Project Report

Course: CS 6320 – Natural Language Processing

Topic: Movie Review Summarization

Name: Aditya Borde

Net Id: asb140930

Project Report

Title: **Movie Review Summarization**

Name: **Aditya Borde (asb140930)**

1. Abstract:

This project focuses on providing summary of a movie review provided by the user on website. We generally observe, reviews are long and it takes time to read whole content. Even hasty reading won't give the exact idea about what review-writer wants to tell in the complete review. Sometimes, new person who is reading the review is unaware of few facts that writer keeps in mind and write whole review. My approach here is to provide effective summary of a movie review extracted from a web-page. It requires to provide a web-page where movie review is jotted down. Based on user preference, summarizer reduces the size of the review to make it easy to read the summary of the review which would have all the gist of the matter presented in whole movie review. Review summary can be extracted using (a) basic naïve word frequency approach, (b) semantically testing movie title with line of reviews and last one is (c) each sentence token is used to generate a graph to find similarity with others and showing best out in the final summary. This project report explains all the three parts mentioned above one by one.

2. Introduction:

Text summarization is a classic problem in Natural language programming. Now-a-days, we have to deal with huge amount of information. For any further process, human efforts will take much amount of time to just deal with this information. Main goal of summarization is to reduce the content of the document and retrieve all the important points from the document. Using natural language processing approach, this huge amount of informational data can be reduced to most important points which provides whole idea about the document. Different types of summarization applications are available. This includes languages other English too. Generation of summary of review can definitely yield improvement feedbacks from users and also it would a way to determine the people and their interests.

Given a body of text, this is automated way to generate a few sentences that sum up its content. There could be many possible good sentences that can contribute to the movie review, but the choosing the best amongst them will be moved to the summary part. On starting the summarizer, user is prompted to enter the URL link to receive the movie review as input. Basic idea is that while generating review summary, one has to take care about not to lose the main content from the review. So, to make this more effective,

reduction ratio would be asked from user by which user wants to reduce the actual size of the review. There are three different ways through which movie review summarization is done. First approach suggests the basic naïve approach to detect the important sentences. In second approach, one may desire to see the content summary relevant to the movie title. Third approach builds the graph linkage between each pair of tokens of sentences, with each link having similarity weightage. All these three approaches have different features and performance effectiveness depending on data provided as an input. Of course, each one have distinct latency times to perform the summarization process.

Making summary out of the huge text works differently than manual summarization. There are different types of summarization techniques available that can make the summary out of the text. It includes extraction-based, abstraction-based and aided summarization. Extraction-based is the automated system extracts objects from the entire collection, without modifying the objects. Where as in abstraction-based involves paraphrasing selections of source document text. In aided-based, machine learning techniques are used which are closely related to the field. In this project, my approach uses extraction-based techniques to find out the best possible summary out of the movie review.

3. Related Work:

Numerous approaches for identification of important content for automatic text summarization have been developed to date. Summarization work is expanded in various fields which would provide effective summary of the content. In many cases, providing addition information about finding out the important words from the document can result in getting the relevant summary from the content. For example, in web page summarization, the augmented input consists of other web pages that have links to the pages that someone might be interested to summarize. In blog summarization, the discussion following the blog post is easily available and highly indicative of what parts of the blog post are interesting and important. User interests are always taken into account in query-focused summarization, where the query provides additional context.

3.1 Web Summarization:

One type of web page context to consider is the text in pages that link to the one that has to be summarized, in particular the text surrounded by the hyperlink tag pointing to the page. This text often provides a descriptive summary of a web page (e.g., “Access to papers published within the last year by members of the NLP group”).The earliest work was carried out to provide snippet of each result from a search engine. Later work has extended this approach through an algorithm that allows selection of a sentence that covers as many aspects of the web page as possible and that is on the same topic.

3.2 Summarization of Scientific Articles:

Impact summarization is known for the task of extracting sentences from a paper that represent the most influential content of that paper. Language models provide a natural way for solving the task. For paper summarization, impact summarization finds other papers in a large set of papers n that cite that paper and extract the areas in which the references occur. A language model is built using the collection of all reference areas to a paper, giving the probability of each word to occur in a reference area. This language model provides a way of scoring the importance of the sentence in the original article.

3.3 Query Focused Summarization:

Query focused summarization works differently. In this type of summarization, the importance of each sentence is determined by considering two factors: how relevant is that sentence to the user question and how important is the sentence in the context of the input in which it appears. There are two types of approaches to this problem. The first adopts techniques for generic summarization of news. For example, an approach using topic signature words is extended for query-focused summarization. Graph-based approaches have also been adapted for query-focused summarization with minor modifications. New approaches have been developed for specific types of queries which identifies relevant and salient features.

