On Extractive and Abstractive Neural Document Summarization with Transformer Language Models

Paper link-https://arxiv.org/pdf/1909.03186.pdf

Objective and Contribution

This review is based on the extensive study of documents on summarisation method techniques. Literature has proposed a method to achieve summarisation through a process consisting of two exclusive steps i.e extractive and abstractive summarisation. The extractice step has an upper and or we can say an important advantage toward the summarization result at the end. The reviewed document has suggested that the summaries obtained trough the abstractive step are better than what we were obtaining through the method of copy mechanism. Abstraction step has achieved higher ROUGE scores when compared to previous works. The outcome of this work has been listed down :

1. Better performing transformer language model than Seq2Seq models when used in summarising long scientific articles. Effectiveness was demonstrated.
2. Present model has achieved a higher ROUGE score as well as being able to produce larger abstractive summary when compared to previous work.

Datasets

Following are the four different datasets used for long document summarisation :

1. PubMed
2. Newsroom
3. bigPatent

4. arXiv

Dataset	Documents	Comp Ratio	Sum Len	Doc Len
arXiv	215913	39.8	292.8	6913.8
PubMed	133215	16.2	214.4	3224.4
Newsroom	1212726	43.0	30.4	750.9
BigPatent	1341362	36.4	116.5	3572.8

Table 1 : Statistics from (Sharma, Li and Wang 2019) for the datasets used in this work. The number of document/summary pairs, the ratio of number of words in the document to the abstract and the number of words in the summary and document.

Structure

In the given work the methodology is structured in following components:

1. ***Extractive summarisation*.** This model is structured on different layers and in it the sentences from the documents are either being copied or classified to obtain an extractive summary.
2. ***Abstractive summarisation***. The transformer model is conditioned based on the obtained extractive summary and documents.

EXTRACTIVE SUMMARISATION

In the extractive step two different hierarchical models are used to extract sentences from the document. These models are : hierarchical seq2seq sentence pointer and sentence classifier. The purpose of extraction is to get the important sentences and get rid of noisy sentences. It helps us to better train our transformer language model. Encoder decoder architecture of hierarchical seq2seq model involve :

1. The encoder is a bidirectional LSTM at both the word and sentence level (hierarchical)
2. The decoder is an autoregressive LSTM

Words and sentence level directional LSTM is combined by the hierarchical encoder. Every sentence in the documents is encoded by the tokan - level biLSTM for obtaining sentence embeddings. Further document representation is being done by the sentence level biLSTM encodes. Hidden statement of the past sentences that were extrates previously is taken by the autoregressive LSTM as input and used to predict next sentence that were to be extracted next. Both the pointer network and sentence classifier encodes the document using hierarchical. A sigmoid function is used to feed the final document representation after concatenating each sentence embedding and thus the probability of sentence inclusion in extractive summary is calculated.

ABSTRACTIVE SUMMARISATION

The single transformer language is being trained from scratch by using “formatted” data. GPT-2 language model is used for transformers that were trained by factorising joint distribution of words autoregressive, based on the above outcome the training data was organised in a certain format in which the summary we put is ground truth, and after that summaries were being generated using the same model. The joint distribution and summary was modeled in the same way and during training we used conditional distribution to generate summary at interface. Following are the four different sections on which training data is formatted. :

1. *Paper Introduction*. (Enough data is available to generate abstract)
2. *Extracted summary (from extractive summarisation)*
3. *Abstract (ground-truth summary)*
4. *Rest of the paper*. Serve to train language model to understand domain language

There are datasets for which the entire section is the introduction only and the rest of the paper is not available or not written. Following figure shows the overall structure.

