Automatic Text Summarizing using graph based algorithm

Mohamed Farhan
mfarhan@knights.ucf.edu

Mithun Mohanraj
mohanmithun005@knights.ucf.edu

Pushkar
pushkar1992m@knights.ucf.edu

1 MOTIVATION AND PROBLEM STATEMENT

Automatic text summarization[1] involves condensing a document or a set of documents to produce a coherent, logical and comprehensible summary. Summarization technologies are very popular and are in demand due to large volumes of textual information available on the internet. In our project, we aim at implementing an unsupervised method for automatic sentence extraction using a graph-based ranking algorithm like Page Rank. Our aim is to generate a set of logically coherent sentences from a huge document that abridges the content of the document.

2 RELATED WORK

2.1 Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization

Graph-based ranking algorithm [3] is a way of deciding on the importance of a vertex within a graph, by considering global information recursively computed from the entire graph, rather than relying only on local vertex-specific information. In this paper[3], the author investigated several graph-based ranking algorithms and evaluated their application to unsupervised sentence extraction in a context.

2.2 The Evaluation of Sentence Similarity Measures

In this paper, several text similarity measures have been used to calculate similarity score between sentences in many text applications. The author of this paper emphasizes that the correct similarity judgment should be made even if the sentences do not share similar surface form.

2.3 TextRank- Bringing Order into Texts

The author discusses TextRank[4] : a graph-based ranking model for text processing, and show how this model can be successfully used in natural language applications. The paper proposes two innovative unsupervised methods for keyword and sentence extraction, and shows that the results obtained compare favorably with previously published results on established benchmarks.

3 PROPOSED METHOD

We intend to modify the PageRank algorithm [3] to make it compatible with our problem definition which is to rank each sentence in our summary to get the most expressive one. This method can be extrapolated to other domains easily even though we are confining ourselves to using data from movie reviews domain. To implement our algorithm we first identify sentences through tokenizing and retrieve all the sentences present in our text document. After that, similarity is measured between different pairs of sentences using several similarity measuring metrics[6]. Some of the scoring metrics for measurement of similarity are as follows:

- `Jaccard Coefficients Measure`
- `IDF-Cosine Similarity Measure`
- `Word Overlap Measure`
- `IDF-Overlap Measure`
- `Phrasal Overlap Measure`
- `Word2vec Cosine Similarity Measure`
- `Cosine Similarity Measure`

A weighted undirected graph is created using the results of the similarity measures calculated between each sentence pair. Then we will apply a modified version of PageRank (to take into consideration edge weights) to rank our sentences (thereby scoring the most important sentences with a higher score). Sentences with the highest score are returned.

4 EVALUATION

4.1 Dataset

We plan on using a dataset which is a set of reviews collected from RottenTomatoes[5]. It contains relevant information such as movie id, movie name, RottenTomato 1-line summary, critic reviews of movies. RottenTomatoes is a reliable source for our project as it is an online movie review database which contains reviews by movie critics and audiences.

4.2 ROUGE Evaluation

ROUGE[2] is a set of metrics used for evaluating automatic summarization in natural language processing. The metrics compare an automatically produced summary against a reference or a set of references (human-produced) summary or translation.

ROUGE-N- measures unigram, bigram, trigram and higher order n-gram overlap. ROUGE-L- measures longest matching sequence of words using LCS (longest common subsequence)

We plan to use ROUGE to evaluate our text summaries and will compare the recall and precision of our system with respect to the online text summarizers.

$$ROGUE\ recall = \frac{number\ of\ overlapping\ words}{total\ words\ in\ reference\ summary} \quad (1)$$

$$ROGUE\ precision = \frac{number\ of\ overlapping\ words}{total\ words\ in\ system\ summary} \quad (2)$$

5 EXPECTED OUTCOMES AND RISK MANAGEMENT

We expect to build a system that generates logically coherent and a comprehensible summary of movie reviews using various scoring metrics for measurement of similarities.We expect to get a clear idea of how well our model is performing using the evaluation metrics we stated above.

Evaluating such a model is challenging as a perfect summary varies among different people.Sentences generated by our model may contain mixture of relevant, non-relevant information which sometimes may contain partially redundant information.

The above mentioned limitation can be overcome using abstractive summarization (supervised) technique which we plan on implementing if time permits.

6 PLAN AND ROLES OF COLLABORATORS

6.1 Work Distribution

1.)Pushkar- Data collection, Data preprocessing, Similarity measures
2.)Mithun Mohanraj- Implementaion of TextRank algorithm, Similarity measures
3.)Mohamed Farhan- Data preprocessing, ROUGE evaluation metrics, Similarity measures, write-up, presentation slides

6.2 Timeline

1.)Data collection, Data preprocessing, Coming up with psedo-code for PangeRank algorithm :- Februrary
2.)Implement TextRank algorithm using various similarity measures :- March
3.)ROUGE evaluation metrics for all similarity measures, paper write-up, presentation slides.:-March-April

7 REFERENCES

1.)A Survey on Automatic Text Summarization-https://www.cs.cmu.edu/ afm/ Home-files/Das-Martins-survey-summarization.pdf

2.)Extractive Summarization Using Supervised and Semi-Supervised Learning-http://anthology.aclweb.org/C/C08/C08-1124.pdf

3.)Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization - http://www.aclweb.org/anthology/P04-3020

4.)TextRank: Bringing Order into Texts - https://web.eecs.umich.edu/ mihalcea/papers/mihalcea.emnlp04.pdf

5.)Movie Review Mining and Summarization -https://pdfs.semanticscholar.org/d576/d9ea5cc898d2fb4e833a630e59ff02edc7a8.pdf

6.)The Evaluation of Sentence Similarity Measures http://www.cis.drexel.edu/fac papers/dawak-547.pdf