3.4 Email Summarization:

Summarization must be sensitive to the type of characteristics of email which can be uniquely separated. A mailbox contains one or more conversations between two or more participants over time. While summarization of spoken dialogs, summarization needs to take the interactive nature of dialog into account; a response is often only meaningful in relation to the utterance it addresses. However, the summarizer need not concern itself with speech recognition errors, the impact of pronunciation also availability of speech features. In early research on email summarization, it uses extractive summarization to generate summary for first two levels of discussion thread tree. Later email summarizers have also been developed for a full mailbox or archive instead of just a thread.

4. Approaches:

Project uses three different types of algorithms to summarize the movie review.
For all the three document these are the common steps will be taken:
(A) *Parsing Input*: This step traverse the input document in html format. Parses the input file and identifies the movie title and actual movie review.
(B) *Punctuation and Stop Words Removal*: Next step contains detection of stop words and punctuations in the review and removing them from before applying any algorithm.

Now, we will see all the three approaches:

4.1 Naïve Based Review Summary Approach:

Algorithm uses following steps to generate the movie review summary:

(a) *Forming Word Dictionary:* After performing above steps, this algorithm create the distinct word dictionary from the document for summary. This dictionary has all the words and number of occurrence of that word in the review.

(b) *Normalization of Count:* Each value in the dictionary then normalized by highest occurring word count.

(c) *Managing Threshold Values:* This step recognizes the minimum and maximum threshold mentioned by program and removes the values from dictionary those are not greater than minimum threshold and less than maximum threshold.

(d) *Sentence Importance:* This step evaluates the sentence importance based on the summing up all the frequency values from the dictionary of words and generates the rank of each statement.

$$\text{Sentence Rank} = \sum_{i = word\ token\ in\ sentence} frequency[i]$$

(e) *Generating Summary:* Top n-rank statements are grouped to form the final summary. This value of *n* is provided through user preference.

4.2 Title Based Review Summary Approach:

Title based review summary is very effective approach when reviews mentioned are aligned with actual title semantic meaning. Note this approach would be useful when there are more than a single word (after removing stop words) present in the title of the movie. These are the following steps taken to approach towards extracting the summary out of review:

(a) *Semantic Meaning of Words from Title:* First step in this algorithm is to determine semantic meaning of each word relevant to the title. This forms the token of words from the title and semantic meaning of each of them in the title.

(b) *Collection of Relevant Word Set:* Now we can form a collection of signature of each sense of the word including definition and examples of each word. This collection needs to include actual words from the title.

(c) *Stop Words Removal from Relevant Word Set:* This collection is filtered to remove all the stop words from the collection.

(d) *Deciding Sentence Importance:* Each statement is parsed and word of tokens are generated for that sentence. Importance of the sentence is found out by number of intersecting words present in that statement.

$$Sentence\ Rank = Count\big(Set(Relevent\ Words) \cap Set(Words\ Tokens\ of\ Sentence)\big)$$

(e) *Final Summary:* Top n-rank statements are grouped to form the final summary. This value of *n* is provided through user preference.

4.3 Text Ranking Based Review Summary Approach:

Algorithm uses generation of graph having each sentence as a node of a graph. This graph is supposed to be complete graph built at run time. If weight between edges is calculated to zero, it can be again considered as no edge between those nodes:

(a) *Sentence Tokenization:* Whole input text is tokenized into sentences which are used in formation of our graph nodes.

(b) *Building a graph:* Each sentence token participates in formation of graph nodes. To create a complete graph, all possible groups having two nodes (sentence tokens) are generated.

(c) *Calculating Edge Weights:* This step estimates the edge weight between any possible two nodes. Edge weight has been calculated using Lin Similarity. Similarity expresses how two nodes are related to each other.

$$\text{Lin Similarity } (X, Y) = \frac{\text{common}(X, Y)}{\text{description}(X, Y)}$$

(d) *Node Importance:* Node importance is deciding factor to come up with the final summary. Page rank is used in determining the node (sentence token) importance. A node is said to be more important than another by checking all the weighted links going out from that node shows more weight than another.

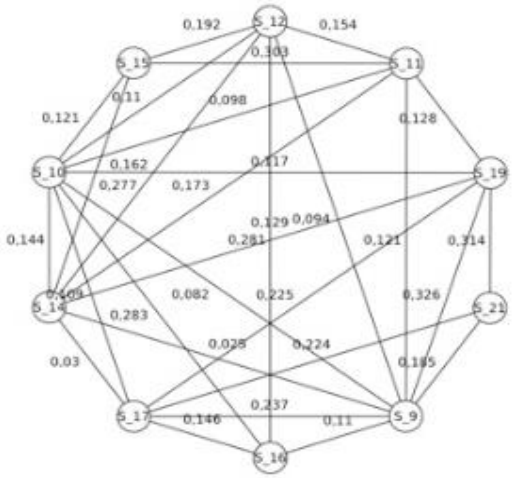

Fig. 4.3.1. Node Formation in Text Rank

(e) *Generating Final Summary:* Top n-rank nodes, which are in turn our sentence tokens, are grouped to form the final summary. This value of *n* is provided through user preference.

