Muhammad Tanveer
Ansh Bhargava
CSCI 49362

NLP Project Proposal

Goal

Our project is centered around the natural language processing task of summarization and exploring the differences between the two main approaches of abstractive and extractive summarization.

Abstractive summarization involves creating a summary of a text which captures the essential information and may include phrases that were not present in the source text to do so. On the other hand, extractive summarization creates a summary only by extracting 'important' phrases or sentences from the source text.

We believe this is an interesting focal point because the model architectures and training for these approaches differ vastly and are both active areas of research. Summarization is also a key task within natural language understanding and we hope to glean a better understanding of these two fundamentally different approaches. Further, text summarization is an important task in its own regard because it can reduce the amount of time required to process information, improve the speed of searching for information, and make learning a topic easier.

Research Question

In modern systems, abstractive text summarization tends to perform better than extractive summarization on popular metrics and is also favored by researchers because of its blue sky potential. However, a modern challenge is extending summarization techniques to longer documents. This exists because current state of the art models are based on Transformers and this architecture has a finite and short context window. Further, these models may be limited on how they take into account document structure and layout, which is particularly important for longer texts.

To this end, we want to explore the efficiency of three approaches of summarizing long documents:

1. Using a pre-trained, publicly available abstractive summarization model to directly summarize the long text.
2. Using a pre-trained, publicly available extractive summarization model to directly summarize the long text.
3. Two step approach: first using the extractive summarizer to shorten the source text and then using the abstractive summarizer to create the final summary.

Our research question asks which approach performs better, as measured by readability and factual consistency metrics.

Our hypothesis is that the two-step process as described above will perform better than either simply running the abstractive or extractive summarizer on the whole document.

Prior and Related Work

1. "Extractive Summarization of Long Documents by Combining Global and Local Context" - Wen Xiao, Giuseppe Carenini
    a. Summary: This paper seeks to train a neural extractive model to generate summaries for long texts (tested on the PubMed and arXiv datasets) by incorporating features that take into account both local and global context. The paper finds that, in general, modeling local context benefits the model results more than global context.
    b. Strengths: This paper extends extractive summarization techniques by leveraging section level information to guide the generated summary but also retaining global context. The model also performs well on popular metrics like ROGUE.
    c. Weaknesses: The model suffers from some redundancy issues. Further, while it incorporates section level discourse elements, it could incorporate structure in a more fine-grained manner, perhaps by way of a discourse tree.

2. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization" - Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu
    a. Summary: PEGASUS is a state-of-the-art LLM trained for various NLU tasks. It leverages semi-supervised pre-training in the form of masked sentences while training.
    b. Strengths: When tasked with abstractive summarization, it performs really well on major metrics. In particular, it is able to be fine tuned to a task with very few examples and hence generalizes well.
    c. Weaknesses: As a Transformer based model, PEGASUS has issues with finite context windows. As such, this model may lose performance over longer documents. Further, this model doesn't explicitly take into account document structure and sectioning which may inhibit its performance.

3. "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents" - Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, Nazli Goharian
    a. Summary: This paper introduces a hierarchical encoder to account for document structure in an abstractive summarization model. They do so by adapting the encoder-decoder framework to be discourse-aware. The paper shows outperformance over state-of-the-art over the arXiv and PubMed datasets.
    b. Strengths: This model addresses some issues of previous abstractive summarization methods by effectively introducing document structure and discourse awareness into summary generation.
    c. Weaknesses: As with other models, the performance is measured by ROUGE which may discount factual consistency in evaluation. Further, this paper doesn't directly address the lack of global context in a Transformer model.

Our proposed project adds to the literature by providing an overview of the latest models across summarization and how they perform on longer texts. We intend on a systematic evaluation across readability and factual consistency metrics and identifying possible avenues of further research for this specific subtask.

Data

We plan on using the PubMed and arXiv datasets. This is reflective of the papers above and are also easily available datasets. Ideally, we would test on datasets from other domains, perhaps finance and legal documents but reliable datasets are difficult to find.

Overview of Approach

We intend on approaching this project as follows:

1. Gathering base datasets and exploring options of other data to test on
2. Accessing appropriate extractive and abstractive models that represent state of the art models.
3. Determining appropriate metrics and evaluation tools.
4. Setting up our testing framework in code, systematically combining our extractive and abstractive models for a hybrid model.
5. Analyzing results and iterating on models / approaches.

Evaluation Plan

We will measure the models on general metrics like ROUGE and more fine-grained metrics like compression rates, density, coverage. We also would like to evaluate the readability and accuracy metric for the abstractive summarization versus extractive summarization.

Timeline

March 23 - Baseline and Data
April 3 - Exploratory Data Analysis and Project Plan Finalized
April 17 - Preliminary Results
April 27 - Finalize Report & Final Results
May 8 - Project Presentation

Team Member Responsibilities

Muhammad Tanveer - tasked with creating an outline for the project, the evaluation metrics we will be using, and the models that will be evaluated, along with data cleaning and preprocessing, creating the final report with results.

Ansh Bhargava - creating the framework for summary generation, performing tests on long corpus, evaluating based on metrics selected.

https://paperswithcode.com/task/abstractive-text-summarization/codeless