Generic Multi Document Summarization Based
On STARLET Approach

Mr. Prashant Jawalkar* Prof. Kirti Korabu*

Information Technology Department
STES, Sinhgad College of Engineering, Vadgaon (Bk.), Pune, India
1prashant.jawalkar@gmail.com
2kirti_korabu@yahoo.com

Abstract - Document summarization is an emerging technique for understanding the main purpose of any kind of documents. To visualize a large text document within a short duration and small visible area like PDA screen, summarization provides a greater flexibility and convenience. Summarization may be single or multi document summarization. If summary is to be generated for single document then it is called as single document summarization. If summary is to be created for multiple documents then it is called as multi document summarization. STARLET is a novel approach to multi-document summarization for evaluative text that considers the rating distribution as summarization feature to consistently preserve overall opinion distribution expressed in the original review. In STARLET approach A* Search Algorithm is used. STARLET is a multi document summarization of Service and Product Reviews with balanced rating distribution. A Generic Multi Document Summarization is a graph based multi document summarization technique. In this technique, directed acyclic directed graph will be constructed and A * search algorithm is used for selecting high scoring sentences and summary is generated. Finally, we shows that how Generic Multi Document summarization is efficient than other available multi document summarization techniques such as RANDOM and MEAD.

Keywords: Summarization, Multi Document Summarization, GMDS.

I. INTRODUCTION

A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or user) and task (or tasks) [7]. Roughly summarization is the process of reducing a large volume of information to a summary or abstract preserving only the most essential items. Summarization is the restating of the main ideas of the text in as few words as possible. It can be done in writing, orally, through drama, through art and music, in groups and individually.

An extractive summarization [9] method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

If summarization is performed for a single text document then it is called as the single document text summarization. Single document summarization techniques have the potential to simplify information consumption on mobile phones by presenting only the most relevant information contained in the document. If the summary is to be created for multiple text documents then it is called as the multi document text summarization technique [9]. Our main focus is on extractive multi-document summarization.

The rest of the paper is organized as follows: Section II consist of Literature survey. It describes existing multi document summarization techniques and STARLET approach for service and product reviews using balanced rating distributions. Section III describes the GMDS summarization model. Section IV describes conclusion and section V consist of references.

II. LITURATURE SURVEY

A. Existing Multi Document Summarization Techniques

There are many multi document summarization techniques available. Following are the some of multi document summarization techniques we studied for Generic Multi Document Summarization as follows.

1. RANDOM based technique

The RANDOM based technique [7] is the simplest of all the other as it randomly selects lines from the source document, depending upon the compression percentage and put them inside the summary. In this technique, a random value between 0 and 1 is assigned to each sentence of the document. A threshold value for length of the sentence is
provided. We will assign a score of 0 to all sentences that do not meet this length cutoff. Finally we choose required sentences according to assigned highest score for desired summary.

2. **LEAD based technique**

LEAD based technique [7] is a technique in which first or first and last sentence of the paragraph are chosen depending upon the compression rate (CR) and it is very good for news articles as they have the main theme set in the first lines of the articles. So, it can be reasonable that n% sentences are chosen from beginning of the text e.g. selecting the first sentence of each document, then the second sentence of each, etc. until the desired summary is constructed. This method is called LEAD based method for summarization. In this technique we assign a score of 1/n to each sentence, where n is the sentence number in the corresponding document file. This means that the first sentence in each document will have the same scores; the second sentence in each document will have the same scores, etc. We also provide a threshold value for sentence's length. The sentences with lengths less than the specified value are thrown out.

3. **MEAD based technique**

MEAD [7] is a centroid-based extractive summarizer that scores sentences based on sentence-level and inter-sentence features which indicates the quality of the sentence as a summary sentence. It then chooses the top-ranked sentences for inclusion in the output summary. MEAD extractive summaries score sentences according to certain sentence features - Centroid, Position, and Length. In this technique the score of a sentence is calculated using the following formula as follows [7].

\[
\text{Score} (S_i) = \begin{cases} 
\sum (W_c \cdot C_i + W_p \cdot P_i) & \text{If Length} (S_i) > \text{Threshold} \\
0 & \text{If Length} (S_i) < \text{Threshold}
\end{cases}
\]

Here,

- \(W_c\) = The weight for the Centroid feature.
- \(W_p\) = The weight for the Position feature.
- \(C_i\) = The calculated Centroid value for ith sentence.
- \(P_i\) = The calculated Position value for ith sentence.
- \(S_i\) = The ith sentence of the document.
- \(i\) = Sentence number within the cluster
- \(n\) = Number of sentences in a single or multiple text documents.