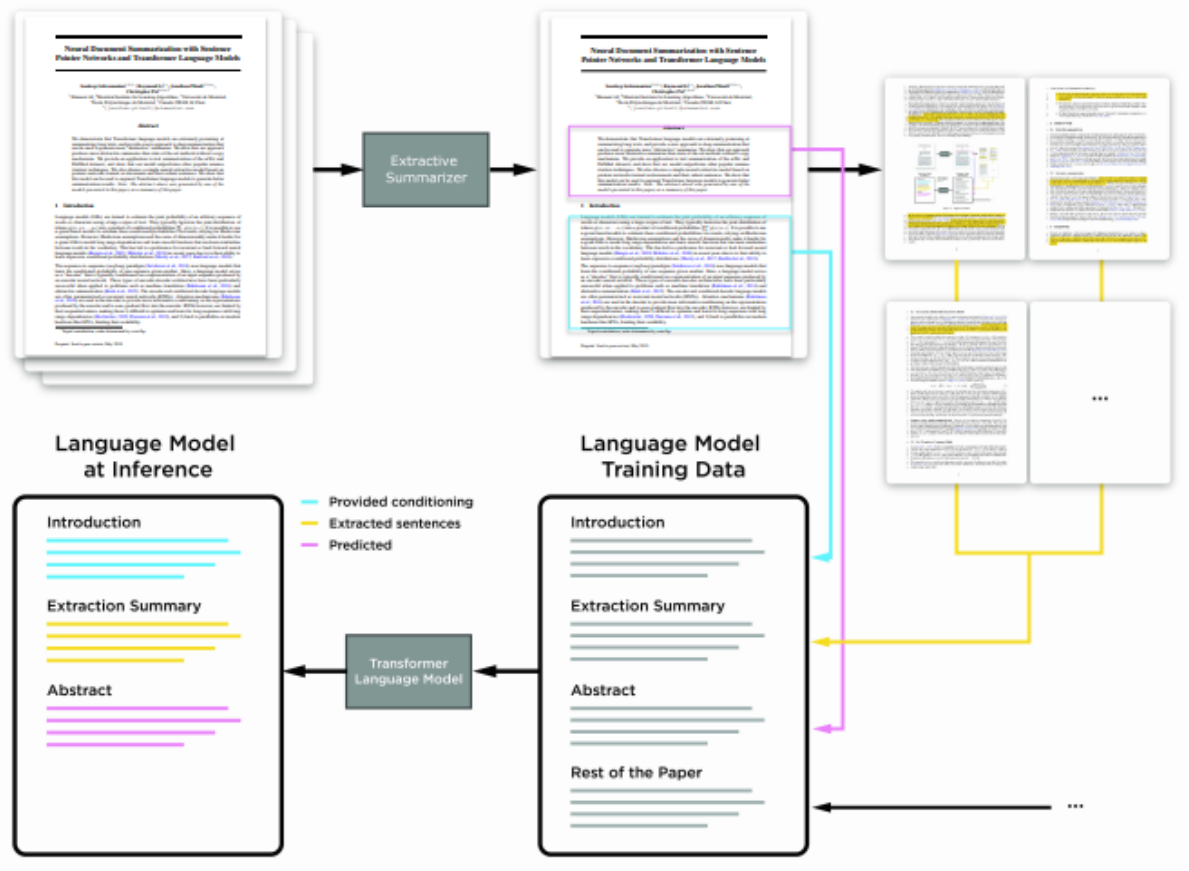

Figure 1: Proposed model for abstractive summarization of a scientific article. An older version of this paper is shown as the reference document. First, a sentence pointer network extracts important sentences from the paper. Next, these sentences are provided along with the whole scientific article to be arranged in the following order: Introduction, extracted Sentences, abstract & the rest of the paper. A transformer language model is trained on articles organized in this format. During inference, the introduction and the extracted sentences are given to the language model as context to generate a summary. In domains like news and patent documents, the introduction is replaced by the entire document.

Results and Analysis

The proposed model in the given work has outperformed all the previous extractive models and table 2 and table 4 shows the supporting data for our claim. Following tables are for arXiv and PubMed dataset. TLM has also outperformed the other abstractive models in case of Newsroom date set, and the margin is pretty high as compared to Seq2Seq.It has also turned out to be better than pointer - generator network. With the Exconsumm model we have obtained some mixed results.

