# Comp550 Natural Language Processing

Assignment 4

Yiran Mao

Mcgill ID: 260850827

Date: 2018-11-24

# Q1: Reading Assignment --Multi-document Summarization

In this paper the authors mainly want to confirm the impact of frequency in summary creation by isolating the contribution made by frequency information from that of other features. Then they created a summarization system------SumBasic.

They firstly investigated the impact of frequency at the word level. They calculated the percentage of the top n frequency words from the input that appear in the human models and in an automatic summarizer. They found that the high frequency words are more likely to appear in the human model, indicating that frequency does impacts a human's decision. Then they also found that the words that human summarizers agreed to use in their summaries include the high frequency. After doing the word level, they turned to the semantic content units. They measured how predictive the frequency of content units is for the selection of it in a human summary and finally found that content unit frequency is more highly predictive for inclusion in a human summary than in an automatic summary. Also, by comparing a pyramid derived from the input documents and a pyramid built from human summaries as the original method prescribes, they found that frequency alone cannot fully explain human behavior. Based on the research above, they introduced SumBasic. The algorithm for SumBasic uses a greedy search approximation, and it has two components: a sentence weighting mechanism based on word frequency, and a mechanism to adjust weights after each selection(Pnew = Pold*Pold). By updating the probabilities in this intuitive way, they allowed words with initially low probability to have higher impact and gave a natural way to deal with the redundancy in the input. When using the generic summary task in 2004 DUC as test data and the ROUGE automatic metric for evaluation, SumBasic performed significantly well. And when using the MSE evaluation, the differences were not significant. At the end of the paper, they indicated that re-ranking and duplication removal can have a major impact on the final performance.

This model has some limitations. Firstly, it only used frequency as feature, however, frequency is not the only factor which impacts the human summarization behavior. In future application we can implement some other features such like position of the sentence in the passage, sentences simplification, and some other features which can help promote less frequent content. Secondly, it can only extract important sentences but mostly human summarizers would generate new sentences. I think we can combine this method with some generative model.

Using ROUGE as an evaluation measure has several advantages: it has been shown to correlate well with human judgements, it is a cheap and fast way to compare systems. However, it does not allow us to examine the issue of duplication removal in further detail.

Three questions: Would it possible to use the same algorithm to isolate the contribution of other features? Is there any better method to calculate the sentence weight? How to deal with the orders of selected sentences in the summary?

## Q2: Multi-document Summarization

In this experiment, I implemented 4 methods: original SumBasic, best-avg (picks the sentence that has the highest average probability in step 2, skipping step 3), simplified(without update for word_probs and sentence_prob), leading(takes the leading sentences of one of the articles).

Comparing the method original SumBasic with best-avg, the difference is that the result of original SumBasic always includes the current best word. From the perspective of crawling the centre meaning of the article, I think the result of orig performs better which indicates this constraint really helps and is useful. Using update method, the best word is different in every iteration and to a large extent implied the focus of the article. And in all my clusters, the results of these two methods have no duplicate sentences, which indicates that the non-redundancy update really works perfectly.

When it comes to the simplified method, it always holds the word scores constant and does not incorporate the non-redundancy update. Without any other designs, the model would choose the same sentence again and again. So I decided to pop the sorted sentences queue to remove the redundant sentence. This method is simple and can roughly extract some important sentences. But its result was wired in some of the clusters.

And the common disadvantage of the three methods above is that they have no mechanism for processing the order, so the sentences were presented in an illogical order (As mentioned in Q1). Therefor the summary is not well organized. I think maybe we can use other features (such as the position of the sentence in the article) to solve the problem.

Relatively speaking, the leading method seems more logically for it directly follow the sentence order of the original article. However, its performance is not satisfactory. 100 words' leading sentences is too short compared with a long article, the summary sometimes have not yet touched on the central idea of the article, but only some background introduction.

Overall, I believed the best method is the combination of the original SumBasic and a proper mechanism for dealing with sentences order.