D Y Patil International University

GenCompareSum: Text Summarization

Dheeraj Parmar | 20190802113

ABSTRACT - Pre-trained Language Models (PLMs) have been used to improve the performance of TS. However, PLMs are limited by their need of labeled training data and by their attention mechanism, which often makes them unsuitable for use on long documents. To this end, we propose a hybrid, unsupervised, abstractive-extractive approach, in which we walk through a document, generating salient textual fragments representing its key points. We then select the most important sentences of the document by choosing the most similar sentences to the generated texts, calculated using BERTScore. We evaluate the efficacy of generating and using salient textual fragments to guide extractive summarization on documents from the biomedical and general scientific domains. We compare the performance between long and short documents using different generative text models, which are fine tuned to generate relevant queries or document titles. We show that our hybrid approach outperforms existing unsupervised methods, as well as state-of-the-art supervised methods, despite not needing a vast amount of labeled training data.

INTRODUCTION - Whilst transformers have made great advancements in their ability at capturing semantic knowledge, they have also introduced new limitations. Firstly, they are restricted by the number of tokens that they can process at any one time. Another issue is the computational cost of fine tuning the attention mechanisms embedded in transformers. These constraints are challenging for recent text summarization methods, often resulting in analysis being done on a truncated version of a document. Since summarization should be able to succinctly capture the meaning of very long documents in a few sentences, the requirement to truncate a document before summarization is a major disadvantage.

These are mostly supervised methods, requiring large amounts of labeled training data, which are often unavailable or time-consuming and costly to produce. We address the challenges of supervised methods by adopting a hybrid unsupervised approach, where the PLMs are required only to act on short sections of the document at any time, meaning that our method can be extended to any document length. Furthermore, by nature of it being an unsupervised approach, it does not require manually labeled training data for the extractive summarization task. To-date, unsupervised methods for text summarization have generally used graph-based methods, the more recent of these using transformer-based embeddings to calculate weights between the nodes in the graph.

We differ from these previous approaches as we do not use a graph-based model and instead evaluate the effectiveness of a novel approach – generating and using salient textual fragments to guide the extractive summarization.

Text summarization methods are divided into extractive and abstractive groupings. Extractive methods select the most relevant sentences from a document and abstractive methods consider the most relevant pieces of information to produce new textual fragments which convey the core message. Although abstractive summarization has the potential to be more succinct and readable, in its current state it cannot be trusted to be factually consistent, making it unsuitable in many practical applications, such as summarization of biomedical articles for use by clinicians.

Specifically, we use transformer-based models for the generation of salient points, but ultimately, we generate an extractive summarization to ensure factual consistency.

MOTIVATION - Our method, GenCompareSum, is a two-step hybrid summarization approach. GenCompareSum first splits a document into sections of several sentences and walks through them, generating salient textual fragments which represent each section. We experiment with different generative models, which are fine tuned to predict either queries or document titles, that best represent a section of the document. Our method then uses these generated textual fragments to guide an unsupervised extractive summarization by calculating the BERTScore similarity between each of the generated texts and each of the sentences in the source document. We then select the sentences with the highest scores to form the extractive summarization.

We evaluate our approach on short and long versions of data sets from the biomedical and scientific domains. Furthermore, we compare the use of different PLMs for generating salient textual fragments.

CONTRIBUTIONS -

1. A novel two-step unsupervised hybrid abstractive-extractive summarization method, which generates salient textual fragments - queries and document titles - which represent sections of a document, and then uses them to guide the extractive summarization step.
2. The fusion of state-of-the-art PLMs with unsupervised approaches, to achieve a summary which harnesses the semantic knowledge of transformer-based models, whilst being extendable to any length document, without requiring a large corpus of training data.
3. Evaluation results demonstrate our hybrid method outperforms both existing unsupervised methods and state-of-the-art supervised methods, both on long and short documents.

STATE OF THE ART -

Text Splitting T5 is a sequence-to-sequence model, pre-trained on a cleaned and pre-processed version of the Common Crawl2 dataset. The T5 model uses an encoder-decoder architecture and is pre-trained via an unsupervised task in which 15% of tokens are masked.
We use docTTTTTquery model trained on the MSMARCO data set. Then fine tune our own model on long-answer - query pairs from the biomedical domain. We refer to this model as ‘t5-med-query’.
N-Gram Blocking - N-gram blocking is a technique which is applied to reduce redundancy and improve coverage in summarization models.
Text Vector Comparison - BERTScore uses BERTbased token embeddings, calculates the cosine similarity between them and uses greedy matching to match each token in the first text to its most similar token in the second; these scores are averaged across the sentences to give precision, recall and F1 scores which quantify the similarity between two texts.

