Summarizing Abstract for Automatic Title Assignment (SAATA)

W266 - Final Project Proposal

Dickey Woo, Marshal Ma, Waqas Ali

Introduction

The goal of our project is to build a text summarization model, which can generate an appropriate (context-focus) title given a piece of academic paper's abstract. Text summarization is considered one of the more challenging NLP problems due to low text resources and its requirements to extract context from a sequence of text i.e. abstract in this case, and express them in brevity. This is especially challenging in long documents, where context potentially needs to be "ranked/weighted" (not necessarily according to the frequency of appearance) to create a summary text i.e. title in this case, that can express the most important context of the given long document.

We believe that using academic papers provide the following advantages in training a text summarization model:

- Academic papers are often more vigorously reviewed to ensure that their title matches the content of the paper as compared to the other sources of text such as news.
- Many academic papers/journals are open-source and relatively resources-rich
- All academic papers have similar structure and their abstracts are relatively short. It reduces the computational requirements for training neural network model.
- Academic papers' titles are context focus which allows the proposed trained text summarization model for potential down-stream tasks such as context highlighting

Potential challenges to build reliable text summarization models specific to our data source can be due to

- Appearance of unfamiliar (technical) terms in academic papers/journals that our pre-trained models may not have encountered before

- Names/acronyms from the paper titles or abstracts that may need special handling

Dataset:

The dataset we are planning to use as a starter would be the Kaggle arXiv Paper Abstracts, which includes 38,972 entries of academic papers (written in English) sourced from free distribution, open access archives arXiv portal. However, we may also expand our dataset to include more papers from other sources. Our text summarization model would generate a predicted title as summary to the input corpus given the abstract of a paper. The title candidates are then measured against the actual paper title (as "label") evaluated by metrics such as BLEURT scores.

Methodology:

Our text summarization model architecture of choice would be a sequence-to-sequence (transformer-to-transformer) architecture for text generation. Pre-trained transformers models specifically for text summarization tasks (i.e., PEGASUS) would be examined and fine-tuned for our dataset. Beam search algorithms will be used to evaluate multiple candidates at each training step.

References:

1. Y. Liu, M. Lapata. 2019. Data summarization with pretrained encoders. arXiv: 1908.08345v2
2. R. Nallapati, B. Zhou, C. Santos, C. Gulcehri, B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv:1602.206023v5
3. L. Wang, J. Yao, Y. Tao, L. Zhong, W. Liu, Q. Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *Int. Joint Conf. on AI and European Conf. on AI*
4. Rahul, S. Adhikari, Monika. 2020. NLP based machine learning approaches for text summarization. Proc. of the Fourth Int. Conf. on Computing Methodologies and Communication
5. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328-11339). PMLR.