A Comparison of Neural and Unsupervised Text Summarization Techniques

Video Link: https://www.youtube.com/watch?v=rv0KJZyJg_E

CS230 Spring 2019

Devin Cintron

Department of Computer Science

CS230

Abstract

There is an ever-growing abundance and long form documents available on the web. The ability to synthesize subsections of large volumes of texts into a concise, summarative format will enable experts and novices alike to quickly review large amounts of information in a reasonable time. In this vein, a variety of text summarization techniques have been advanced which effectively reduce long form texts into much shorter formats. The nature of the outputs of these models, however, is notably diverse.

We carry out an assessment of the differences between the summaries produced by two different techniques: the unsupervised *TextRank* algorithm and a neural Pointer-Generator Network. We evaluate the performance of these methods using the commonly used ROUGE score as well as through a surveying of human preferences. Interestingly, we find that, while the Pointer-Generator Network performs better as measured by ROUGE score (average ROUGE-1 F-score of 0.44 vs 0.35), that human evaluation found TextRank summaries to be superior.

Introduction

The number of pages on the web surpassed a count of 1 billion as early as 2014. As emerging markets continue to gain web access and as IoT devices continue to increase time spent on the web, growth will surely continue. Due to its obvious value, text summarization has been a long standing and continuous focus of much of Natural Language Processing research.

Modern summarization techniques can be considered in two particular classes: extractive and abstractive techniques. Extractive summarization techniques follow a process in which the most valuable elements of a portion of text are selected and *extracted* to form a summary shorter than the original portion. The *TextRank* algorithm as well as a number of many other lightweight unsupervised techniques have been created for extractive text summarization. Abstractive text summarization, on the other hand, is a more complex procedure in which new language and terms can be introduced which are not drawn directly from the text itself.

On the surface, abstractive text summarization is arguably much closer to human performed summarization. This assertion is drawn from the fact that it can bring out of vocabulary terms whereas extractive techniques cannot. Naturually, the more complex task of abstractive summarization has constituted the majority of recent research into the task, though extractive techniques continue to be visited for their robustness and effectiveness in particular applications.

Main Objectives

1. Summarize with Abstractive Approach
2. Summarize with Extractive Approach
3. Evaluate with Automated Evaluation (ROUGE)
4. Evaluate with Human Evaluation
5. Compare Performances of Models and Evaluation Results

1 Dataset

We use the *CNN/ Daily Mail* data set in order to compare these methods. The data set consists of articles drawn from the two news services with accompanying summaries for each article. The average length of the articles is approximately 800 words and the average length of the summaries is approximately 60 words. In order to compare against a state-of-the-art model, we compare the results of the Pointer-Generator model taken from the highest achieving parameters found by the author. This leaves us with a subset of 11,490 articles and summaries drawn from the larger total CNN/Daily Mail.

2 Methods

2.1 Extractive Baseline First-n

As a baseline for comparison of the two other models, we first design a naive "summarization" algorithm which simply takes the first n sentences of an article, where n is a random choice among integers such that the expected value of the length of the summary is consistent with our other extractive approach.

2.2 Extractive Approach: TextRank

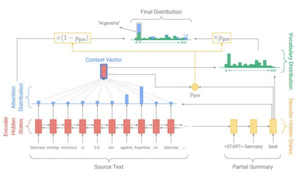
In (Mihalcea and Tarau, 2004), the authors define *TextRank*, a graph-based extractive summarization approach adapted from Larry Page's *PageRank* algorithm. A directed graph is created from a longform text in which summaritive text units represent vertices and edges represent relationships between text elements. Thereafter a ranking algorithm is applied and the best scoring elements are kept. This algorithm is similar to PageRank as scores are reflective of the number of incoming connections and the scores of the source-verticies of those incoming connections, the difference being that, in TextRank, edges values are weighted on a basis of the strength of the relationship. The full process of TextRank is then:

1. *Identify text units that best define the task at hand, and add them as vertices in the graph.*
2. *Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or unweighted.*
3. *Iterate the graph-based ranking algorithm until convergence.*
4. *Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions.*

Our implementation uses a variation of this scoring function found in (Barrios, et al. 2016) and taken from the *GenSim* python library.

2.3 Abstractive Approach: Pointer Generator with Coverage

For our abstractive summarization example, we chose the pointer-generator network approach as advanced by (See et al., 2017), in particular the authors' pointer-generator with coverage technique. Token-wise encoding is performed by a single layer bidirectional LSTM and decoding by a unidirectional LSTM. Bahdanau attention is used to compute an attention distribution and context vector. A probability for *generating* a new term is calculated at each time step via a sigmoid of the weighted product of the context vector, decoded state, and decoded input, as well as a bias. Coverage is an additional component which functions to prevent any particular section of the document from receiving an imbalanced amount of attention – it is the ongoing sum attention distributions of all previous time steps. The coverage value is then included in future calculations of attention to inform about the past distributions. In our training, we were unable to reach a performance exceeding that of the authors' so we opt to use their test results directly when comparing against the output of our extractive and naive approaches.



3 Results

3.1 Automatically-Produced Results

	Average Rouge F-Scores		
	Rouge-1	Rouge-2	Rouge-L
PG-Cov	0.4367	0.2039	0.4098
GenSim	0.3506	0.1461	0.2749
Naïve	0.39816	0.17351	0.35995

For an automated approach to comparison, we used the standard ROUGE score. Specifically we use ROUGE-1, ROUGE-2, and ROUGE-L metrics – these metrics are calculated via overlap of unigrams, overlap of bigrams, and longest common subsequence, respectively. The pointer generator with coverage approach outperforms the TextRank approach by all measures.

Our calculation of ROUGE is performed via National School of Computer Science and Applied Mathematics of Grenoble PhD student Paul Tardy's File2Rouge implementation. We believe that this implementation may be somewhat inflationary as we note that the scores appear to be inflated relative to other ROUGE implementation measurements – however, since our objective is comparison between the abstractive and extractive techniques, we find this to be permissable for our application.

3.2 Human-Produced Results

In order to provide a human measurement of evaluation, we surveyed 22 individuals on their preferences between the methods. Each individual was asked to read a selection of 3 articles, then to select which summary they preferred between that computed by the TextRank approach and the Pointer Generator approach. The order of the summaries was randomized when read to decrease the effect of any sequential bias.

Interestingly, the TextRank approach was rated as more preferred in 43 of the 66 assessments.

Preferred Abstractive:	43
Preferred Extractive:	23
Total Surveys:	66

Conclusions

- The abstractive techniques produce better results by all ROUGE metrics
- The extractive techniques produce better results by human evaluation
- Further research is neccessary to reconcile the inconsistency. Important to survey variety of contexts of summarization and methods.

Future Research

Valuable future work would include an assessment of a wider range of models for summarization. Additionally, it would be valuable to perform a greater investigation into comparisons of human evaluation versus ROUGE scoring across a diverse range of summarization techniques and contexts. While research comparing ROUGE and human scoring has been performed previously, it is important to compare these methods across different summarization contexts and methods as there is a great diversity in the nature of summaries depending upon the application.

Acknowledgements of Work

All work herein was performed alone by Devin Cintron