A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents

(NAACL 2018)

Paper Link: https://www.aclweb.org/anthology/N18-2097.pdf

Introduction:

This huge availability of documents has demanded exhaustive research in the area of automatic text summarization. According to Radef et al. a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that". Automatic text summarization is the task of producing a concise and fluent summary of the document while preserving overall meaning and the key information content.

In recent years, various approaches have been developed for automatic text summarization and they are applied widely in various domains. Automatic text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task.

Automatic text summarization gained attraction as early as the 1950s. An important research of these days was for summarizing scientific documents. In general, there are two different approaches for automatic summarization: extraction and abstraction. Extractive summarization methods work by identifying important sections of the text and generating them verbatim; thus, they depend only on extraction of sentences from the original text. In contrast, abstractive summarization methods aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text.

Objective and Contribution

This is the first model to perform abstractive summarisation on long-form documents (research papers). The architecture consists of a hierarchical encoder that captures the discourse structure of a research paper and an attentive discourse-aware decoder to generate the summary.

The contributions of this paper are:

1. Proposed an abstractive model for summarising research or scientific papers.

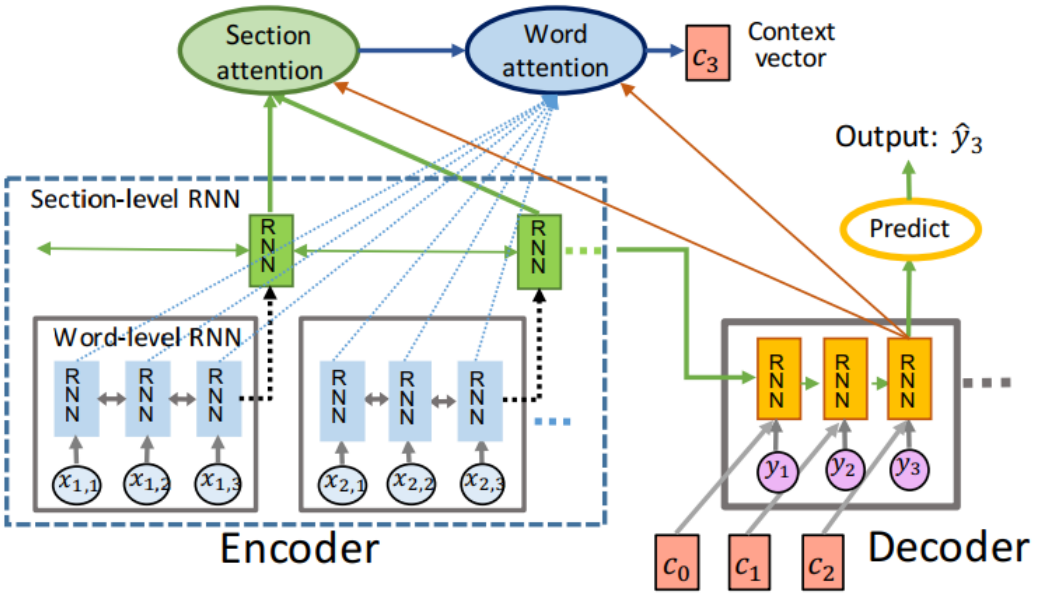
2. Introduced two large-scale datasets of long structured research papers obtained from datasets arXiv and PubMed.

Usually Scientific papers are an example of documents which are significantly longer than news articles. Also these papers follow a standard discourse structure describing the problem, methodology, experiments/results, and finally conclusions. Neural methods have a general framework of sequence-to-sequence (seq2seq) models. In these models a text document is given as input to the encoder and the output of it is fed to the decoder which in turn will generate the summary of the given text document.

But such models mainly focus on summarizing news articles which are relatively short. But we know all the documents cannot always be short and concise. There can be many other types of documents. They may be long documents and they might be structured. In such types of documents the sequence-to-sequence model may not be suitable. Because in these models at the decoder, it needs to capture important and relevant information by looking at all the tokens from the input sequence. And it is difficult to look at all the tokens of a long document and hence it might not work well.

This is where the main contribution of the paper lies. This model includes a hierarchical encoder. This encoder will be able to detect the discourse structure of the document. Also the decoder is discourse-aware decoder which then can generate the summary depending upon the discourse of the document. The decoder can examine different sections of the document and take out important and relevant information for that section. So a meaningful context vector can be obtained.

Discourse-aware Summarisation Model



The figure represents the entire discourse-aware attention model. We can see there are two word-level recurrent neural networks in blue color. The combined output is given to section level RNN for both of them. The decoder also has an RNN network and for getting the summary there is a predict network. At each time step ‘t’ the decoder will form a context vector ‘Ct’ for given text. This context vector captures the relevant and important information from input text. So it is actually a weighted sum of encoder states. In order to compute section attention weights, section-level RNN’s weights are fed to section attention block and to calculate word attention weights, word-level RNN weights are fed to word attention block.

- **HIERARCHICAL ENCODER:**

Our encoder is a hierarchical RNN that captures the discourse structure of the document. The encoder first encodes each discourse section by parsing all the words into their respective section RNN as shown in the figure above. It then takes the outputs of all section RNNs and feed the hidden states into another RNN to encode the whole document.

- **DISCOURSE-AWARE DECODER:**

At each decoding step, the decoder takes in the words of the document and the relevant discourse section. Then the discourse-related information can be used to modify word-level attention function. At each decoding step, the decoder would use the decoder state and context vector to predict the next word in the summary. So this decoder will ensure that the important points from the document from different sections are taken into account.

- **COPY MECHANISM**

Authors have added an additional binary variable to the decoder to decide whether the decoder should generate a word from vocabulary or copy a word from the source. Because if the word is important then the decoder should simply copy that word and if the document has some sentences which correspond to a specific word then using that contextual information decoder should be able to generate the word from vocabulary. The copy probability is learned and optimised during training.