Table 2: Summarization results on the arXiv dataset. Previous work results from (Cohan et al. 2018). The following lines are a simple baseline Lead-10 extractor and the pointer and classifier models. Our transformer LMs (TLM) are conditioned either on the Introduction (I) or along with extracted sentences (E) either from ground-truth (G) or model (M) extracts.

Model	Type	ROUGE			
		1	2	3	L
Previous Work					
SumBasic	Ext	29.47	6.95	2.36	26.3
LexRank	Ext	33.85	10.73	4.54	28.99
LSA	Ext	29.91	7.42	3.12	25.67
Seq2Seq	Abs	29.3	6.00	1.77	25.56
Pointer-gen	Mix	32.06	9.04	2.15	25.16
Discourse	Mix	35.80	11.05	3.62	31.80
Our Models					
Lead-10	Ext	35.52	10.33	3.74	31.44
Sent-CLF	Ext	34.01	8.71	2.99	30.41
Sent-PTR	Ext	42.32	15.63	7.49	38.06
TLM-I	Abs	39.65	12.15	4.40	35.76
TLM-I+E (M,M)	Mix	41.15	13.98	5.63	37.40
TLM-I+E (G,M)	Mix	**41.62**	**14.69**	**6.16**	**38.03**
Oracle					
Gold Ext	Oracle	44.25	18.17	9.14	35.33
TLM-I+E (G,G)	Oracle	46.40	18.15	8.71	42.27

Table 4: Summarization results on the PubMed dataset. Previous work results from (Cohan et al. 2018). The following lines are a simple baseline Lead-10 extractor and the pointer and classifier models. Our transformer LMs (TLM) are conditioned either on the Introduction (I) or along with extracted sentences (E) either from ground-truth (G) or model (M) extracts.

Model	Type	ROUGE			
		1	2	3	L
Previous Work					
SumBasic	Ext	37.15	11.36	5.42	33.43
LexRank	Ext	39.19	13.89	7.27	34.59
LSA	Ext	33.89	9.93	5.04	29.70
Seq2seq	Abs	31.55	8.52	7.05	27.38
Pointer-gen	Mix	35.86	10.22	7.60	29.69
Discourse	Mix	38.93	15.37	**9.97**	35.21
Our Models					
Lead-10	Ext	37.45	14.19	8.26	34.07
Sent-CLF	Ext	45.01	19.91	12.13	41.16
Sent-PTR	Ext	43.30	17.92	10.67	39.47
TLM-I	Abs	37.06	11.69	5.31	34.27
TLM-I+E (G,M)	Mix	**42.13**	**16.27**	8.82	**39.21**
Oracle					
Gold Ext	Oracle	47.76	20.36	11.52	39.19
TLM-I+E (G,G)	Oracle	46.32	20.15	11.75	43.23

Table 5: Summarization results on the bigPatent dataset. Previous work results from (Sharma, Li, and Wang 2019). Our transformer LMs (TLM) are conditioned on the whole document or additionally with extracted sentences (E) either from ground-truth (G) or model (M) extracts.

Model	Type	ROUGE		
		1	2	L
Previous Work				
Lead-3	Ext	31.27	8.75	26.18
TextRank	Ext	35.99	<u>11.14</u>	29.60
SumBasic	Ext	27.44	7.08	23.66
LexRank	Ext	35.57	10.47	29.03
RNN-Ext	Ext	34.63	10.62	29.43
Seq2Seq	Abs	28.74	7.87	24.66
Pointer-gen	Mix	30.59	10.01	25.65
Pointer-gen (Cov)	Mix	33.14	11.63	28.55
Sent-rewriting	Mix	37.12	11.87	32.45
Oracle				
Gold Ext	Oracle	43.56	16.91	36.52
OracleFrag	Oracle	91.85	78.66	91.85
Our Models				
Sent-CLF	Ext	<u>36.20</u>	<u>10.99</u>	<u>31.83</u>
Sent-PTR	Ext	34.21	10.78	30.07
TLM	Abs	36.41	11.38	30.88
TLM+E (G,M)	Mix	**38.65**	**12.31**	**34.09**
TLM+E (G,G)	Oracle	39.99	13.79	35.33

Table 6: Summarization results on the Newsroom dataset. Previous work results from (Grusky, Naaman, and Artzi 2018) and (Mendes et al. 2019).

