Bottom-up Attention for Multi-Document Summarization

Alexander R. Fabbri, Irene Li, Tianwei She and Dragomir R. Radev

Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

Multi-document summarization (MDS) aims to generate a text abstract from a set of documents. Although the development of neural encoder-decoder model has benefited single-document summarization (SDS) a lot, the results in MDS have not shown satisfactory improvement mainly because of the lack of large dataset. In this project, we propose three models for multi-document neural abstractive summarization. The first model is memory networks, which aim to encode context via multi-hop attention over memories, or previous input/output and have not been previously used in summarization. The second model, bottom-up attention model, which is extended from SDS, incorporates a content selector to the normal summarizer model. The third model applies a hierarchy encoder with a special attention mechanism called MMR-attention (Maximal Marginal Relevance-attention) to generate summarizations. My work is mostly on bottom-up attention model.

Methods

The bottom-up attention model is a two-stage model for single document summarization, which was proposed in [1]. It first uses a word-level tagger as a content selector, and then incorporates the selection probability, which is the output of the tagger, into a Pointer Generator network to generate the summary. The content selection is treated as a binary tagging problem, where for each of the source tokens, 1 if a word is copied in the target sequence and 0 otherwise. The supervised data for this task can be generated by aligning the corresponding summaries to the documents. We first trained the model on relatively large SDS dataset and then fine-tuned it on our MDS dataset. Given a threshold, the selection probability then is applied on the copy attention of Pointer Generator network as a hard mask, which means that only the tokens selected by the tagger can be included in the summary. The Pointer Generator network is also trained on SDS dataset and fine-tuned later.

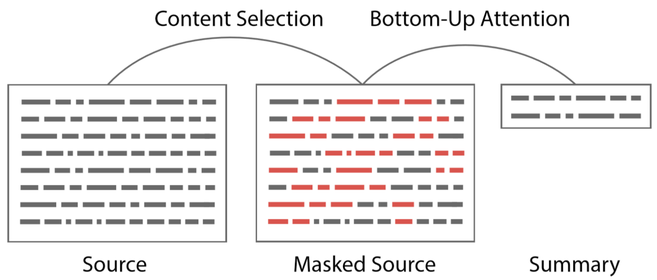

Figure 1. [1] Overview of the selection and generation processes

$$p(\tilde{a}_j^i|x, y_{1:j-1}) = \begin{cases} p(a_j^i|x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

Figure 2. [1] The calculation of bottom-up attention based on the selection probability. Given a threshold ϵ, the selection is applied as a hard mask.

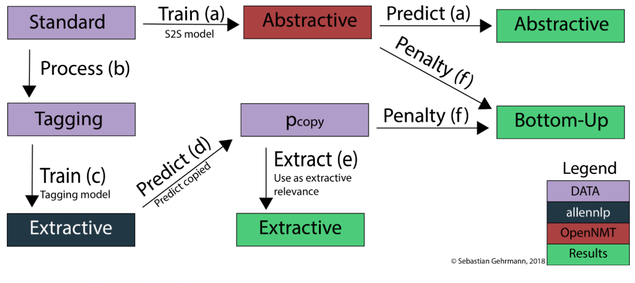

Figure 3. [1] Bottom-up attention implementation structure.

sample%	# pairs	doc # sents	sum # sents	doc # tokens	sum # tokens
0.50	5950	154.14	6.18	3937.01	112.97
0.80	595	246.68	6.18	6297.68	112.97
0.90	1190	277.55	6.18	7094.54	112.97

Table 1. The statistics of merged DUC corpus.

	R-1	R-2	R-L
Extractive(phrases) thres=0.25	16.03	3.04	9.12
Extractive(3-sent)	17.12	4.81	10.25
Abstractive	9.58	2.41	6.67

Table 2. ROUGE score of summaries generated by pre-trained tagger and summarizer.

	threshold / #sents	R-1	R-2	R-L
Extractive (Phrases)	0.4	20.07	1.68	16.56
	0.3	25.17	2.82	20.28
	0.2	**28.9**	4.00	**22.66**
	0.1	21.58	**4.14**	17.17
Extractive (Sentences)	3	28.68	5.43	22.83
	5	31.89	6.00	25.1
	6	**32.42**	**6.19**	**25.42**
	8	31.76	6.09	24.76

Table 3. ROUGE of extractive summaries, which are generated based on the output of fine-tuned tagger.

threshold	R-1	R-2	R-L
0.00	28.89	5.16	24.53
0.15	19.56	3.27	17.24
0.25	15.92	2.36	14.07

Table 4. ROUGE of abstractive summaries, which are generated by pre-trained Pointer Generator.

References

[1] Gehrmann, S., Deng, Y., & Rush, A. (2018). Bottom-Up Abstractive Summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4098-4109).

Datasets

Because of the lack of large datasets on MDS, we generally trained the models on SDS data and then fine-tuned them on SDS data. The SDS dataset we used is the CNN-DM corpus, which contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average). It has 287,226 training pairs, 13,368 validation pairs. As for the MDS dataset, we used a relatively larger training and validation set by merging and sampling the MDS task data of DUC01, DUC02 and DUC03, and tested our model on DUC04. The merging dataset contains 1190 pairs in total. Table 1 shows some statistics of our MDS dataset.

Results and Conclusion

Besides the abstractive summaries, we can also generate extractive summaries based on the selection probabilities. At present, I've got results for extractive summaries generated from fine-tuned tagger and abstractive summaries generated from pre-trained Pointer Generator, which incorporates the selection probabilities output by pre-trained tagger.

By comparing the ROUGE scores (Table 2) of the summaries directly generated from the pre-trained tagger and summarizer (which are trained on CNN-DM corpus) with the performance of fine-tuned models, we can see that fine-tuning can greatly improve the results. We tried 2 ways to generate summaries from the selection probabilities, including extracting phrases and selecting sentences. The promising performance of extractive summaries (Table 3) shows the effectiveness of our fine-tuned tagger. The abstractive summaries are not so good currently, we'll work on improving it in the future.

Acknowledgement

I would like to acknowledge professor Dragomir Radev for his supervise of my project. Thanks Alex and Irene for their guidance and encouragement through the progress.