Text summarization from basic to advanced approaches

Nattapat Juthaprachakul
Simon Fraser University
301350117, njuthapr@sfu.ca

Siyu Wu
Simon Fraser University
301395909, swa246@sfu.ca

1 Motivation

Text summarization is the task of producing a concise and fluent summary of texts or documents while preserving the key information and overall meaning. The goal of this project is to implement different text summarization techniques on the same dataset and compare their performance based on the same evaluation metrics such as ROUGE. There are three main reasons why we choose this project.

First, thanks to the huge amount of available data, Text summarization is becoming an important and useful task in several applications such as business analysis. For example, we always hear someone say, “I don’t want a full report, just give me a brief summary of the results.” With Text summarization, we can gain the key information from articles and documents without reading through every word. This helps save much time and effort.

Second, Text summarization is a relatively new and very interesting topic in this course. We have gone through the list of final projects from the last term and term before last term. We found out that this topic has never been implemented before in this class.

Lastly, we now have an opportunity to implement several techniques we have learnt so far from this course with a real-world dataset such as frequency-driven and sequence-to-sequence techniques. Also, we are going to implement different techniques beyond scope of this course such as a pre-trained attention-based network and Pointer-generator network. Hence, we will compare how much these more advanced methods could improve the performance over the general approaches.

2 Approach

There are two main categories of techniques we are going to implement in this project, which is generic approach and Deep learning-based approach.

2.1 Generic approaches

Frequency-driven approach: Term Frequency Inverse Document (TF-IDF)

Latent Semantic Analysis (LSA): LSA is an unsupervised method for extracting hidden semantic structures of words and sentences based on observed words. It uses the context of the input document and extracts information such as the words that occur together. The meaning of a sentence could be determined by the words it contains. There are three main steps in LSA namely: creating an input matrix, using Singular Value Decomposition to model a relationship among a word and sentence, and selecting an output sentence. The sentence selected by this method could be recognized as the text summary containing the most important information.

Latent Dirichlet allocation (LDA): LDA, known as a Bayesian Topic model, is a generative probabilistic model over the collection of text documents. Each document is modeled as a combination of topics which represents groups of words that tend to occur together. In this project, we want to compare the performance of LDA and LSA.

2.2 Deep learning-based approaches

Vanilla sequence-to-sequence model: The vanilla sequence-to-sequence model such as LSTM is the general Deep learning-based architecture for NLP. We would like to implement this model for our project because it could provide us a baseline of the training time and accuracy for our Deep learning-based approach. Therefore, this helps us better understand how much the improvement of other Deep learning-based approaches could offer regarding both accuracy and efficiency.

Pre-trained network with an Attention-based model as encoder such as BERT: Bidirectional Encoder Representations from Transformers (BERT) is a language representation model that uses Attention mechanism called Transformers without using RNN networks. In addition, since this architecture is pre-trained on a huge amount of data, it could be used as contextual word embeddings. Additionally, it is very powerful and widely adopted in the NLP community. This could result in a big improvement in performance for Text summarization.

Pointer-generator network: Pointer-generator network is an extension of sequence-to-sequence models, which helps fix their shortcomings. In the general sequence-to-sequence architecture, the network tends to reproduce factual details inaccurately and repeat output several times. To solve these problems, pointing operation is used for copying words directly from the source text while the network generator can still produce new words. In addition, the network uses coverage to keep track of what has been summarized so far. Therefore, our project could implement this novel architecture and compare the result with other models described before.

3 Data

The dataset we are going to use in this project is CNN/Daily Mail which is a non-anonymized news dataset widely used in Text summarization. It has more than 300,000 rows of data and two features namely: article and highlights. The first feature is a text of the online news articles which will be used as the documents to be summarized while the second feature is the joined text of highlights which is the target text summarization.

4 Evaluation

The widely used general evaluation metric for Text summarization is Recall-Oriented Understudy for Gisting Evaluation (ROUGE) which automatically determines the quality of a summary by comparing the output texts produced by algorithms and the reference summaries. There are several variations of ROUGE that we will use as follow:

ROUGE-n: it is a recall-based metric that is based on comparison of n-grams between reference summaries and candidate summaries. We will use ROUGE-1 (Unigram) and ROUGE-2 (Bigram).

ROUGE-L: it uses the concept of the longest common subsequence (LCS) between the two sequences of texts. The intuition is that the longer the LCS between two summary sequences, the more similar they are.

5 Timeline and work breakdown

The timeline and work breakdown of our project are divided into several weeks. In the first to second week (24th October -7th November), we will research more about the models we are going to use by doing literature reviews. Also, we will obtain the datasets and set up training facilities (GPU and Cloud) especially for Deep learning-based models. In the third to fourth week (8-20th November), we will finish the generic models and have initial results for Deep learning-based models. In the later week, we can fine tune our models and write reports. Since in case there are some unexpected events, we will have time to solve those problems after the milestone report. Note that this is just an estimated timeline.

6 Work allocation

The allocation of work in this project is based on Text summarization techniques we would like to implement. Since we want to compare a variety of methods for text summarization, each person in our group of two people will train 2-3 models on the same dataset.