5. Experiment:

1. Data:

Data used for project is taken from Cornell Movie Review Data website. I used the movie review containing 450-500 words. Please find below sample data where we can observe results of each discussed approaches.

EVIL DEAD II

A film review by Mark R. Leeper

Capsule review: There's more budget than logic to this stringing together of off-beat and semi-humorous horror scenes. Creative visual concepts abound and the pace is frenetic and that makes up for a multitude of sins. I guess in some sense THE EVIL DEAD II is the ultimate horror film...sort of...I guess. Well, what can I say? It does not have much of a plot. It has very little acting, no stars, little continuity, and no logic. But it has action, horror, and black humor in massive doses. Now, THE EVIL DEAD II did have some plot. It was not it's strong suit but it was there. What THE EVIL DEAD boasted most was wit. An attacking corpse would be thrown into the fire. Then some living person would have a bout of remorse and pull it out of the fire. The corpse would look up and politely thank its benefactor for pulling it out of the fire, then continues to try and kill the living. I guess there is some wit in a scene like that and some willingness to experiment with the horror medium. The sequel is one strange semi-horror scene after another. The plot is that some professor of some sort has translated the Necronomicon (of H. P. Lovecraft fame). He recorded an incantation on tape and now whenever anyone plays the tape it's Anything-Can-Happen-Day. A young couple find the cabin and think it might be an ideal trysting place. Most of one of them is left the next night when the professor's daughter shows up with a friend and two rather strange locals. By that point we have already seen a beheaded corpse climb out of the ground and do a charming dance with its head. We've seen a lot more than that, but that would be telling. And we will see a whole lot more, but that, too, would be telling. The actors of this piece were, I think, chosen for the terrorized looks they could get on their faces and for how ghoulish they could make themselves look. The script is incredibly contrived, including such touches as having a bridge that would have cost in the millions that leads to nowhere but a shack in the woods. I didn't think boondoggles got that big. For those who like gore and creative off-beat horror, this one's for you. As a fan of the latter,

though not of the former particularly, I will give this a +1 on the -4 to +4 scale. If you like the bizarre, give it a try.

Fig. 5.1.1. Sample Movie Review Data

2. Results:

Results are observed on the above movie summary. Steps provided in the each approach will give following results. I have used reduction size as $(1/5)^{th}$ of the original size. My approach considers a sentence (even it have more number of words than other) as a unit of size.

5.2.1 Naïve Based Review Summary Approach:

Well, what can I say? It has very little acting, no stars, little continuity, and no logic. But it has action, horror, and black humor in massive doses. Now, THE EVIL DEAD II did have some plot. And we will see a whole lot more, but that, too, would be telling. The actors of this piece were, I think, chosen for the terrorized looks they could get on their faces and for how ghoulish they could make themselves look. As a fan of the latter, though not of the former particularly, I will give this a +1 on the -4 to +4 scale.

Fig. 5.2.1. Naïve Based Review Summary

5.2.2 Title Based Review Summary Approach:

I guess in some sense THE EVIL DEAD II is the ultimate horror film...sort of...I guess. It has very little acting, no stars, little continuity, and no logic. Now, THE EVIL DEAD II did have some plot. What THE EVIL DEAD boasted most was wit. An attacking corpse would be thrown into the fire. Then some living person would have a bout of remorse and pull it out of the fire. The corpse would look up and politely thank its benefactor for pulling it out of the fire, then continues to try and kill the living.

Fig. 5.2.2. Title Based Review Summary

5.2.3 Text Rank Based Review Summary Approach:

I guess there is some wit in a scene like that and some willingness to experiment with the horror medium. Then some living person would have a bout of remorse and pull it out of the fire. The corpse would look up and politely thank its benefactor for pulling it out of the fire, then continues to try and kill the living. Creative visual concepts abound and the pace is frenetic and that makes up for a multitude of sins. By that point we have already seen a beheaded corpse climb out of the ground and do a charming dance with its head. The script is incredibly contrived, including such touches as having a bridge that would have cost in the millions that leads to nowhere but a shack in the woods. The Internet Movie Database accepts no responsibility for the contents of the review and has no editorial control.

