measuring the quality of the generated summaries. The project also includes an analysis section for comparing the performance of extractive and abstractive approaches in different applications, based on factors such as summary quality, speed, accuracy, and interpretability. The results of this project will help identify the most appropriate technique for specific use cases and contribute to the development of efficient and accurate NLP tools for text summarization.

1 Motivation

Meetings can often be long and tedious, making it challenging for individuals to maintain focus and attention for extended periods. Additionally, taking minutes during meetings can overload the designated note-taker, leading to an incomplete or inaccurate record of the discussion. As students, we frequently encounter a vast amount of reading materials that exceed the time we have available to review them. In response to these challenges, we propose to develop a tool that can automatically summarize large volumes of text to help individuals save time and increase productivity.

2 Algorithms

We plan to explore both extractive and abstractive text summarization techniques. For the extractive method, we will examine algorithms such as LexRank, LSA, Luhn, and KL Sum. For abstractive methods, we will focus on Transformer-based architectures.

3 Datasets

While many text-based datasets are suitable for natural language processing, not all are appropriate for text summarization. We will use the CNN/Daily Mail dataset, a collection of 300,000 news articles written in English by journalists at CNN and Daily Mail. This dataset is well-suited for machine learning applications, and we believe it will be useful for training our text summarization models.

While most of the datasets are in English, the majority of our group speaks a language other than English as their first language. Another dataset we are considering will address this; the Hindi Text Short Summarization Corpus is a collection of 330k articles with their headlines collected from Hindi News Websites.

Another dataset we are considering is the BBC News Summary dataset. This dataset for extractive text summarization has four hundred and seventeen political news articles of BBC from 2004 to 2005 in the News Articles folder. For each article, five summaries are provided in the Summaries folder. The first clause of the text of articles is the respective title.

4 Measurements

The primary metric for evaluating text summarization is ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation. There are several variations of ROUGE that we will consider depending on the specific goal of the summarization task. Additionally, we will also measure performance using BLEU, another common metric used in text summarization applications. By leveraging these metrics, we can evaluate the effectiveness of our proposed natural language processing techniques for text summarization.

5 Analysis

As extractive and abstractive summarization techniques are the two main approaches in this domain, it is essential to conduct a comprehensive analysis of how algorithms that implement these approaches compare with each other. The quality of the summaries generated by each approach should be evaluated using metrics such as ROUGE (discussed above) to compare the effectiveness of different summarization techniques. Both extractive and abstractive approaches can be applied in various applications such as news summarization, chatbot conversation summarization, and document summarization. An analysis of the performance of both approaches in different applications can be conducted to determine the most appropriate technique.

To determine the most appropriate technique for a specific use case, an analysis of the strengths and weaknesses of both approaches can be conducted. Factors such as summary quality, speed, accuracy, and ease of use can be considered.

Both extractive and abstractive approaches rely on deep learning models such as transformers and recurrent neural networks to generate summaries. Thus, an analysis of the effectiveness of these models in generating summaries can be conducted to determine the most appropriate model for a specific use case.

Furthermore, an analysis of the interpretability of summaries generated by both approaches can be conducted to determine the approach that generates summaries that are easier to understand and interpret. This analysis can be based on factors such as coherence, readability, and relevance to the source text. By conducting a thorough analysis of the strengths and weaknesses of both approaches, we can determine the most appropriate technique for a specific use case.