ASSIGNMENT 4: [COMP550]

JOSEPH D. VIVIANO

1. Book Report: Multi Document Summarization

1.1. **Summary.** SumBasic exclusively uses term frequencies when deciding what to include in a summary. Previous methods also used frequency information, but the authors established that humans use **10-20% more** of the highest-frequency words when compared with machines. Also, the probabilities of unigrams that were used by more humans were on average higher. But, while there was a clear preference across summarizers to use high probability words, all summarizers made use of a larger number of lower frequency words. The authors conclude that a better algorithm would use term frequency to select sentences for summarization, with a mechanism for selecting lower frequency words when appropriate. A similar pattern emerged when the authors looked at 'frequency content units', which are atomic facts expressed via a collection of words.

The authors introduce SumBasic, which **A)** calculates the unigram probability distribution over the input words, **B)** selects the sentence with the highest mean probability for the summary, **C)** updates all *selected* word probabilities by squaring them (effectively upweighting sentences with unseen words), and **D)** if required, goto B. Using the DUC data and the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, the authors establish the good performance of this simple method.

I thought the rationale was solid and thought using human-machine experiments was a clever way to inspire SumBasic's features. It would have been nice to see some examples of sentences including low frequency words that were chosen by humans but ignored by SumBasic, as table 4 makes it clear that SumBasic relies more heavily on frequency than humans do.

1.2. **ROUGE Applicability.** ROUGE, being a recall-based measure, is well-suited to evaluating auto-summarization because it evaluates how much the the resulting summary overlaps (at some N-gram level) with the target summaries. It also implicitly weights n-grams that occur in multiple documents higher, which is a reasonable thing to do as it relies on consensus. However, it does not penalize the inclusion of n-grams that are not in many target documents explicitly. This is only controlled by setting the maximum length of the summary. A measure that takes both precision and recall into account would be able to more directly control this. Furthermore, a good summary might make use of all the *disjoint* information across the target documents, i.e., to give a compact overview of all the unique pieces

Date: Nov 2018.

of information across the targets. In certain cases this would be more useful than focusing on the n-grams that occur in all documents, as this might represent the least interesting information. In the end, this is application specific.

1.3. **Questions.**

- The lower-frequency terms only used by a single user: are these synonyms or similar? It is possible that 'lower probability words' are sometimes simply words that have multiple equivilant entries in the language and do not really represent unique concepts?
- What features of the DUC (computerized) summarizers that beat SumBasic do you think were instrumental in their success?
- Word probabilities were down-weighted after inclusion by squaring them. Is it possible that the downweighting could be a hyperparameter to increase performance of the algorithm?

2. SubBasic Quality Assessment

2.1. **General Quality Assessment.** In general SumBasic summaries were serviceable, but did not always contain enough information for the reader to fully understand the original text, and were not often coherent across sentences. This isn't suprising because the algorithm does little to enforce multi-sentence coherence aside from down-weighting previously-included words. The original algorithm produced generally better summaries than either of the tweaks I implemented, as they tended to contain more different pieces of information, and were generally more coherent. None of them matched the coherence of the 'leading' method, although the leading method rarely contained enough information to be a useful summary. The two features of the SumBasic algorithm manipulated (requiring the best word, and down-weighting used words) proved useful to generating coherent summaries with maximal information content, as they both managed to remove redundancy (either linguistic or content-based) from the summaries. All sentences were grammatical at the sentence level, as they were simple duplication of the original documents, but no methods were able to deal with coreference problems and other multi-sentence coherence issues.

2.2. **Inclusion of Best Word.** Inclusion of the best word was extremely important for producing summaries with the most important content, as well as coherent, in the multi-document setting. Otherwise, nearly identical phrases between documents tended to be repeated, even after probability downweighting, which had another side effect of eating up the majority of the length of the summary. E.g.,

Officials in Riyadh did not respond to a request for comment. Officials in Riyadh did not respond to a request for comment. Officials in Riyadh did not respond to a request for comment. Residence there would keep the king out of the loop on most affairs of state, one of the sources close to the royal family said. The site is isolated, the closest city of Tabouk more than 100 kilometres away. Some insiders believe he built his father a new but remote Red Sea palace in Sharma, at the Neom City development site — thrown up in a record one year at a cost of $2 billion — as a gilded cage for his retirement.

Versus the original SumBasic algorithm, which obviously used the 'best word' to select entirely different sentences:

The Saudi sources say MbS has destroyed the institutional pillars of nearly a century of Al Saud rule: the family, the clerics, the tribes and the merchant families. While the council accepted King Salman's wish to make MbS crown prince, it would not necessarily accept MbS becoming king when his father dies, especially given that he sought to marginalize council members. Residence there would keep the king out of the loop on most affairs of state, one of the sources close to the royal family said. He argues the US can't afford to alienate Riyadh due to oil and Iran.

2.3. **Non-Redundancy Update Performance.** The removing the redundancy weight updates, in my estimation, led to slightly less coherent summaries, with more jarring transitions between sentences that each produced individual atomic facts. I think this is due to many sentences containing atomic facts bring similar: e.g., *The Dow and SP 500 were down about one per cent from the start of the year, while the Nasdaq was just points higher* and *In New York, the Dow Jones industrial average lost 551.80 points or 2.2 per cent to 24,465.64* don't contain the exact same information, but are similar enough to not warrant being placed one after the other. They also share a decent number of words. When the words contained in a sentence like this are down-weighted, the next sentence is more likely to contain different kinds of information. So I would say this method primarily increases non-redundancy and coherence, and tends to contain the most information from the original content.

McGill University
Email address: joseph@viviano.ca