SI630 Project Proposal
version 1.0

Zheng Yuan

1 Introduction

The increasing availability of digital data has led to an increase in the amount of text data that needs to be processed and analyzed. This has created a demand for methods that can quickly and effectively summarize large amounts of text data. Automated text summarization is a field of NLP that aims to automatically generate a condensed representation of a document or set of documents while preserving its most important information. The goal of this project is to develop an automated text summarization system that can generate high-quality summaries of large amounts of text data.

2 NLP Task

The specific NLP task for this project is automated text summarization, which involves the automatic generation of a concise representation of a document or set of documents while preserving its most important information. The system will be trained to identify the most important sentences and phrases in a document, and generate a summary that accurately captures the essence of the document. The project will explore several NLP techniques, including extractive summarization and abstractive summarization, to determine the best approach for automated text summarization.

3 Data

The data used for this project will be a large collection of text documents, such as news articles, scientific papers, or product reviews. The text documents will be pre-processed to remove any irrelevant information, such as headers and footers, and split into sentences. The sentences will be used as the input for the text summarization system, and the ground-truth summaries will be manually generated by a team of annotators. I tried to google some datasets of news articles and papers, but the results seemed appear as "famous". I am not quit sure if I can use them, so I decided to use a news API from NEWSDATA .IO, which includes historical news data for the past 2 years.

4 Related Work

1. "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond" by Alexander Konstantinov et al. (2016): This paper (Nallapati et al., 2016) presents an abstractive summarization model based on sequence-to-sequence Recurrent Neural Networks (RNNs), which can generate new phrases and sentences to form a summary. The authors evaluated their model on a Gigaword dataset, which is a large corpus of news articles, as their training data. And they use ROUGE-1, ROUGE-2, and ROUGE-L, which are metrics that compare the summaries generated by the model with the reference summaries.

2. "Get To The Point: Summarization with Pointer-Generator Networks" by Abigail See et al. (2017): This paper (See et al., 2017) proposes a novel attention-based neural network architecture for summarization, which can selectively copy words from the source text, generating a summary that is a combination of summary-specific words and words from the source. The authors used the CNN/Daily Mail dataset as their training data and evaluated their model based on ROUGE scores.

3. "Sentence simplification with deep reinforcement learning" by Wei Zhang et al. (2017): This paper (Zhang and Lapata, 2017) presents a novel reinforcement learning-based approach to sentence summarization, where a deep neural network is trained to make a sequence of decisions to choose a subset of the input sentences to form a summary, based on the rewards given by a reward function that evaluates the quality of the summary. The authors used three datasets: Wikismall(benchmark), Wikilarge(traning set), and Newsela. And they used Flesch-Kincaid Grade Level index and BLEU to measure the readability of the output.

5 Evaluation

As I lack a clear understanding of how to approach this project, I am uncertain of what constitutes a good performance in text summarization. However, after reviewing the papers I have listed, I have learned that there are quantitative methods available for evaluating the effectiveness of the results. Common evaluation metrics for text summarization include ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and F1 score. These metrics compare the generated summary to a reference summary and provide a numerical score that indicates how closely the generated summary matches the reference summary. In my project, I have decided to use the evaluation metrics of ROUGE and BLEU to assess the performance of my text summarization system. ROUGE evaluates the similarity between the generated summary and a reference summary by calculating the overlap, either in terms of word overlap or sentence overlap. On the other hand, BLEU assesses the overlap between the generated summary and the reference summary by determining the n-gram match, where an n-gram represents a sequence of n words.

6 Work Plan

I guess the project will be completed over the course of the rest semester and will be divided into several phases. The phases of the project are as follows:

6.1 Pre-processing Data (within 1 weeks)

Pre-process the text documents to remove any irrelevant information. Split the text documents into sentences. (Clean and tokenize the text)

6.2 Text Summarization System Development (2-4 weeks)

- Decide on the architecture and framework for the model. For example, determine the type of neural network architecture and the framework (such as PyTorch) that I want to use for my text summarization model. Also, I will need to weigh the pros and cons of each type of architecture.

- After choosing the architecture and framework, I will now train the model on the preprocessed data. This involves feeding the data into the network, updating the model parameters based on the error between the predicted outputs and the ground truth labels, and repeating this process for several epochs until the model has learned to perform the text summarization task effectively.

6.3 System Evaluation (1 week)

Evaluate the performance of the text summarization system using ROUGE and BLEUmetrics. Analyze the results and make any necessary modifications to the text summarization system

6.4 Final Report (1 week)

Write a final report on the results of the project. I may present the results and findings of the project to a panel of experts in the NLP community.

This project aims to provide a comprehensive exploration of automated text summarization and develop a high-quality text summarization system that can be used in real-world applications. The results of this project will contribute to the advancement of the field of NLP and provide valuable insights into the challenges and opportunities for automated text summarization.

References

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.