Kathleen Schaefer

Johnathan Ashley

Automatic Summarization of Documents Through Sentence Selection

Introduction:

In this project we investigated summarizing text documents by selecting sentences with high information content. Using the frequency of words within the document and a training corpus as well as the position of sentences in the document, the program scored sentences and compiled the results into a summary. We then scored our different techniques using ROUGE the document summarization scorer.

Experimental Setup:

For this project we used the Legal Case Reports from the Irvine Machine Learning Repository. This dataset was chosen because it tagged each sentence for easy parsing and contained golden standard summarizations of the individual sentences. However, since our method of summarization involved choosing sentences from each document rather than words from each sentence as in the golden standard, the summarizations were not an ideal representation of what our results should look like. Nevertheless, they provided a reasonable summary for the ROUGE document summarization to score against.

We ran our evaluator on 79 documents to gather our data. Then, we inputed each generated summary into the ROUGE tool using a compilation of all the golden standard summaries of the individual sentences for the tool to compare against. As a benchmark, we also created summaries by choosing sentences randomly and comparing ROUGE’s score for both these summaries and the summaries generated by our choosing algorithm.

Results:

We used ROUGE to compare our summaries against the compilation of the provided golden standard summaries of each sentence in the document. Each summary was generated by choosing three sentences from the document. For each document, we scored a choice of random sentences to use as a baseline, a choice of sentences based on word frequency values only, a choice of sentences based on their position only, and a choice of sentences based on an equal weighting of position and word frequency. The results are shown in the Table 1 below.

The way that ROUGE evaluates summaries is by computing the precision and recall of the computed summaries versus the golden standard summaries. Because our method of summarizing involved choosing three sentences based on some metric, the likelihood of our summaries having high precision and recall with the gold standard summaries (created by summarizing each sentence individually) is incredibly low.

|  |  |  |  |
| --- | --- | --- | --- |
| runid | recall | precision | f-score |
| combined\_ROUGE-SU\* | 0.01047 | 0.04502 | 0.0066 |
| keyword\_ROUGE-SU\* | 0.00192 | 0.03404 | 0.00236 |
| positional\_ROUGE-SU\* | 0.01554 | 0.05164 | 0.01044 |
| random\_ROUGE-SU\* | 0.00491 | 0.08174 | 0.00657 |

Table 1: The results of running ROUGE on summaries of 79 Documents

Conclusions:

Examining the table above, we see that the f-score (a weighted average of precision and recall) indicates that a system based on position only is the best for summarizing this corpus. Interestingly, the random method has the highest precision, which indicates that it is best for getting a summary with the least amount of wrong information. However, for all methods, the values returned are rather low. This is a result of the gold standard summaries provided by our corpus. Where we took a holistic approach and summarized the entirety of documents, the provided summaries summarized each individual sentence. These different methods make getting a high f score difficult. However, it can be said with reasonable confidence that, of summarization methods analyzed, a position-based method is ideal, as the first and last sentences of a document tend to have the highest information density.