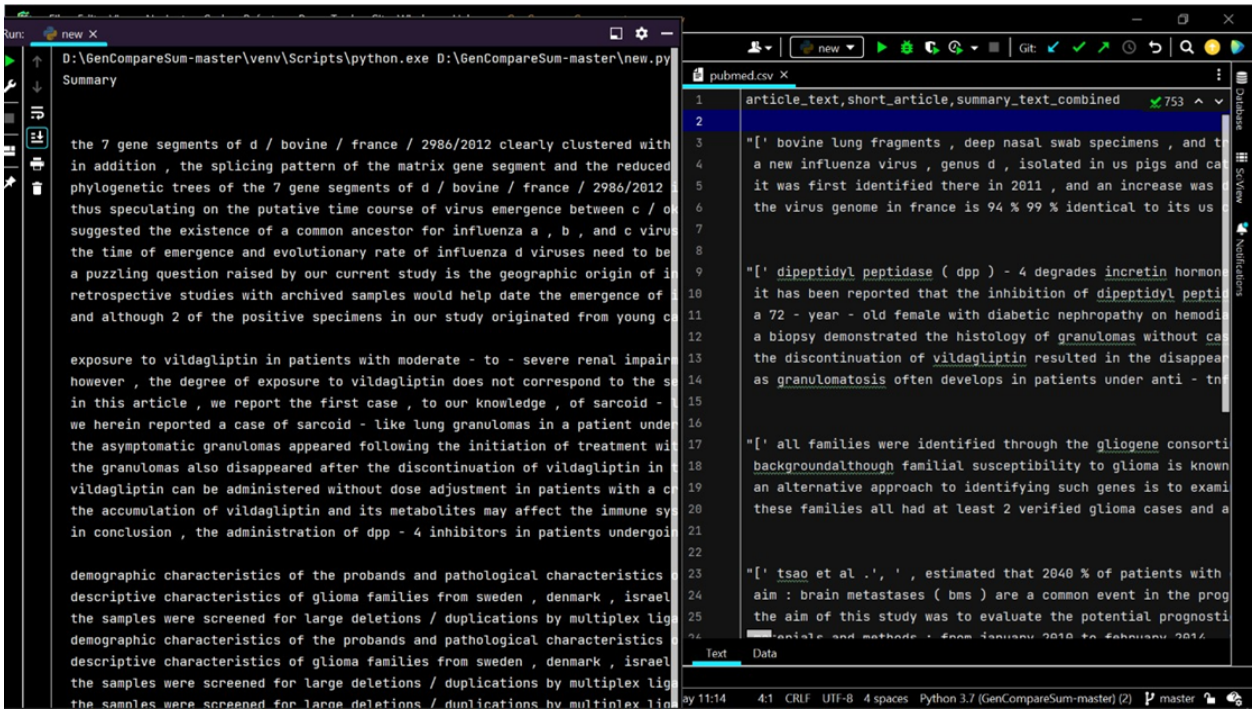
EXPERIMENTAL RESULTS -

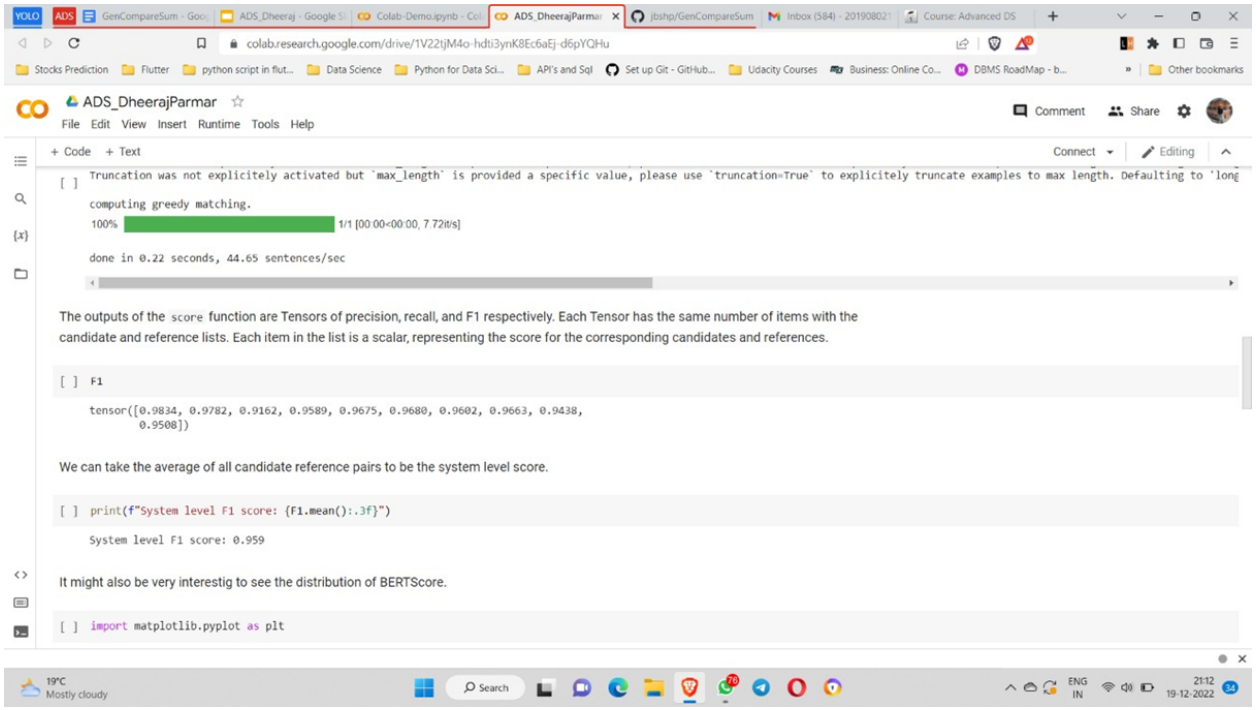
Model	PubMed			S2ORC			CORD-19			ArXiv		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Short Document												
ORACLE	47.27	22.85	43.20	49.29	25.42	45.52	43.47	17.75	39.28	47.29	18.49	41.90
RANDOM	34.98	10.82	31.37	34.69	11.06	31.30	31.64	7.91	28.10	34.53	8.88	30.26
LEAD	35.39	12.07	32.28	**40.50**	**16.72**	**37.68**	34.80	10.17	31.67	34.35	8.75	30.61
LexRank	38.48	13.05	34.92	39.44	14.57	36.13	35.65	10.17	32.11	38.98	11.44	34.64
TextRank	38.15	12.99	34.77	40.17	14.84	36.63	36.25	10.61	32.53	37.97	11.58	33.53
SumBasic	36.11	11.06	32.67	35.99	11.99	32.87	33.63	8.82	30.22	37.14	9.83	33.06
BERTExtSum	38.78	14.47	35.43	39.41	16.14	36.38	34.68	10.34	31.42	39.36	11.74	35.09
GenCompareSum (docTTTTTquery)	37.82	13.12	32.41	38.31	14.27	35.17	33.77	9.73	30.66	38.59	11.49	34.50
GenCompareSum (t5-med-query)	38.54	13.67	35.06	38.96	14.78	35.80	36.77	11.24	33.29	38.92	11.59	34.76
GenCompareSum (t5-s2orc-title)	**39.19**	14.35	**35.65**	40.16	15.84	36.91	**36.84**	**11.35**	**33.35**	**39.66**	**12.30**	**35.38**
Long Document												
ORACLE	61.76	36.78	57.61	64.11	39.21	60.16	59.10	32.09	54.63	60.16	32.17	54.97
RANDOM	37.26	11.19	33.66	37.12	10.23	33.73	33.37	7.70	29.98	34.20	8.70	30.64
LEAD	37.23	11.11	33.67	40.50	16.72	37.68	34.61	10.17	31.68	34.70	10.27	31.37
LexRank	41.02	15.83	37.18	42.60	15.84	38.97	39.50	12.65	35.68	33.94	12.09	30.62
TextRank	34.53	12.98	30.99	36.58	13.23	33.10	32.99	10.39	24.47	26.57	9.20	23.74
SumBasic	40.61	12.42	36.54	36.63	10.43	33.68	33.88	8.24	30.86	33.18	7.75	30.29
BERTExtSum*	41.87	16.01	**38.51**	**43.56**	**17.85**	**40.40**	38.95	12.17	35.48	**40.65**	14.01	**36.89**
GenCompareSum (docTTTTTquery)	40.54	14.77	36.83	40.78	14.24	37.43	36.84	11.19	33.51	38.19	12.76	34.55
GenCompareSum (t5-med-query)	41.60	15.67	37.79	41.84	15.10	38.35	39.33	12.31	35.74	37.17	11.97	33.95
