Individual Report - Group Project

Group - 2

Mohiddin Bacha Shaik

News Article Summarization using Sequence-to-Sequence Models

DATS 6312 - Natural Language Processing

Spring - 2024

Dr. Amir Jafari

Table of Contents

1. Introduction:

In the digital age, staying updated with news is crucial but time-consuming. The vast amount of information available online often overwhelms readers, making it challenging to extract key insights efficiently. This project addresses this issue by exploring Sequence-to-Sequence (Seq2Seq) models for automating news article summarization using advanced Natural Language Processing (NLP) techniques. By condensing lengthy articles into concise summaries, our system aims to provide efficient access to essential information, enabling users to stay informed without spending excessive time on reading.

1.1. Objective

Through the exploration of Seq2Seq models, we aim to develop a robust text summarization system capable of accurately capturing the essence of news articles. By leveraging these models, we seek to generate summaries that maintain coherence, relevance, and accuracy, thereby enhancing the accessibility of news content for readers, researchers, and professionals. Additionally, this project contributes to the advancement of NLP technologies, particularly in the domain of text summarization, by pushing the boundaries of what is achievable in creating informative and concise summaries from large bodies of text.

Additionally, we intend to create a user-friendly application that enables individuals to submit either a direct link to an article, multiple articles, or a text prompt outlining their specific interests. This application will then leverage the trained Seq2Seq model to generate concise and informative summaries of the provided content.

2. Dataset

2.1. CNN Daily News Dataset

The CNN Daily News dataset comprises a collection of news articles from CNN and Daily News, covering various topics such as politics, business, sports, and entertainment. It contains a vast array of articles published over a long period, providing rich and diverse content for analysis. Metadata, including the publication date, title, text body, and text summaries accompanies each article.

2.2. Multi-News Dataset

The Multi-News dataset is a comprehensive collection of news articles and summaries from multiple sources, curated specifically for text summarization tasks. It features articles from various news outlets, ensuring a wide range of perspectives and topics. Each article in the Multi-News dataset is paired with a human-written summary, serving as a reference for summarization algorithms.

3. Description of Individual Work

3.1 Data Processing

In the data processing phase of our news article summarization project, I focused on ensuring the quality and consistency of the datasets utilized, namely CNN Daily News and Multi-News. One significant challenge encountered was the presence of duplicate summaries and numerous null values in the articles' datasets, which necessitated meticulous data cleaning to prepare for effective model training.

During the examination of the Multi-News dataset, it became evident that it contained articles from various sources, often resulting in excessively long texts that exceeded the processing capabilities of our summarization model. To address this, a critical decision was made to standardize the maximum text length across all datasets to optimize the model's performance. Consequently, any articles exceeding 5000 tokens were identified and removed from the dataset. This step was vital not only for maintaining consistency in data input but also for enhancing the computational efficiency and accuracy of the summarization model.

This process of data normalization and cleaning was instrumental in preparing our datasets for the subsequent stages of model training, ultimately contributing to a robust framework capable of generating concise and accurate summaries of news articles.

3.2 Encoder-Decoder Using LSTM

I implemented an encoder-decoder architecture using LSTM networks. The encoder consists of three LSTM layers with dropout settings to prevent overfitting, processing embedded representations of input sequences. The decoder uses these representations to generate summaries, employing its own LSTM layer followed by a dense softmax layer for word prediction. I compiled the model using RMSprop optimizer and sparse categorical crossentropy loss, observing significant reductions in loss during training, which indicates effective learning and summarization capabilities. This setup allowed for accurate and efficient summary generation from the input news articles.

For the summary generation component of our project on news article summarization, I encountered challenges in implementing the encoder-decoder model effectively. The intended approach was to use the trained encoder to transform input sequences into a consistent feature vector, which would then initialize the decoder. The decoder was supposed to continue the sequence generation by using the states from the previous timestep as initial conditions, thereby predicting the next word in the summary. However, despite setting up the architecture with embeddings and a dense softmax layer for word prediction, I was unable to achieve functional integration and accurate summary generation. This issue highlighted the complexity of correctly implementing LSTM-based encoder-decoder frameworks, especially in maintaining state continuity between the encoder and decoder for coherent text generation.

3.3 Code Contribution 30%