CE9010 PROJECT PROPOSAL: DEEP-LEARNING BASED TEXT SUMMARIZATION

Han Simeng

Ruan Dezheng

Tan Yong Jia Glenn

Xu Yanbin

ABSTRACT

With the amount of information booming both online and offline, text summarization has become important for people to digest a large amount of information while reading as little amount of information as possible. This data science project aims to produce high-quality summaries of news articles with techniques taught in the class and beyond.

1 INTRODUCTION

There are two main approaches to text summarization, extractive and abstractive. Extractive summarization chooses individual whole sentences from an article while abstractive summaries first choose sentences and then paraphrases them. In this project, we will explore extractive summarization because it is extremely hard for machines to produce coherent and grammatically correct abstractive summaries. To produce better summaries, we propose to apply sentence compression at the same time.

2 DATASET AND EVALUATION METRIC

2.1 DATASET PREPRATION

We will build a data scraper that scrapes data from the Daily Mail news website (https://www.dailymail.co.uk/home/index.html). We will use the open-source DailyMail dataset with the standard split for training, evaluation and testing(196,961/12,148/10,397). We will then analyze the data statistics like how many words there are in the news on average, and how many words there are in the gold summary on average.

2.2 EVALUATION

We will use the ROUGE score(http://www.aclweb.org/anthology/W04-1013), the standard evaluation metric for text summarization.

3 BUILDING THE MODEL

We will build a sequence-to-sequence model.

- Encoder: pretrained word-embedding, LSTM.
- Decoder: LSTM(also try RNN).
- Loss function optimizer: categorial cross-entropy, Adam optimizer.

4 More Details and How to divide the work

What we will learn to complete this project:

- Text analytics and pre-processing.
- Building neural networks with tensorflow and keras.
- Training, hyper-parameter tuning on a large dataset
- Using existing library to calculate the evaluation score.
- Setting up the cloud GPU.

We aim to divide the work such that each team member is actively participating in each and every stage of the project.

- Tan Yong Jia Glenn: Data Scraping, visualization and analysis; initial model building.
- Han Simeng: Data pre-processing, Initial model building, finalizing the code.
- Ruan Dezheng: Initial model building, cloud environment setup, cross-validation.
- Xu Yanbin: Initial model building, hyper-parameter tuning