Comp 550 Assignment 4

Jonathan Pearce, 260672004

November 28, 2018

Problem 1a. This paper by Nenkova and Vanderwende examines the impact of frequency in various features of summarization and the role of frequency in summarization system design. In the paper, Nenkova and Vanderwende describe their summarization system SumBasic, and compare its performance against other summarization techniques. The authors begin by discussing the methods that have been used historically to identify sentences for inclusion in multi document summaries. The papers that explain these past methods tend not to report the relative contribution that term frequency makes in the summarization process, even though it is commonly used as a feature to help determine sentence importance. Using the multi-document summarization task from DUC 2003 Nenkova and Vanderwende show that frequency is one factor that impacts a human's decision to include content in a summary and that current top performing automatic summarizers place less importance on these high frequency words, relative to humans. They conclude that the overlap between human and automatically generated summaries can be improved by targeting high frequency words, this serves as the main motivation for their summarization technique, SumBasic. SumBasic is a greedy search approximation that utilizes frequency based sentence selection along with a method to reweight word probabilities during the summarization process in order minimize redundancies. Nenkova and Vanderwende, show that SumBasic uses high frequency words in it's summaries at a more even proportion to that of humans. They proceed to rigorously evaluate SumBasic against other automatic summarizers using data from DUC 2004 and the ROUGE-1 automatic metric. SumBasic performs significantly better than 12 other systems and it's mean score is better than all but one system (peer 65). Nenkova and Vanderwende further evaluate SumBasic with another dataset from a 2005 conference. In this second evaluation SumBasic is seen to have successful results in comparison to other systems using evaluation metrics such as ROUGE-2 and the pyramid score. The authors finish their paper by demonstrating that SumBasic's method for dealing with duplication removal is more effective than techniques used in more complex automatic summarizers and that duplication removal has a significant impact on SumBasic's performance. Nenkova and Vanderwende conclude that SumBasic exceeded their expectations and that future work should focus on isolating other interesting linguistic features (similar to their work) and that duplication removal should be given more attention.

Problem 1b. Because SumBasic's design was driven by word frequency it relies on the fact that the frequencies obtained from the input are representative of the importance of those words. Therefore, it makes sense that SumBasic's performance is proportional to the amount of input data it receives, since the more data it obtains the more accurate the word frequencies will be at representing the true importance. A solution to prevent this issue would be to compare word frequencies in the input to word frequencies from a large corpus of text. This would allow the summarization system to identify which words in the input text are being used more frequently than usual and place more weight onto those words. Further it can be seen from Table 5 and 6 in the paper that SumBasic is not the absolute best automatic summarization system. Therefore it is reasonable to conclude that using frequency alone is not quite enough to create a state of the art summarizer. An attempt to combine SumBasic's framework with other features that have been demonstrated to be successful in summarization would be interesting, as it's possible that including a few more basic features to SumBasic would elevate its' performance.

Problem 1c. The greatest advantage to using ROUGE as an evaluation metric is that it is automatic and does not require costly human analysis or intervention that other metrics require such as the pyramid method. This advantage is demonstrated in Table 6 of the paper where the pyramid method was only used on 10 test sets whereas the automatic methods such as ROUGE-2 were applied to all 25 test sets.

A disadvantage to ROUGE is that it is recall oriented, therefore it is advantageous for summary systems to extend their summaries all the way to the word limit even if it means cutting a sentence off halfway through. These extra words at the end of the summary will not improve the summary quality because it's a broken sentence, but ROUGE only notices the extra n-grams. Adding a partial sentence to the end of your summary will only increase the summaries' ROUGE score.

Another disadvantage of ROUGE is that it does not consider the summary as a whole, ROUGE only evaluates a summary by n-grams. Therefore ROUGE does not ensure that the summary is in fact an accurate representation of the input text.

Problem 1d.

- How was the formula in Step 4 of SumBasic determined? Were other mathematical scaling techniques examined?
- Is there a correlation between percentage of the top n frequency words used in summary and summarizer performance? On page 2 they show that good automatic summarizers use high frequency words less than humans, but they do not examine whether increasing the use of high frequency words directly causes an increase in automatic summarizer performance. Is it possible there are hidden variables not being accounted for?
- Would changing the sentence score formula to only average the n highest word probabilities improve summary quality? This would prevent one or two words with very low probability from keeping an otherwise good sentence out of the summary.

