Outperforming Long-form Neural Abstractive Summarizers with Simple Unsupervised Baselines

Shira Eisenberg and Sam Wiseman

University of Chicago and TTIC

Motivation and Overview

Neural generation methods have made great strides when applied to the abstractive summarization of news documents, but their effectiveness at summarizing longer form scientific articles is less well understood.
We show that:

- When applied to the scientific article summarization tasks recently proposed by Cohan et al. (2018), neural methods only minimally outperform simple lede-style baselines
- Neural methods are outperformed by LexRank (Erkan and Radev, 2004), an unsupervised, non-neural, extractive baseline
- LexRank with original TF-IDF-based similarity scores outperforms similarity scores based on BERT

Task Background

- The summarization of longer form scientific articles has received less attention than the summarization of news articles.
- Cohan et al. (2018) have recently introduced two datasets for the summarization of longer form scientific articles: one of PubMed articles paired with their abstracts, and one of arXiv articles paired with their abstracts. (Statistics shown in Table 1)
- Cohan et al. (2018) propose an attentional sequence-to-sequence style model with copy attention and coverage penalty, augmented with a hierarchical document encoder, and a discourse-aware decoder.
- We compare the Cohan et al. (2018) model, taken to embody many aspects of state-of-the-art neural text generation systems, with the LexRank model Erkan and Radev (2004), which is neither nueral nor supervised

LexRank Background

- LexRank (Erkan and Radev, 2004) is an unsupervised extractive approach to text summarization, which attempts to score sentences in terms of eigenvector centrality, using a slight modification of the PageRank algorithm (Page et al., 1998).
- LexRank first generates a graph representation of an input document, where nodes represent sentences and weighted edges represent the pairwise similarity between sentences, ($\mathbf{B} \in \mathbb{R}^{N \times N}$; $\mathbf{B}_{ij} = \mathbf{B}_{ji}$ is the non-negative normalized similarity score between sentences x_i and x_j)
- The eigenvector centrality of each node is calculated by normalizing the rows of $\mathbf{B}$ to form a stochastic matrix,then finding the principal eigenvector typically using power iteration.

Examples of Generated Summaries of Long-Form Scientific Articles

Pubmed

- **Abstract:** background : timely access to cardiovascular health services is necessary to prevent heart damages . the present study examined inequality in geographical distribution of cardiovascular health services in iran . methods : present study is a cross - sectional study conducted using demographic data from all iranian provinces (31 provinces) from 2012 census by the statistics center of iran (sci) ...
- **This Work:** since accessibility to intensive healthcare , in particular for cardiovascular conditions , is of utmost importance and since the appropriate distribution of coronary care unit (ccu) beds and cardiologists can be taken as a measure , the present study aims to examine the inequality of the geographical distribution of ccu beds and cardiologists in iran using the gini coefficient and the lorenz curve . the number of ccu beds and cardiologists in public sector by province in 2012 was obtained from iran ministry of health and medical education ...

arXiv

- **Abstract:** we show how to control spatial quantum correlations in a multimode degenerate optical parametric oscillator type i below threshold by introducing a spatially inhomogeneous medium , such as a photonic crystal , in the plane perpendicular to light propagation . we obtain the analytical expressions for all the correlations in terms of the relevant parameters of the problem and study the number of photons , entanglement , squeezing , and twin beams ...
- **This Work:** we showed that the quantum correlations can be tuned by means of this pc , obtaining noise reduction in field quadratures , robustness of squeezing in a wider angular range , and , most remarkably , an improvement of entanglement above threshold @xcite . in this paper , we present analytical results valid below the parametric threshold and based on linear and few - modes approximations in good agreement with numerical simulations of the full model . we have considered a pc modulation ([2kc]) and five modes , @xmath117 for the pump and @xmath118 for the signal ...

Comparison of a Generated Summary

- **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 ...
- **Cohan et al. (2018):** 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 ...
- **This Work:** hash table entry in wikipedia. there are various hash functions on strings , such as crc ,we call these methods unlimited. a method is limited , if the number of probes can not exceed some limit . any of these methods may probe indefinite number of locations , even as many as (@xmath1) in the worst case ...

* Note: Although our summaries sometimes appeared less natural, they scored higher on ROUGE metrics.

Conclusions and Key Points

- On longer summarization datasets, current abstractive neural generation models underperform unsupervised, non-neural baselines. We believe these results offer a challenge to the next generation of neural generation models, which must be able to handle longer documents and longer summaries
- Based on the success of LexRank on these larger problems, it would be interesting to see to what extent improving similarity scores, perhaps with improved sentence representations, can improve summarization performance
- Scaling neural abstractive methods up may require explicitly modeling both the graph-structure of the documents (or collections of documents) they are attempting to summarize, as well as the graph strucutre of the summaries they are attempting to generate

Dataset Statistics

Dataset	# Docs	Doc. Len	Summ. Len
CNN/DM	312K	781	56
PubMed	133K	3.0K	203
arXiv	215K	4.9K	220

Validation Results

Model	Thresh.	RG-1	RG-2	RG-L
CNN/DailyMail				
LexRank	0.1	35.26	13.09	31.56
LexRank+BERT	0.8	33.86	12.07	30.20
PubMed				
LexRank	0.1	41.75	15.60	37.51
LexRank+BERT	0.8	39.83	13.77	35.69
arXiv				
LexRank	0.1	38.69	14.22	34.80
LexRank+BERT	0.8	36.34	10.11	32.18

Test Results

Model	RG-1	RG-2	RG-L
PubMed			
Lede-6	37.11	12.85	33.78
Attn-Seq2Seq	31.55	8.52	27.38
Pntr-Gen-Seq2Seq	35.86	10.22	29.69
Cohan et al.	38.93	15.37	35.21
LexRank	**42.09**	**15.91**	**37.84**
arXiv			
Lede-5	34.25	8.70	30.44
Attn-Seq2Seq	29.30	6.00	25.56
Pntr-Gen-Seq2Seq	32.06	9.04	25.16
Cohan et al.	35.80	11.05	31.80
LexRank	**37.91**	**14.34**	**33.82**