COMP 550 Assignment 4

Yves Blain-Montesano 260745418

November 28, 2018

1 The Impact of Frequency on Summarization

The paper explores the inclusion of frequency information of text to the quality of a document or multi-document summarization. It begins by explaining that high-frequency words are most likely to appear in summaries generated by human strategies. Then-state-of-the-art models did not include as many high frequency words (roughly 84%) as humans (roughly 90%), and they included content words that human strategies did not. They take into account the agreement between human strategies and find that words unique to a human strategy are usually low-frequency.

The paper also looks beyond only word frequency and co-occurrence in human summaries by examining co-occurrence of semantic content units in input text and human summaries. They find that the frequency of content units (an atomic fact in text form) points to their importance, and to their inclusion in human strategy-generated summaries. Nevertheless, they find that frequency alone is not sufficient to explain summarization behavior.

They set out to build a summarization system, named SumBasic, incorporating only frequency information as a baseline to isolate its contribution. The system takes, from all input documents, the relative frequency of each word as its probability. It assigns to each sentence the average probability of all words in it, and picks, from the sentences containing the highest probability word, the best-scoring sentence. Then, to reduce redundancy of sentences in the summary and allow less frequent words to impact the summary subsequently, it squares the probability of all words in the chosen sentence. It repeats this until the desired summary length is reached.

A limitation of this approach is that it may be possible to consider more than unigram word probabilities. If frequency is meant to be tested with this baseline method, an extension may be to include longer n-grams, and to explore the inclusion or exclusion of some stop words. While they provide motivation for the redundancy update, it is still a heuristic to square a word's probability. As well, surface frequency and probability is used to indirectly model semantic content units' frequencies. Perhaps this can be more directly modeled.

They evaluated SumBasic's performance on the ROUGE-1 metric, which has the advantage of being a simple metric and to correlate well with human judgements. However, it is still limited to unigram overlap between input and summary text, and does not explicitly model semantic similarity or even sentence well-formedness which may be useful for modern sequence to sequence summarizers.

Avenues of query, while remaining a frequency-oriented baseline, might be to explicitly model the frequency of semantic units from parses of input documents, to explore what degree of markovization is useful for n-gram word frequency, or to include more surface information such as parts of speech.

2 SumBasic

My intuitive judgement of the summaries of the methods is that the inclusion of the best word helps overall. It seems to remove noticeably more redundancy than the `best-avg` method. I noticed little difference between `simplified` and `orig`, though it seems that the redundancy update provides less context, as it might select a sentence with high average probability but less information than a subsequent sentence with some of the same words, whose probabilities are now too small for it to be selected. This hurts coherence and informativity. All three of the above methods have issues with grammaticality and coherence, as quoted speech including multiple sentences was split. Segments of these splits are sometimes given in the summary alone. The `leading` baseline is certainly grammatical and as coherent as the source article's leading text, though does not include as much information as the other three methods. Overall, I judge the `simplified` method to be the best.