Extractive News Summarization

Ajanthy Jayarajan and E.Y.A.Charles
Department of Computer Science, Faculty of Science, University of Jaffna
ajanthyjaya@gmail.com

Abstract

Autonomous text summarization is an area of research with the aim to reduce the content of one or more documents. This study explores the graph based, feature based and cluster based approaches and proposes an ensemble model with combined features. The proposed model was evaluated using ROUGE Metric on BBC News Summary and found to be outperforming individual models.

Introduction

Text Summarization is a method to shorten the large amount of facts into a concise form by the process of selection of vital information and neglecting of insignificant information. Thus, it reduces the reading time [1]. Automatic text summarization process can be classified on multiple basis such as input type, output type or purpose [2]

- Single and Multi-Document Summarization
- Extractive and Abstractive Summarization
- Inductive and Informative Summarization

Objective

The main objective of this research project is to develop a summarization method that can produce informative and effective summaries to single document with high precision.

Methodology

The system generates an extractive summary for an input article through the steps [Figure -2]

- The news articles are pre-processed
- Then, an intermediate summary is created individually by using graph based approach, feature-based approach and cluster based approach
- The individual summaries are aggregated and the redundant sentences are removed
- At the end, the final summary is produced

- In graph theoretic approach [Figure-1], the text is word tokenized and similarity matrix across sentences is generated using the cosine similarity between sentences. The sentences are represented as nodes and the similarities between the sentences are represented as edges in the sentence similarity graph. Then, each node in the graph is assigned a score using the PageRank algorithm which is an indication of how significant a particular sentence in the summary. The sentences with highest PageRank value is added in the intermediate summary which is generated by graph based approach

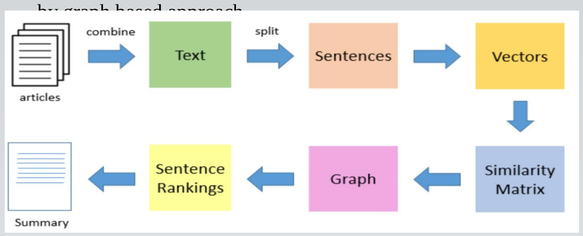

Figure 1: Flow chart of the Graph based Approach

- In the cluster based approach, a pre-trained Word2Vec embedding model is used to represent the words. By combining the word vectors of each word in a sentence, sentence vectors are created to represent each sentence. k-means clustering method is used to group similar sentences. The sentence which is closest to the cluster centre is selected as the most significant one in a cluster. The summary is made by one such sentence per group.
- In the feature based approach, the scoring of sentences is done using predetermined features such as Title word Feature, Sentence Length Feature, Sentence Location Score, Sentence Uppercase Availability, TF-ISF Score, Jaccard Similarity, Proper Noun Score, Cosine Similarity and Numeric Token Availability. The sentences are ranked according to their total feature score values.

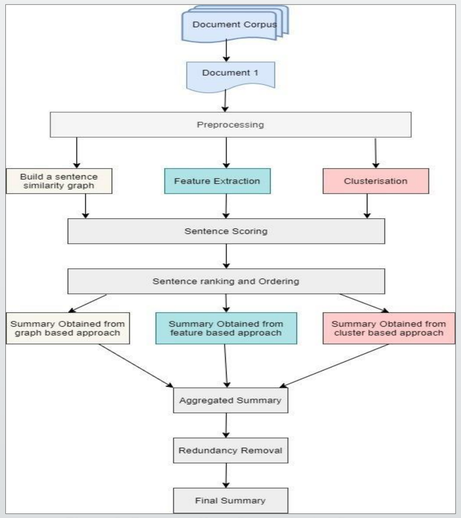

Figure 2: Extractive News Summarization Process

Experimental Setup

Dataset

The BBC News Summary Extractive Summarization of BBC News Articles [7] is used as main dataset for this work. BBC News Summary Extractive Summarization was created using a dataset used for data categorization that consists of 2225 documents from the BBC news website corresponding to stories in five topics.

Evaluation

ROUGE metrics was used as an evaluation metric. ROUGE-1 metric refers to the overlap of unigrams between the system summary and reference summary. ROUGE-2 metric refers to the overlap of bigrams between the system summary and reference summary. ROUGE-L metric determines longest matched sequence in both summaries.

- **Recall** is the ratio between number of overlapping words and total number of words in reference summary
- **Precision** is the ratio between number of overlapping words and total number of words in system generated summary
- **F-Measure** is [2] the harmonic mean of precision and recall

Results

In the cluster based approach, similarity between a sentence and a cluster centre was calculated by using the cosine similarity measure which provided the highest accuracy when compared with other similarity measures. The figures 3, 4 and 5 shows the precision, recall and f-measure values of each approaches. The Table – 1 shows the f-measure values of approaches.

