Emily Mulhall
COMP 550
Assignment 4
November 30, 2017

Question 1

Overall, the three summarization methods vary greatly in their success. Surprisingly, I found that the simplest method was actually the most effective. The leading summarization method, which simply took the leading sentences of one of the articles up until the word count was, in my opinion, the most successful. The beginning of an article often acts as a summary of the main points, hooking the reader and presenting the topic. Therefore, the leading sentences summarized the articles rather well. Additionally, because they were all chosen one after the other, their order was logical. In the SumBasic original and simplified versions because the sentences were chosen by their word choice, they were often presented in a very odd order. Often, some sentences, particularly quotes, could not be made sense of because they were lacking context.

In most cases, the non-redundancy update worked better than the simplified method, but was still not perfect. In the first cluster, reducing the probability of words already chosen for the summary worked very well. In particular, it prevented a repeated sentence from appearing in the summary. Often the non-redundancy helped diversify the sentences and material covered in a summary. Because the summary is built from a cluster of documents, some sentences that appear in the simplified summarization are extremely repetitive. While the main idea of the articles is stated, it is often stated multiple times rather than expanding on the topic further. This is likely due to this method's inability to account for redundancy between documents. Additionally, the simplified method does not account for conflicting information in articles. While all of the articles chosen for each cluster revolve around the same topic, some of them differ in focus. For example, in the fourth cluster one article is mainly on the last time the King of England wanted to marry an American, another focuses on the price of the ring Prince Harry proposed with, and the final one focuses on the engagement. On this cluster the original SumBasic method is far superior to the simplified method. While there are some sentences that are out of context, the majority of the summarization resulting from the original method flows very nicely. The simplified version, on the other hand, is quite difficult to understand and does not even mention the Prince or his family.

Overall, I would rank the methods from best to worst in the following order: leading, original SumBasic, simplified SumBasic. However, I was extremely impressed with how well the leading and original SumBasic methods did. The leading method had the advantage that all of the sentences were already in the right order, and this made the summarization far easier to follow than summarizations resulting from other methods. However, the original SumBasic method, in some cases, included more detail on the topic than the leading, because it picked out sentences with frequent words as opposed to the beginning of an article. Because the leading method chose the beginning sentences of an arbitrary article, it also, at times, was not representative of all articles in cluster. Though the leading tended to produce better summaries than the original SumBasic on this particular cluster of articles, I can see the SumBasic performing better than the leading on a cluster which contains articles that differ in viewpoint on the same topic because it would be more representative of all articles. However, these three methods all did far better than would one, looking at their simplicity, would have expected.

Question 2

In this paper, authors Trevor Cohn and Mirella Lapata approach the task of sentence compression through operations including deletion, substitution, reordering, and insertion. The task of sentence compression is the challenge of reducing the length of a sentence while maintaining the important information and its grammaticality. Because this is a natural language generation task, the main challenges are selecting appropriate content and an appropriate form to express that content. Typically, this task is done using only word deletion, as this is the most simple way to reduce size while maintaining information. However, when humans reduce the size of a sentence they do not strictly delete words; they reword the sentence using many

different operations. Thus, the authors seek to approach sentence compression through the generation of abstracts as opposed to extracts.

Due to the lack of corpora available for the task, the first challenge the authors face is to create a corpus. As stated earlier, much of the work on this task has been in extractive compression, which involves copying and extracting parts of the source text. The authors, on the other hand, seek to use abstraction, which is the synthesis and production of new text. To build their corpus the authors obtained manual compressions of newspaper articles. To assess the quality of these compressions the compression rates of the annotators was compared, where a compression rate is the percentage of words retained in a compression. They also compared compressions using BLEU.

The next task was building the model. The authors expanded on previous work which approached sentence compression as a "tree-to-tree rewriting task" (pg. 3). They use a "synchronous tree substitution grammar" where "each grammar rule is assigned a weight, and those weights are learnt in discriminative training" (pg. 3). This is a probabilistic context-free grammar. To extend this framework first a grammar was extracted from the corpus they had created. It was then augmented with a grammar obtained from a parallel bilingual corpus, which contained only paraphrasing rules. The model was then trained and tested. To assess the performance of it participants rated the grammaticality of the compressions and how well they maintained important information. While the model received significantly higher ratings than the baseline in terms of maintaining importance, it did not in terms of grammaticality. Yet, both the baseline and the model performed significantly worse than manmade compressions in both areas.

Lack of available data is one of the largest limitations for this task. There are many different ways to rewrite a sentence, and the documents were compressed by just one annotator. While this maintains consistency, it can result in overfitting to a particular annotator's tendencies, and it will not include different possible compressions. More time and resources must be put into building the corpus. A larger number of annotators should be hired, and they should annotate all articles. That way, each sentence has multiple possible compressions in the training data.

An additional limitation is available evaluation metrics. For summarization tasks human ratings is one of the most used methodologies. However, humans can be extremely unreliable. For example, there were only 22 volunteers who rated these sentences, all of whom self-reported being native speakers of english. Yet, there is no way to guarantee that this is true. Additionally, it is possible that participants disagree on ratings. While the authors use the mean rating to analyze results, they do not mention how related all of the responses were. One of my questions for the authors would be how similar these rankings were among different participants. While I do not believe that there is currently a better method of evaluation, having people come in person to rank these sentences could at least guarantee that they are all native speakers of english.

Additionally, I do not understand the purpose of the inclusion of the grammar from the bilingual corpus. If the grammar includes paraphrases, then doesn't this defeat the goal of this paper, to approach the task with rewrite operations in addition to deletion? Additionally, won't translating a text to a foreign language and back likely result in ungrammaticality?