Project Proposal for ASTON (Automatic SummarizaTion fOr News)

Ruosi Lin, Xichao Chen, Shijin Tang

Motivation

We live in an era of big data. According to Marr, 16 million text messages, 156 million emails, and 456,000 tweets are sent every minute as of May 2018. With the advancement of technology, the news cycle becomes even shorter: other than newspapers, we can get news from text messages, social media, online subscriptions... It is impossible for human beings to consume every single news in its original form. Plus, it would be laborious to manually summarize every news, regardless of its length. With automatic news summarization, we can skim through the summary and stay current within the areas of interests. Our efficiency improves as we are then able to process more information given the same time window. Torres-Moreno lists several other advantages of automatic text summarization in general, including fewer biases, more personalized recommendations, and so on. Our team would like to explore the latest methodologies for automatic text summarization and apply it on a large-scale dataset.

Approach

Generally, there are two types of approaches for auto text summarization: extraction and abstraction. In this section, overviews for both methods are given. It is noted that current research has extensively focused on extractive methods.

Extraction-based

A subset of existing words, phrases or sentences are selected from the original text and used to form the summary. Generally, the most important content will be chosen from the input. The process can be divided into 3 sub tasks:

1. Create an internal representation of the input. It should express the main aspects.
2. Calculate sentence scores based on the representation.
3. Select a number of sentences to create the summary.

Abstraction-based

This method tries to build a semantic representation for the original text; then uses this information and natural language techniques to generate a summary in a human-readable manner.

The first processing part of this approach is similar to extractive methods. However, this approach can be considered harder compared with extractive methods. It requires natural language generation, which is not quite mature currently.

Dataset & Metrics

Dataset

Google Deepmind made Open sources DeepMind Q&A Dataset open to the public. Each dataset contains a handful of documents (90k and 197k each), and each document includes approximately 4 questions on average. Each question is a sentence with one missing word/phrase, which can be found from the accompanying document/context.

This dataset contains documents and accompanying questions from CNN news articles. There are approximately 90k documents and 380k questions.

We are going to focus on the story data of the CNN News part.

Evaluation metrics

Mean and standard deviation of ROUGE-1 and BLEU scores are going to be used to evaluate the performance of our model.

Rouge-N is the ratio of the count of N-gram phrases which occur in both the model and gold summary, to the count of all N-gram phrases that are present in the gold summary

BLEU is the ratio of the number of words that co-occur in both gold and model translation/summary to the number of words in the model summary. Unlike ROUGE, BLEU directly accounts for variable length phrases – unigrams, bigrams, trigrams etc. by taking a weighted average.

Timeline

Time	Task
Week 1	Literature review, understanding the data
Week 2-5	Implementation
Week 6	Testing & further optimization
Week 7	Summary and warp-up
11/22	Presentation, report, code, data due

Reference

- Text Summarization in Python: Extractive vs. Abstractive techniques revisited
- How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read
- A Gentle Introduction to Text Summarization
- Allahyari, Mehdi, et al. "A brief survey of text mining: Classification, clustering and extraction techniques." arXiv preprint arXiv:1707.02919 (2017).
- Wikipedia contributors. (2018, September 11). Automatic summarization. In Wikipedia, The Free Encyclopedia. Retrieved 19:05, September 29, 2018, from https://en.wikipedia.org/w/index.php?title=Automatic_summarization&oldi
- Lin, Chin-Yew. (2004). ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.