Figure 3 : Evaluation results - Precision

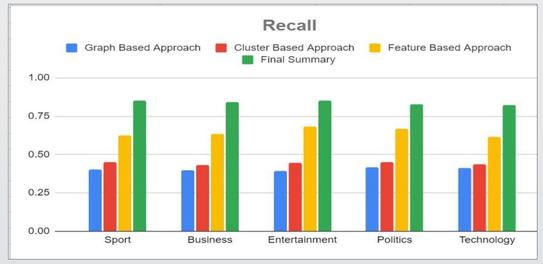

Figure 4 : Evaluation results – Recall

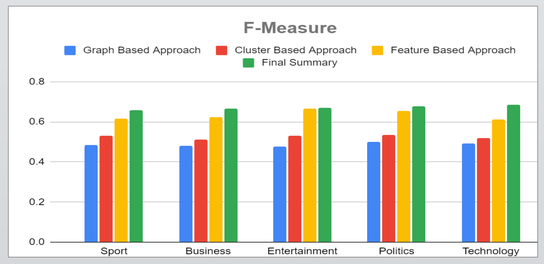

Figure 5 : Evaluation results – F-Measure

Literature Review

- ✓ Hans Christian (2016) et al used [4] Term Frequency - Inverse Document Frequency method and got 67% accuracy.
- ✓ K. Shetty et al (2017) [5] used a Cluster-based approach on CNN Corpus and got a precision of 0.284 and recall with 0.304
- ✓ Rajas P. Chiney, R. Prasanna Kumar. (2020) [2] used a feature-based approach on BBC datasets with the precision of 0.7404 and recall of 0.6885
- ✓ Devi Krishnan et al (2019) [6] used a feature-based approach on BBC News Summary and got a precision of 0.597 and recall of 0.488
- ✓ V. Alwis (2018) [7] used Graph and feature-based approach on the news articles from different e-news sites about the same topic and got a precision of 0.752 and a recall of 0.813

Clusters	Graph based Approach	Cluster based Approach	Feature based Approach	Brute Force Method	Final Summary
Sport	0.5148	0.5536	0.6476	0.3564	0.6725
Business	0.5121	0.5399	0.6614	0.3356	0.6745
Entertainment	0.5073	0.5536	0.6462	0.3503	0.6745
Politics	0.5690	0.5349	0.6887	0.3473	0.6928
Technology	0.5339	0.5535	0.7038	0.3570	0.6943

Table 1 : F-Score of Clusters on each approaches

Discussion & Conclusion

- This research has proposed and evaluated a method to produce a condensed form of a text by combining graph based, feature based and cluster based approaches
- Rather than using the individual implementation of each approaches the combined approach comparatively produces high precision, recall and f-measure
- In the cluster based approach different clusters may represent different subtopics. This approach does not depend only on similarity to cluster for sentence selection but also considers the overall document content similarity.
- Graph based methods are much more easy to visualize, understand and are more redundant than the other summarization methods.
- Our proposed method combined the three individual approaches; produced the final summary; evaluated using ROUGE Metric and produced 0.587 as precision value, 0.794 recall value and f-measure as 0.672, which produces comparatively better result than using single approach
- Future work
  - The summary needs to be evaluated not only syntactically but also semantically to check whether it produces good results
  - In this research, only the BBC News Summary was used. However, this approach can also be converted to any domains by not considering specific features like Sentence Location Feature

References

[1] S.Babar and P.D.Patil,"Improving performance of text summarization", Procedia Computer Science, vol.46, 2015.
[2] Rajas P. Chiney, R. Prasanna Kumar, "Extractive summarization approach for news articles based on selective features", International Journal of Advanced Science and Technology, vol.29, no.6, pp. 8215-8224, 2020.
[3] H. Christian, M. P. Agus and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (tf-idf)", ComTech: Computer, Mathematics and Engineering Applications, vol. 7, 12 2016.
[4] K. Shetty and J. S. Kallimani, "Automatic extractive text summarization using K-Means clustering", in 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017.
[5] D. Krishnan, P. Bharathy, Anagha and M. Venugopalan, "A Supervised Approach For Extractive Text Summarization Using Minimal Robust Features", *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 521-527, 2019.
[6] V. Alwis, "Intelligent e-news Summarization", in 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), 2018.
[7]Pariza Sharif, BBC News Summary [Dataset] https://www.kaggle.com/pariza/bbc-news-summary, 2018.