GenCompareSum (t5-s2orc-title)	**42.10**	**16.51**	38.25	43.39	16.84	39.82	**41.02**	**13.79**	**37.25**	39.96	**15.15**	36.19

For the short documents, our method GenCompareSum (t5-s2orc-title) performs best across three out of four of the data sets, and second-best for the fourth data set. There is no clear 'second-best' model out of the methods compared for the short data sets.

For the long document data sets, GenCompareSum (t5-s2orc-title) outperforms all other unsupervised models. A strong unsupervised baseline, LexRank has been shown in prior literature to give competitive performance when compared to supervised approaches. In-line with these works, we show LexRank to be the best-performing unsupervised method after our own.

SUMMARY TEXTUAL





The precision, Recall and the F1 Score calculated using the bert score on the pyrouge evaluation metric.

CONCLUSION - In this work we propose GenCompareSum, a novel two-step unsupervised hybrid abstractive-extractive method for text summarization. We evaluate the efficacy of using PLMs to generate salient textual fragments which represent the key points of a document – experimenting with generation of both queries and document titles -- and using them to guide the second step, extractive summarization. We show that our unsupervised method, which can be extended to any length of document and does not require a corpus of annotated training data, outperforms over both strong supervised and unsupervised baselines on long and short documents. Furthermore, we show that our best-performing model uses title-document pairs for the generative task, which are readily available across many domains without the need for manual labelling effort.

REFERENCES -

1. **https://commoncrawl.org**
2. **https://www.bing.com**
3. **https://huggingface.co/doc2query/S2ORC-t5- base-v1**
4. **https://github.com/bheinzerling/pyrouge**
5. **https://github.com/miso-belica/sumy**
6. **Etc.**

THANK YOU !!