Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev
Department of Computer Science, Yale University

LILY Lab

Introduction

Automatic generation of summaries from multiple news articles is a valuable tool as the number of online publications grows rapidly. Single document summarization (SDS) systems have benefited from advances in neural encoder-decoder model thanks to the availability of large datasets. However, multi-document summarization (MDS) of news articles has been limited to datasets of a couple of hundred examples. In this paper, we introduce Multi-News, the first large-scale MDS news dataset. Additionally, we propose an end-to-end model which incorporates a traditional extractive summarization model with a standard SDS model and achieves competitive results on MDS datasets. We benchmark several methods on Multi-News and hope that this work will promote advances in summarization in the multi-document setting.

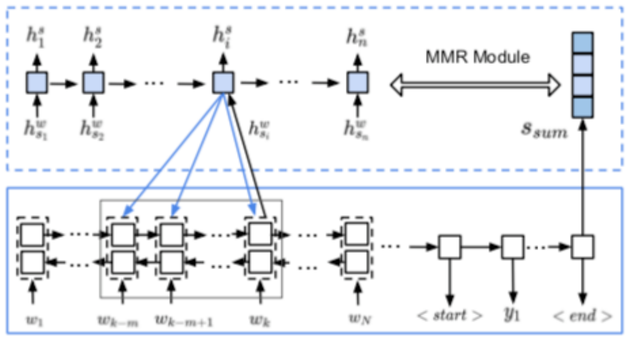
Dataset and Models

Our dataset, which we call Multi-News, consists of news articles and human-written summaries of these articles from the site newser.com. Each summary is professionally written by editors and includes links to the original articles cited. We will release stable Wayback-archived links, and scripts to reproduce the dataset from these links. Our dataset is notably the first large-scale dataset for MDS on news articles. Our dataset also comes from a diverse set of news sources; over 1,500 sites appear as source documents 5 times or greater. See Table 1.

We expand the existing pointer-generator network model into a hierarchical network, which allows us to calculate sentence-level MMR scores (see Figure 1). MMR is a metric which balances the relevance of a given sentence with its redundancy among the current output summary. Given a sentence-level representation from both the articles and the current summary, and then we apply MMR to compute a ranking on the candidate sentences Intuitively, incorporating MMR will help determine salient sentences from the input at the current decoding step based on relevancy and redundancy.

Dataset	# pairs	# words (doc)	# sents (docs)	# words (summary)	# sents (summary)	vocab size
Multi-News	44,972/5,622/5,622	2,103.49	82.73	263.66	9.97	666,515
DUC03+04	320	4,636.24	173.15	109.58	2.88	19,734
TAC 2011	176	4,695.70	188.43	99.70	1.00	24,672
CNNDM	287,227/13,368/11,490	810.57	39.78	56.20	3.68	717,951

Table 1: Comparison of Multi-News to other summarization datasets



Pointer-generator Network

Figure 1: Our Hi-Map model incorporates MMR into an end-to-end pointer generator network

Method	R-1	R-2	R-SU
First-1	26.83	7.25	6.46
First-2	35.99	10.17	12.06
First-3	39.41	11.77	14.51
LexRank (Erkan and Radev, 2004)	38.27	12.70	13.20
TextRank (Mihalcea and Tarau, 2004)	38.44	13.10	13.50
MMR (Carbonell and Goldstein, 1998)	38.77	11.98	12.91
PG-Original (Lebanoff et al., 2018)	41.85	12.91	16.46
PG-MMR (Lebanoff et al., 2018)	40.55	12.36	15.87
PG-BRNN (Gehrmann et al., 2018)	42.80	14.19	16.75
CopyTransformer (Gehrmann et al., 2018)	**43.57**	14.03	17.37
Hi-MAP (Our Model)	43.47	**14.89**	**17.41**

Table 2: Automatic evaluation of our model and several baselines

Method	Informativeness	Fluency	Non-Redundancy
PG-MMR	95	70	45
Hi-MAP	85	75	100
CopyTransformer	99	100	107
Human	150	150	149

Table 3: Results of pairwise human evaluation

Results

Our model outperforms PG-MMR when trained and tested on the Multi-News dataset (Table 2). The Transformer performs best in terms of R-1 while Hi-MAP outperforms it on R-2 and R-SU.

For human evaluation (Table 3), human-written summaries were easily marked as better than other systems, which, while expected, shows that there is much room for improvement in producing readable, informative summaries. We performed pairwise comparison of the models over the three metrics combined, using a one-way ANOVA with Tukey HSD tests and p value of 0.05. Overall, statistically significant differences were found between human summaries score and all other systems, CopyTransformer and the other two models, and our Hi-MAP model compared to PG-MMR. Our Hi-MAP model performs comparably to PG-MMR on informativeness and fluency but much better in terms of non-redundancy, likely due to the incorporation of learned parameters for similarity and redundancy reduces redundancy in our output summaries.

Conclusion

In this paper we introduce Multi-News, the first large-scale multi-document news summarization dataset. We hope that this dataset will promote work in multi-document summarization similar to the progress seen in the single-document case. Additionally, we introduce an end-to-end model which incorporates MMR into a pointer-generator network, which performs competitively compared to previous multi-document summarization models. We also benchmark methods on our dataset. In the future we plan to explore interactions among documents beyond concatenation and experiment with summarizing longer input documents.

Acknowledgement

Thank you for everyone who helped with the project and for Drago for advising the project!