**Scientific Document Summarization**

**1. Introduction**

A summary is a text that represents and preserves the important information conveyed by a document in a concise manner. Automatic summarization has been an area of research for the last couple of decades and has been tackled with different approaches.

However, summarization of scientific documents is a relatively less studied application. In this report, we will address this problem and put forward a solution in the form a pipeline to process scientific documents and output their summaries.

The rest of the report is organised as follows. In Section 2, we will mention references to literature describing the work done by different research groups in scientific document summarization. In Section 3, we will provide the explanation for the proposed method. Section 4 will detail the empirical observations and comparison with different baselines. Finally Section 5 will conclude the work that has been done.

**2. Related Work**

---to do---

**3. Method**

We started by building our baseline for summarization of scientific documents using the unsupervised mechanism, TextRank. This algorithm is briefly explained next.

*3.1 TextRank*

A graph is constructed using the sentences of the documents as the vertices and the edges connecting these vertices were defined by the similarity between the connected vertices. The similarity function is defined using the vector space model. Each sentence is represented as a vector resulting from considering the tf-isf value (term frequency - inverse sentence frequency) of the words in the contained vocabulary (the words encountered in the document). These tf-isf values make up the dimensions of the sentence vectors. Then, the cosine similarity of between each pair of vectors is calculated and used as connecting edges for the vertices (sentences).

TextRank is a modification of PageRank algorithm popularized in ranking Web documents. The graph in the original PageRank algorithm is a directed graph where edges are defined by the directed hyperlinks from an origin page to a destination page. This is important, as the ranking of pages is done on the basis of the links from different pages. In case of TextRank, the edges are defined by their similarity, which is symmetric. Hence to apply this definition within the PageRank algorithm, these edges are converted into bidirectional ones. Additionally, all the edges are weighted.

Once the graph is constructed, each vertex is initialized with a random *significance value*. This value is iteratively updated depending on the significance values of the connected vertices as well as the weights of these connections. This is represented by the following equation:

----insert TextRank equation here---

----a line or two to explain the variables in the equation above---

The implementation of the iterative process over the entire network, considering that the graph is represented as an adjacency matrix, is done by calculating its eigenvector matrix.

*3.2 Baseline*

For the baseline, we constructed the graph using all the sentences in a scientific document. Then the sentences were ranked according to the TextRank algorithm and the top n sentences were chosen for summary, limited by the word (or sentence) limit on the summary.

Next, we constructed separate graphs with each comprising of sentences from a single section in the document. TextRank was separately run over each of the section graphs to rank sentences for representation from each section. The top ranked sentences were chosen from each section, limited by the word limit on the summary.

**4. Result**

The summaries generated from the baseline and the second method (section based ranking) were compared with human summaries, which were manually extracted by the members of the group. Currently, this has been for 5 scientific articles randomly sampled from the ACL Anthology. The ROUGE scores are presented in Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Recall (n=1) | Recall (n=2) | Precision (n=1) | Precision (n=2) | F-measure (n=1) | F-measure (n=2) |
| All Sentences | 0.500 | 0.301 | 0.545 | 0.330 | 0.520 | 0.314 |
| Section Based | 0.410 | 0.182 | 0.544 | 0.256 | 0.463 | 0.210 |

Table 1.

From the scores, it can be seen that the baseline (ranking all the sentences of the article together) performed better than section-based ranking. However, even with higher ROUGE scores, a qualitative evaluation of the summaries generated by the baseline showed that the statistical importance of sentences did not guarantee that such sentences would be ideal candidates for a summary.

Moreover, from the section-based ranking, it was noticed that often the sentences that ranked in the second or third place within a section seemed to be better candidates. But since the summary length is limited, these would be missed out.

**5. Conclusions**

The observations discussed in the previous section could be explained by the choice of similarity measure used for connecting sentences in the graphs. Cosine similarity, in this case, is based on the occurrence of the same words in the sentences for which the similarity is being computed. Hence, a word that is central to the theme of the paper could appear in a lot of sentences. A sentence with a lot of such common words would be similar to a lot of other sentences and hence could be marked ‘important’. Although, this is exactly what the algorithm aims to achieve, the observations show that sentences were ranked important merely due to the presence of such central words rather than the significance in terms of contribution to the summary.

Another point to note is that unlike the general method of ranking and using the top sentences, appropriate for summarizing new articles or meeting transcriptions, summarizing a scientific article requires information from all the important sections.

This was an inspiration to study the ranked sentences further. It was seen that a single level of ranking on the basis of statistical importance of words in sentences would not be sufficient to produce summaries that closely match the quality of summaries generated by humans.

In the following phase of the project, the focus would be to find features or properties of sentences that could be used to decide on a specific order in which the already shortlisted sentences should be presented so as to best represent the summary of the article.