Multi-Document Summarization Based on Knowledge Graph

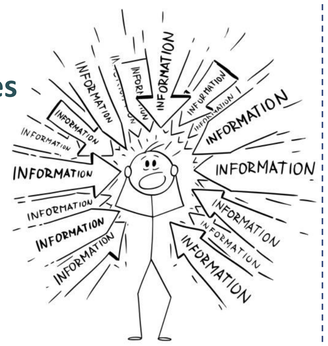
yuanbo tang; zhitao wang; jia guo

TBSI-{tyb22,wzt22,j-guo22}@mails.tsinghua.edu.cn

Introduction

Background & Motivation

- Keep track of the latest advances is becoming more *difficult and time-consuming*.
- Information overload problem can be greatly alleviated by generating succinct and comprehensive *summary*.



Challenges

- Scientific papers contain complex concepts, technical terms, and abbreviations.
- There exist intricate relationships between papers in Multi-Document Summarization task, such as sequential, complementary and contradictory.

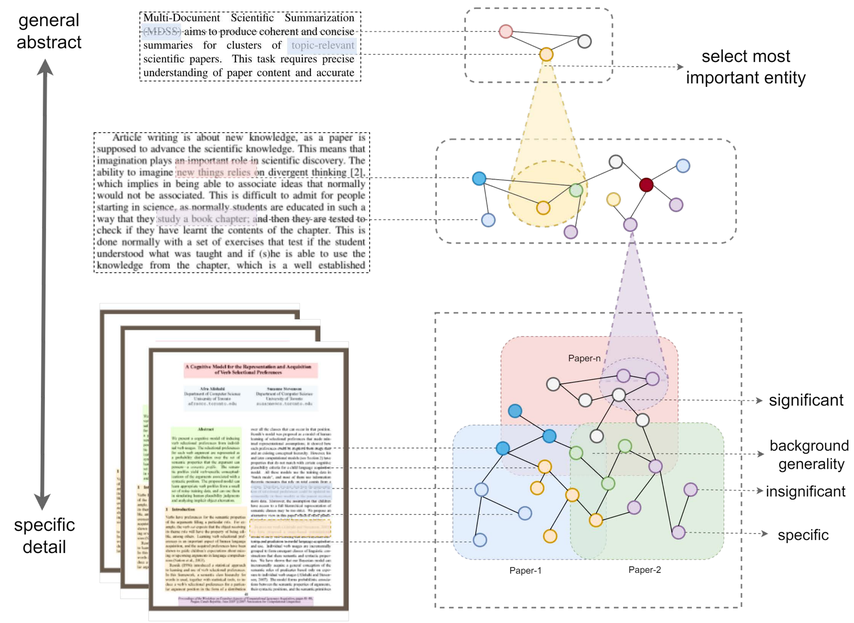
Purpose

- Produce coherent and concise summaries for clusters of topic-relevant scientific papers.
- It is better to have hierarchical and explainable result.

Problem Formulation

Given a set of scientific papers $D = \{d_1, d_2, ..., d_N\}$. Each paper d_i consists of M_i sentences $\{s_{i,1}, s_{i,2}, ...s_{i,M_i}\}$. The gold summary $S = \{w_1, w_2, ...w_{Ns}\}$. The target is to generate a summary $S^* = \{w_1^*, w_2^*, ...w_{Ns}^*\}$ that is close enough to the gold summary S.

Method

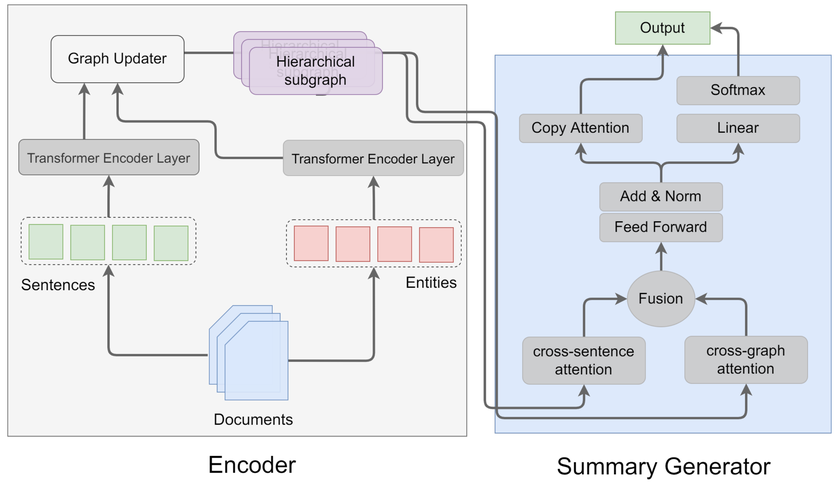


Basic Idea

1) Extract entities and relations from papers and build KG.
2) Score entities and generate hierarchical graphs.
3) Use hierarchical KG to guide the summary generation.

Inspired by ***TF-IDF*** and ***PCA***, we rank the entities iteratively by:

$$score(e_i) = \frac{TF(e_i)}{IDF(e_i)} - \lambda * \sum_{e_j \in C} |e_i - e_j| + \mu * degree(e_i)$$



The overall framework of our proposed model

Result

Dataset[1]

Dataset	# train/val/test	doc. len	summ. len	# refs
Multi-XScience	30,369/5,066/5,093	778.08	116.44	4.42

Evaluation metric

ROUGE-1/ROUGE-2 refers to the overlap of unigram/bigrams between the system and reference summaries.

$$Rouge-N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)}$$

Comparison[2]

Model	R-1	R-2	R-L
Extractive			
LexRank (Erkan and Radev, 2004)	30.19	5.53	26.19
HeterSumGraph (Wang et al., 2020)	31.36	5.82	27.41
Abstractive			
GraphSum (Li et al., 2020)	29.58	5.54	26.52
MGSum (Jin et al., 2020)	33.11	6.75	29.43
Our Model	**33.56**	**8.71**	**20.29**

Conclusion

- A KG-based model is proposed which could produce hierarchical and coherent summary.
- Our method is competitive and promising compared to other abstractive and extractive models.

Limitation & future work

- Due to the limited time, the readability of the result still need to be improved.

References

[1] Y. Lu, Y. Dong, and L. Charlin, "Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles." arXiv, Oct. 27, 2020. doi: 10.48550/arXiv.2010.14235.

[2] Wang Pancheng, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li,Jintao Tang, and Ting Wang. "Multi-Document Scientific Summarization from a Knowledge Graph-Centric View." arXiv, September 9, 2022. https://doi.org/10.48550/arXiv.220