Model	Type	Extractive			Mixed			Abstractive		
		ROUGE								
		1	2	L	1	2	L	1	2	L
Previous Work										
Seq2Seq	Abs	6.1	0.2	5.4	5.7	0.2	5.1	6.2	1.1	5.7
TextRank	Ext	32.4	19.7	28.7	22.3	7.9	17.7	13.5	1.9	10.5
Pointer-gen	Mix	39.1	27.9	36.2	25.5	11.0	21.1	14.7	2.3	11.4
Lead-3	Ext	53.0	49.0	52.4	25.1	12.9	22.1	13.7	2.4	11.2
Exconsumm	Mix	**68.4**	**62.9**	**67.3**	31.7	16.1	27.0	17.1	3.1	14.1
Our Models										
Sent-CLF	Ext	53.0	47.0	52.1	26.8	12.6	23.6	15.4	2.7	12.8
Sent-PTR	Ext	60.7	55.2	59.7	28.9	14.1	25.1	15.9	2.8	13.0
TLM	Abs	49.8	39.7	47.4	27.1	11.6	22.8	**20.4**	**6.9**	**17.1**
TLM+E (G,M)	Mix	63.3	57.3	61.8	**31.9**	**16.6**	**27.4**	20.1	6.5	16.6
Oracle										
Gold Ext	Oracle	68.1	64.5	67.3	40.8	24.6	34.2	21.9	5.2	16.3
TLM+E (G,G)	Oracle	78.8	74.0	77.8	38.6	22.0	33.6	24.5	9.6	20.8

The TLM (TLM - I +E (G,M)) has given a much better ROUGE score and is superior to previous abtractive results just with one exception on ROUGE-L. We infer from this that this could be because of the absence of a copy mechanism in our case, that makes it difficult to get exact matches on large n- grams. Below figure is the conclusive support for this hypothesis as 25-grams can be copied by the discourse-aware model using a copy mechanism. Added to that it also shows that more abstractive models were generated by this TLM when we compared it with other TLM models that were being used in previous work. And also the percentage of n-grams overlap was low between source documents and generated summary.