Problem 2. Part 2 of Assignment 4 had us implementing the original SumBasic summarization algorithm in order to summarize multiple news articles on selected topics. I choose 4 different topics and 3 different articles per topic, all of sufficient length to ensure the summarization technique had enough input. Beyond the standard SumBasic implementation, we also summarized our topics using two different minor modifications of SumBasic. The first being a version of the SumBasic algorithm that picks the sentence that has the highest average probability regardless of what the highest frequency word is. The second was a simplified version of the system that holds the word scores constant and does not incorporate the non-redundancy update. Finally to compare our 3 SumBasic variations more fairly we implemented a basic baseline algorithm that selected one article per topic and simply took the first n sentences such that the last sentence made the summary greater than or equal to 100 words in length. This word limit was used for all 4 methods.

To begin, since all of these methods are extractive summarizers the grammaticality of the individual sentences was as good as the grammaticality of the sentences in the articles. Therefore each sentence in all the summaries felt natural to read on there own. Further there was no effort in attempting to properly order the selected sentences in a coherent manner for the three SumBasic techniques (leading method is already in order), therefore the evaluation below will mostly focus on summary content since linguistic quality has been covered by these two points.

The original (standard) SumBasic summaries proved to be relatively successful. The summary content for cluster 2 and 4 was quite good, and cluster 1 and 3's content was reasonable as well. The original SumBasic technique is able to generate summaries that reflect the original content accurately. It is possible to discern the general topic of the cluster by reading these summaries. As expected, the original summaries do not contain redundant content. The one area where the summaries could be improved is including the most important content from the articles, occasionally details that most likely would have been included by a human are omitted from the original SumBasic summaries.

The next method we explored was the best average sentence version of SumBasic. This method also produced good results in terms of summary content. Similarly to the original method, these summaries were not redundant, and reading these summaries would give someone a pretty good idea of what the general theme of the cluster of articles was. The point where we saw drop off in performance from the original method was in extracting the most important content from the articles, this technique did a worse job than the original method. Therefore we conclude that when the non-redundancy update is being used, ensuring the selected sentence has the highest frequency word in it helps improve summary quality.

The last SumBasic variation we explored was the version where we ignored the non-redundancy update and kept the word frequencies constant throughout the summary generation. As expected the summaries produced are in fact redundant as each sentence contains the highest frequency word from the entire cluster. This means that even with an intelligent method for ordering sentences the summary would read very poorly and would not be anywhere close to a human standard. Withstanding this extreme redundancy the general topic of each cluster is still possible to discern from each summary. However this method is by far the worst of the three SumBasic methods at including the most important details from the articles. These findings support what Nenkova and Vanderwende say in section 5 of their paper and demonstrates that the non-redundancy update is critical in producing summaries of reasonable quality.

Our baseline model produced good summaries as well. Since this method simply took the first few sentences from one of the articles in a cluster and made that the summary, they are all very coherent and flow well when you read them, unlike the summaries from the SumBasic methods. Within each cluster I choose articles that focused around the same event and therefore these baseline summaries capture a fair amount of detail, if my articles had been spread more broadly across an entire topic the summaries would not be as successful. Another downside to this baseline model is the end of the summary is sometimes abrupt and unnatural, this makes sense since the summary simply extends to the first sentence that goes beyond the word limit.

As discussed above there was no effort in attempting to properly order the selected sentences in a coherent manner for the three SumBasic techniques. Therefore the baseline method's summaries seem much more natural when read. A simple idea to help improve the readability of the SumBasic summaries would be to keep track of the position of each selected sentence in its original article, specifically the proportion of the article that proceeds it (e.g. if we select the 3rd sentence from a 20 sentence article, this sentence would receive a score of $3/20 = 0.15$). This would allow us to easily order the selected sentences, the sentence with the lowest proportion would be the first sentence in the summary and we would proceed until the highest proportion sentence is added to the very end of the summary. Therefore sentences found at the beginning of their article would be near the beginning of the summary, similarly for middle and end sentences. Using proportions instead of ranks would prevent us from penalizing short articles (e.g. with ranks a sentence selected from a short article would always be placed near the start of the summary.