

Project Proposal

Team 5: Shivansh S, Tanishq Chaudhary, Rahul Mehta

Background

A core step towards modelling a deep learning model for extractive summarisation is to have an accurate graph representation of it. Traditional methods have applied RNNs as a linked-list kind of graphs, whereas the newer models prefer having fully-connected graphs like that of Transformers. The problem the paper is trying to solve is that of representing hierarchy and heterogeneity of texts, and representing it as a graph.

Paper Summary

Our initial aim is to implement the paper [Heterogeneous Graph Neural Networks for Extractive Document Summarisation](#). The paper tries to solve the question of having a graph representation of text for the problem of summarisation. It proposes to create a word-sentence heterogeneous graph so that we can pass message between the word and sentence nodes iteratively. The model uses Graph Attention Network (GAT) to do the same. Though the paper tackles the problem by considering sentences as supernodes, and words as basic nodes, it also proposes multi-document summarisation by instead considering documents as the supernodes. The model is able to get SOTA R-1, R-2 and R-L scores for CNN/DailyMail dataset by using the proposed model with trigram blocking.

Project Description

Our main goal is to implement the complex model presented in the paper, for single and multi document summarisation. We also plan to improve upon the model based on better representation and different structures.





Objective/Tasks

1. Re-Implementation of the original paper
2. Experimentation and Improvements on the base paper
3. Extrapolating the base paper towards other forms of summarisation

Dataset and Baselines

We are primarily working on CNN/DailyMail dataset, with established scores given in the paper measured by R-1, R-2 and R-L. We will also be trying out our model based on NYT50 for single document summarisation, and Multi-News for multi document news summarisation.

Work Distribution

 Name	 Task	 Work Distribution	 Duration
Phase One	Re-Implementing	Dataset Pipeline: Rahul CNN → LSTM: Tanishq Training and Exp: Shivansh	24th February - 12th March
Phase Two	Improvements	Embeddings: Rahul TFIDF: Tanishq Linguistic Features: Shivansh	12th March - 25th March
Phase Three	Extrapolate	Abstractive: Rahul Homogeneous Graphs: Tanishq and Shivansh Unsupervised: Shivansh	26th March - 23rd April

References

[1] Yang Liu & Mirella Lapata, Text Summarization with Pretrained Encoders, EMNLP 2019.

[2] Peng Cui & Le Hu & Yuanchao Liu & Hideaki Takeda & Yuji Matsumoto, Enhancing Extractive Text Summarization with Topic-Aware Graph Neural Networks, Coling 2020.

[3] Jiwei Tan & Xiaojun Wan & Jianguo Xiao & Hideaki Takeda, Abstractive Document Summarization with a Graph-Based Attentional Neural Model, ACL 2017.

[4] On Extractive and Abstractive Neural Document Summarization with Transformer Language Models, EMNLP 2020.