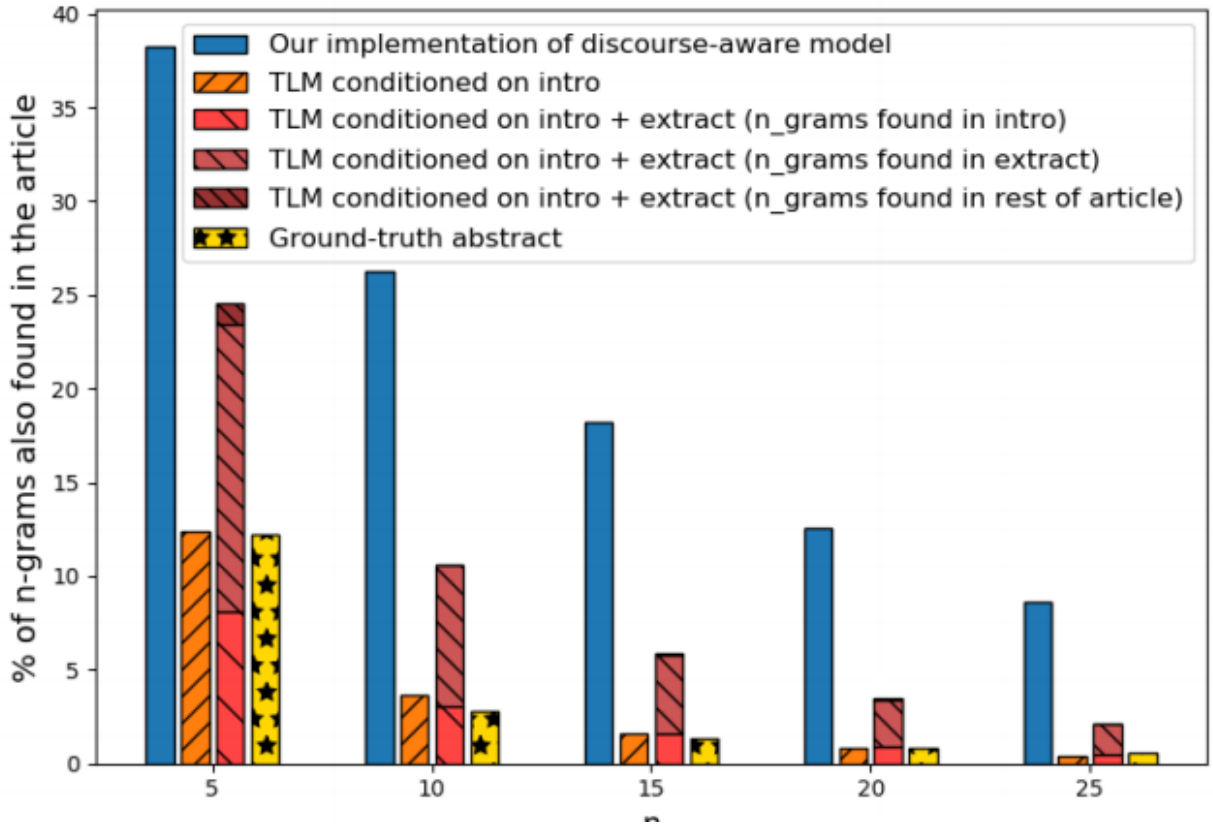

Figure 2: n-gram overlaps between the abstracts generated by different models and the input article on the arXiv dataset. We show in detail which part of the input was copied for our TLM conditioned on intro + extract.

We also measure the upper bound performance of our TLM (TLM-I+E (G,G)) by including the ground-truth extracted sentences in both training and testing. Lastly, the figure below showcases the qualitative results of summaries generated by our TLM.

Table 3: Qualitative Results - News articles and our model generated summaries on the NewsRoom dataset

Document — A new plan from the government of the Philippines would offer free wireless internet to people across the country while also likely eating into the annual revenue of the nations telecoms. Bloomberg reports that the Philippines government plans to roll-out its free Wi-Fi services to roughly half of the countrys municipalities over the next few months and the country has its sights set on nationwide coverage by the end of 2016. The free wireless internet service will be made available in public areas such as schools, hospitals, airports and parks, and is expected to cost the government roughly $32 million per year. [...]
Abstractive — : The government is reportedly considering a nationwide service plan to give free Wi-Fi access to rural areas.
Mixed — The government of the Philippines is considering a new plan to provide free wireless internet to the nation's largest cities and towns.
Extractive — The new plan will include free wireless internet to residents across the country while also probably eating into the annual revenue of the country's telecoms.
Document — (CBS) - Controversy over a new Microsoft patent has people questioning whether or not the intention has racist undertones. CNET reported that Microsoft has been granted a U.S. patent that will steer pedestrians away from areas that are high in crime. [...]
Absractive Summary — The new Microsoft patent claims a device could provide pedestrian navigation directions from a smartphone.
Mixed Summary Microsoft won a U.S. patent for a new way to steer pedestrians out of areas that are high in crime

Conclusion and Future Work

The generated model is giving very strong coherence and fluency. The generation of imaginary / inaccurate content still persists. In future more content with factual correctness can be given focus and coherency still has bottlenecks that can be improved with further model evaluation.