Fig. 5.2.3. Text Rank Based Review Summary

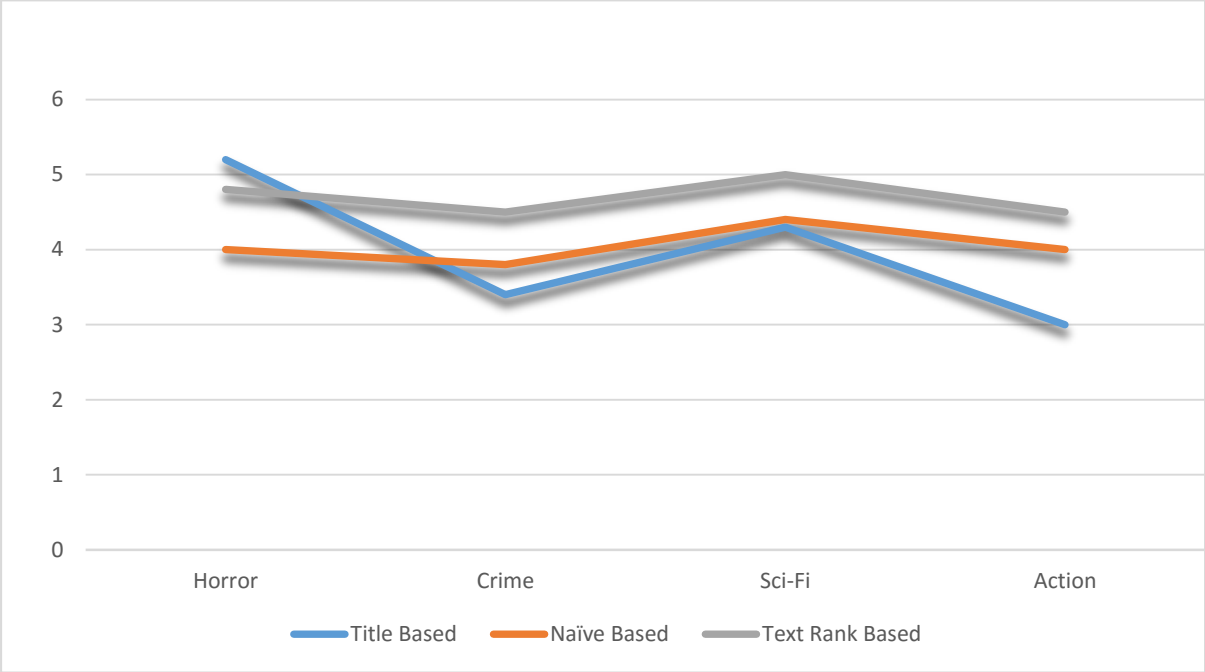

Fig. 5.2.4. Performance Comparison

Approach Name	Execution Time (Seconds)
Naïve Based Approach	0.03
Title Based Approach	5.02
Text Rank Based Approach	0.07

Table. 5.2.1. Execution Time

6. Comments:

I have experimented three algorithms for movie review summarization. Each one would be effective in different types of scenarios. Naïve Based approach as well as Text Rank Based approach would be efficient in terms of time. They also provide relevant summary from the text review. Moreover, Text Rank Based approach gives more promising results when observed for more different types of review data. Sometimes, user might be interested in reading the part which is really related to the title of a movie from its reviews. In such scenarios, I recommend using Title Based approach which provides result summary based on title of the movie. Even though title based has more latency in terms of execution time, result summary, extracted from the whole movie review, is very much relative to the movie title topic.

7. References:

1. Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing text documents: sentence selection and evaluation metrics", ACM SIGIR, 1999, pp 121–128.

2. Kirill Kireyev," Using Latent Semantic Analysis for Extractive Summarization", In Proceedings of Text Analysis Conference, 2008.

3. Karel Ježek, Josef Steinberger "Automatic Text Summarization (The state of the art 2007 and new challenges)".

4. P.J. Herings, G. van der Laan, and D. Talman. 2001. Measuring the power of nodes in digraphs. Technical report, Tinbergen Institute.

5. R. Mihalcea and P. Tarau. 2004. TextRank – Bringing Order into Texts.

6. R. Mihalcea, P. Tarau, and E. Figa. 2004. PageRank on semantic networks, with application to word sense disambiguation. In Proceedings of the 20st International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, August.

7. G. Salton, A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. Information Processing and Management, 2(32).

8. R. Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the 42nd Annual Meeting of the

Association for Computational Linguistics (ACL 2004) (companion volume), Barcelona, Spain.

9. P. Turney. 1999. Learning to extract key phrases from text. Technical report, National Research Council, Institute for Information Technology.

10. C.Y. Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May.

11. Vishal Gupta, Gurpreet Singh Lehal "A Survey of Text Summarization Extractive Techniques'', Journal Of Emerging Technologies In Web Intelligence, Vol. 2, Num. 3, August 2010.

12. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of ACM-KDD 2004, pp.168-177.