1: Automatic Text Summarization of News Articles

Pulkit Khandelwal — Student ID: 260717316 — pulkit.khandelwal@mail.mcgill.ca

***Abstract*—SumBasic and its variants have been implemented to summarize various documents to have a multi-document classification of new articles.**

I. Introduction

Automatic text summarization algorithms have been implemented to automatically summarize the contents of news articles. Four clusters(groups) of news articles have been selected and each cluster has four articles pertaining to the event/news of the cluster. The events are: Demonetization of the Indian Currency, Snapchat's new sunglasses, Indian Space Research Organization's recent success and lastly Donald Trump's remarks on Pakistan's Prime Minister.

Summarization is defined as the shortening of the original article to a brief gist which is both informative and precise. Here the summary's word limit is 100.

II. Algorithms

Extractive Summarization is used to analyze and refine the content of the articles to produce a concise synthesis. The three algorithms are:

1. **SumBasic**: It uses uni-gram frequencies of each word. For each sentence in the source text, the average word probabilities are calculated based on uni-gram probabilities. The sentences are then ranked according to these average word probabilities. The highest ranked sentence is then added to the summary. A non-redundancy update is performed which down weights the words (by squaring their probabilities) which were just used in the summary. This step helps in avoiding repetitiveness of words in the summary. This procedure is continued until the summary length is reached.
2. **simplified**: A variant of SumBasic is the simplified version which does not involve the non-redundancy update and thus simply returns the top ranked sentences according to the average word probabilities until the word limit is reached.
3. **leading**: The leading sentences of each of the articles within a cluster are selected to be included in the summary.

III. Data and its pre-processing

A python library called newspaper[1] is used to extract the News articles. A url to the article is to be provided and it returns the text of the article. Google News is used to track events and news. The data is stored as given in the assignment instructions.

Sentence segmentation, word tokenization, lemmatization according to POS tags, conversion to lowercase, removal of stop words, removal of punctuation has been done as part of pre-processing of text.

IV. Dicussion

1. **SumBasic**: This algorithm is dependent on the frequency of the words. It gives a coherent summary and does not have word redundancies because the words are down weighted after each iteration. The summary is very informative and tries to accurately reflect the original content. I see that the grammaticality of the sentences are apt. The sentences which have the most frequent occurring words are placed at the beginning of the summary. The algorithm might be less robust to contextual information of the source text as it is purely based on word frequencies.
2. **simplified**: A simplified version of the SumBasic algorithm is implemented where the non-redundancy update is not performed. To explore further I have tried two versions for this algorithm. One is where we take the first few sentences of the source text after ranking the sentences without any non-redundancy update. This summary tends to be very repetitive due to redundant phrases. The second where the first sentence gets selected again and again (because there is no non-redundancy update). This summary is not acceptable at all. The original SumBasic outperforms the simplified version.
3. **leading**: This algorithm returns the first few sentences of any one of the document in the given cluster until the word limit is reached. This essentially reproduces the first paragraph or so of the given article. A better approach to this baseline algorithm will be to run the SumBasic algorithm on each article of the given cluster and then take the leading sentences of each of the article in the given cluster to form a summary.

V. Conclusion

So, we have seen how these three methods perform and see that contextual information and critical analysis is not the crux of the summary. I would like to suggest that one should given due importance to each paragraph separately. Within the word limit one should pick up the important sentences from each paragraph because each paragraph has a different role to play. For example: the first paragraph gives an introduction, the middle few paragraphs explains the content in depth and gives critical arguments (both positive and negative). The final paragraph then concludes the articles with a take-home message. Hence, it is very important to give importance to each paragraph. Perhaps, run SumBasic on each paragraph or two separately and see the results! This might also help to remove redundancies and thus give concise summary of the text. Abstraction Summarization should be used to get much more contextual summaries. Also for better judgment ROUGE and Pyramid methods can be used to evaluate summaries.

References

[1] https://github.com/codelucas/newspaper
[2] http://www.cs.bgu.ac.il/ elhadad/nlp09/sumbasic.pdf