Chengkai Zhu
260775967
COMP 550 Assignment 4

1 Question 1: Reading Assignment Multi-document Summarization

- **Summary.** This paper [1] first discusses the role of frequency in designing a summarization system by studying into the impact of using frequency on various aspects of summarization. Then the authors proposed a summarization system called SumBasic and compares its performance on DUC 2004 and 2005 MSE with those from other summarization systems. The impact of duplication removal and re-ranking skills is also discussed in the following section showing SumBasic has a good a strategy of removing redundancy compared with the other three mentioned systems.

  SumBasic mainly has two components: 1) one selecting sentence based on the weighted average of words' probability. 2) The other responsible for re-weighting the word frequency which can deal with redundancy in the input. The underlying theoretical support of SumBasic from the previous discussion which indicates that 1) Words with high frequency from the input are very likely to appear in human model's summary. 2) It is reasonable to maximize the overlap between the summary from human model and automated model by including those high-frequency words.

  Rouge-1 and pyramid scores are adopted as evaluation metric respectively for the experiment on DUC and MSE and remarkable performance is gained under SumBasic System.

- **Limitations.** This approach has several limitations: 1) Words with high frequency does not necessarily appear in the summary 2) It may suffer when topics are composed mostly using synonyms or words substituted by hypernyms/hyponyms, because those words would be counted differently when computing the probabilities.

  One solution I could think of for 1) is to combine selection components from other systems which are not based on word frequency. On the other hand, the problem from 2) could be tackled with some techniques which can merge the probabilities of synonyms or hypernyms/hyponyms using some external knowledge.

- **Pros and Cons of ROUGE.** The main advantage of using ROUGE is that it's simple and efficient without the need of human efforts. And for the evaluation of summary, we usually care more about the recall rather than precision. Some limitations of ROUGE would be 1). It does not take the word order into account so that a poorly organized or syntactic-wrong sentence might also have good metrics 2) It ignores redundant information so that the sentence which is not very concise and contains some less irrelevant information would also get good score. 3) It cannot deal with synonyms.

- **Questions.**

  1. I'm confused whether the stop words should be removed in the phase of calculating word frequency. Because the only mentioning of removing stop words comes from the evaluation part.
  2. I'm curious if there are some correlations between the SumBasic and the systems which are not significantly different in performance in Table 5.
  3. I'm also interested in how to conduct duplication removal techniques on the two other systems under comparison: LexRank and DEMS. And how is the method different from the re-ranking of SumBasic?

2 Question 2: Multi-document Summarization

- **Introduction**
  In this project, four different methods are explored against Multi-document summarization tasks: leading method, original SumBasic, Simplified SumBasic and Modified Sumbasic. The documents are collected from Google News by hand categorized by topics. A total of four clusters while each contains three articles is experimented on for this project.

- **Pre-processing**

  As for pre-processing, sentences are segmented into tokens with punctuations removed. Contractions like "We'll" are separated into two parts ("we" and "'ll"). Stopwords of English are removed before probability is computed for each token.

- **Implementation and Evaluation**

  The baseline method, the leading sentence method, is implemented by iteratively selecting sentence from the head of each article of the same cluster, though the summary is found to be a bit redundant in semantic sense but the general ideas are well included. I have also tried doing by selecting the leading sentence from an article with the maximum length. Since the sentence is chosen from the article in original order, the summary appears in a logical order but gradually goes into details as the length increases, which is not appropriate for a summary.

  This baseline method actually performs very well. This is intuitively reasonable since the leading part of a news article often acts as a summary of the content. As the lack of reference summary, I compared summaries from the rest of the methods with the one from the baseline method using ROUGE-1 metric[2]. Scores by each cluster with all three methods are listed in Table 1 along with their average.

  The original SumBasic method generally performs the best if ROUGE score is only considered.

  The simplified SumBasic method, if not removing selected sentences from candidates, would yield the same sentence repeatedly until the iteration terminates due to the word length limitation. To make the summary in a more reasonable manner, I did a little bit modification to remove selected sentence at each iteration. The outcome shows simplified SumBasic performs closely with the original one and even outperforms the rest two in cluster 1.

  The best-avg method, which only considers the word scores of all sentences, has the lowest ROUGE-1 score compared with the rest two methods.

Cluster	**orig**	**best_avg**	**simplified**
1	0.0536	0.0581	0.0725
2	0.0859	0.0910	0.0834
3	0.0950	0.07097	0.0704
4	0.0869	0.0722	0.0864
Average	0.08035	0.07306	0.0781

Table 1: Rouge-1 Score for All Methods

- **Discussion**

  - **Non-redundancy update.** Non-redundancy do work in most of the cases, compared with simplified method which does not include such feature. In particular, it largely prevents a sentence from appearing again in the summary and helps diversify the summary. But sometimes, a sentence with similar words would be affected by such mechanism and might not be selected in the following iterations, which causes the loss of information.
  - **Including the sentence with highest-probability-word.** Always choosing the sentence with the highest-probability-word seems an important technique. Since a cluster of articles usually revolve around the same event, the highest-probability-word (often the place where the event took place or the name involved in the event) would always appear in the summary part. That is possibly the reason why the best_avg method suffers without using this technique.
  - **Order the summary sentences** Among all methods under comparison, only the summary coming from leading method is presented in a relatively logical sentence order, where the rest usually appears in an odd order. This is reasonable since only the leading method takes order from the original article where sentences are already arranged in a readable order. One approach I could think of to order the summary for the rest of the methods is to consider coreference between sentences. In other words, a sentence containing the anaphor usually appears after the one consisting its antecedent.

References

[1] A. Nenkova and L. Vanderwende, “The impact of frequency on summarization,” *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, vol. 101, 2005.

[2] C.-Y. Lin, “Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?” in *NTCIR*, 2004.