- **COVERAGE MECHANISM**

In case of longer documents where there are longer sequences the neural generation model may repeat words or phrases. In order to avoid this problem attention coverage is checked. This is done by a coverage vector. The coverage will tell the information about the attended document discourse sections and it is incorporated into the attention function.

Dataset:

ArXiv and PubMed

This paper has introduced two datasets collected from scientific repositories. The choice of scientific papers for the dataset is motivated by the fact that scientific papers are examples of long documents that follow a standard discourse structure and they already come with ground truth summaries, making it possible to train supervised neural models.

The datasets are arXiv and PubMed. During the data collection process, the paper has removed any documents that are too long or too short or do not have an abstract or discourse structure. They have removed any figures and tables and normalise any math formulas and citation markers with special tokens. The abstract is the ground-truth. The dataset statistics are displayed below with average document length being 3000 – 5000 words.

Datasets	# docs	avg. doc. length (words)	avg. summary length (words)
CNN	92K	656	43
Daily Mail	219K	693	52
NY Times	655K	530	38
PubMed (this work)	133K	3016	203
arXiv (this work)	215K	4938	220

Table 1: Statistics of our arXiv and PubMed datasets compared with existing large-scale summarization corpora, CNN and Daily Mail (Nallapati et al., 2016) and NY Times (Paulus et al., 2017).

Experiments and Results:

ROUGE is the most widely used metric for automatic evaluation. The Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric is used to automatically determine the quality of a summary by comparing it to human (reference) summaries. There are various versions of ROUGE.

- **ROUGE-n:** This metric is recall-based measure and based

  on comparison of n-grams. Let p be "the number of common n-grams between candidate and reference summary", and q be "the number of n-grams extracted from the reference summary only". Then the score is computed as:

$$\text{ROUGE-}n = \frac{p}{q}$$

- **ROUGE-L:** This measure employs the concept of longest common subsequence (LCS) between the two sequences of text. So, Longer the longest common subsequence between the two summary sentences, the more similar they are. Although this metric is more flexible than the ROUGE-n metric, it has a drawback that all n-grams must be consecutive.

So This paper has used ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L score as evaluation metric and compared the proposed method with different benchmark models as below:

1. LexRank, SumBasic, LSA (extractive)
2. Attention seq2seq, PG network (abstractive)

RESULTS

The two tables below show the results on arXiv and PubMed dataset. The results show that our discourse-aware model was able to outperform all the baseline models, for both extractive and abstractive text summarisation.

	Summarizer	RG-1	RG-2	RG-3	RG-L
Extractive	SumBasic	29.47	6.95	2.36	26.30
	LexRank	33.85	10.73	**4.54**	28.99
	LSA	29.91	7.42	3.12	25.67
Abstractive	Attn-Seq2Seq	29.30	6.00	1.77	25.56
	Pntr-Gen-Seq2Seq	32.06	9.04	2.15	25.16
	This work	†‡**35.80**	†**11.05**	†3.62	†‡**31.80**

Table 2: Results on the arXiv dataset, RG: ROUGE. For our method † (‡) shows statistically significant improvement with $p<0.05$ over other abstractive methods (all other methods).

	Summarizer	RG-1	RG-2	RG-3	RG-L
Extractive	SumBasic	37.15	11.36	5.42	33.43
	LexRank	**39.19**	13.89	7.27	34.59
	LSA	33.89	9.93	5.04	29.70
Abstractive	Attn-Seq2Seq	31.55	8.52	7.05	27.38
	Pntr-Gen-Seq2Seq	35.86	10.22	7.60	29.69
	This work	†38.93	†‡**15.37**	†‡**9.97**	†‡**35.21**

Table 3: Results on PubMed dataset, RG:ROUGE. For our method, † (‡) shows statistically significant improvement with $p<0.05$ over abstractive methods (all other methods).

The paper also performed some qualitative evaluation and observed that the given model was able to generate summaries that not only capture the problem introduction like other SOTA benchmark models but also be able to capture the methodology and impacts of the paper.

Abstract: in this paper , the author proposes a series of multilevel double hashing schemes called cascade hash tables . they use several levels of hash tables . in each table , we use the common double hashing scheme . higher level hash tables work as fail - safes of lower level hash tables . by this strategy , it could effectively reduce collisions in hash insertion . thus it gains a constant worst case lookup time with a relatively high load factor (@xmath0) in random experiments . different parameters of cascade hash tables are tested .

Pntr-Gen-Seq2Seq: hash table is a common data structure used in large set of data storage and retrieval . it has an o(1) lookup time on average , but the worst case lookup time can be as bad as . is the size of the hash table . we present a set of hash table schemes called cascade hash tables . hash table data structures which consist of several of hash tables with different size .

Our method: cascade hash tables are a common data structure used in large set of data storage and retrieval . such a time variation is essentially caused by possibly many collisions during keys hashing . in this paper , we present a set of hash schemes called cascade hash tables which consist of several levels (@xmath2) of hash tables with different size . after constant probes , if an item ca 'nt find a free slot in limited probes in any hash table , it will try to find a cell in the second level , or subsequent lower levels . with this simple strategy , these hash tables will have descendant load factors , therefore lower collision probabilities .

Figure 2: Example of a generated summary

Conclusion and Future Work:

As per authors, this was their first attempt to address neural abstractive summarization of single, long documents. This paper successfully summarizes long and structured documents. This paper has used ROUGE as an evaluation framework but this doesn't capture nuances in coverage of the summaries, also it is non-trivial to evaluate these qualities for long document summarization. So, as per them future work can design expert human evaluations to handle these nuances.