The highest value of the scored sentence is taken in the extract file. Thus the MEAD based summary is generated. The default weights for Centroid and Position are both 1. The default Length cutoff is 9.

**B. Stemmers and Stop Words**

Generic multi document summarization base on STARLET approach requires study of stemmers and stop words.

1. **Stemming**

In information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form, generally a written word form. In documents a word can be seen in different formats, such as plural vs. singular, present vs. past tense, etc. Most of the time these words have the same meaning and treating them differently is unnecessary.

The efficiency of a stemmer is important while performing further calculations. Sometimes stemmers can do over-stemming such that two words are given the same stem, while it should not be. For example, the words "experience" and "experiment" are two different words, which should not be stemmed into the same root. But stemmers can find out their root as "experi". Another stemming problem is related to under-stemming such that two words should have been stemmed into the same word, but have not been. For example, "run" and "ran" can be found as two different stems, instead of one.

2. **Stemming Examples**

A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

3. **Stop word Filtering**

Stop words are words which are filtered out prior to, or after, processing of natural language data (text). It is controlled by human input and not automated. There is not one definite list of stop words which all tools use, if even used. Stop words list is predefined list of words. Input documents usually contain words that do not add information but are necessary for syntactical formation, such as words like "the", "is", etc. Since these words are less useful and less informative, they introduce noise into the document representation. Stop word removal is done using predefined, human-made list of words.

4. **Stop Word Examples**

Following are the examples of stop words as follows. about, above, across, after, again, against, all, am, are, the, was, were, back, backed, can, do, does, done, down, is etc.
C. STARLET approach

Reviews about products and services are abundantly available online. However, selecting information relevant to a potential buyer involves a significant amount of time reading user's reviews and weeding out comments unrelated to the important aspects of the reviewed entity. In this work, we present Starlet, a novel approach to multi-document summarization for evaluative text that considers the rating distribution as summarization feature to consistently preserve the overall opinion distribution expressed in the original reviews [1].

1. Introduction

With the broad availability of always-connected portable devices such as mobiles, tablets, and eReaders, condensing information for displaying in a relatively small screen has become a necessity for the exceedingly demanding population of user's on-the-go. Retail industry and service providers are recognizing that there is a growing crowd of potential customers who are relying on their devices to learn about products and services, discover other user's experiences, and, ultimately, make a decision about spending their money or not.

In the service domain case, many web sites allow reviewers to directly rate pre-defined aspects. E.g., for restaurants typical aspects are atmosphere, food, value, service, and overall with ratings ranging from poor (one star) to excellent (five stars). These rated aspects quantify opinions and polarities expressed in each review by the reviewers, and although there might be inconsistencies, it is safe to assume that the text document associated with the ratings carries the same sentiment contributions quantified by the number of stars. By the same token, aggregating the ratings of the single reviews over the aspects can yield a fair summary of the overall sentiments expressed by the reviewers of the specific service or entity reviewed. Based on these considerations, we can assume that a summary should convey the same distribution of ratings over aspects obtained by combining the rating contribution of each review, so that each opinion contribution, even if controversial, should be represented into the final summary.

Starlet, a novel approach to summarization of evaluative text that utilizes aspects and ratings described in the reviews as features for the summarization process. In the restaurant domain, which we investigate as an example domain for service reviews; Starlet uses atmosphere, food, value, service, and overall aspects to score each sentence in the input documents. For each aspect, Starlet computes a rating indicating how much the current sentence has contributed to that aspect. For this Starlet uses a maximum entropy rating model. The predicted aspect ratings are used in a summarization model to compute a score for each sentence and to derive a summary score. The model is a linear weighted model with aspects as features and associated weights learned using A* search and discriminative training.

2. Feature Weight Learning

An extractive multi-document summary can be created by traversing a directed acyclic graph where each node \( i \) represents a particular summary of length composed by a set of sentences \( S_i \), and a set of edges \( Traversing the edge j \) incrementally adds a sentence from the set of available sentences to the previous sentence set \( S_i \). Figure 1 shows a graphical representation of the process of selecting sentences. Each node in the graph can be evaluated by a scoring function which quantifies how good the node is when compared to a target node [1].

![Fig. 1 Creation of extractive summary.](image)

The A* search algorithm can be used to efficiently traverse the graph and accurately find the optimal path. It applies a best-first strategy to traverse the graph from the initial node (summary of length zero) to the final node (summary of length \( W \)), and uses a heuristic function to determine the order of the nodes to explore first. The search algorithm is guaranteed to converge to the optimal solution if the heuristic function is monotonic or follows the admissible heuristic requirements. That is, the estimating path cost function from the current node to the goal never overestimates the actual cost. Starlet used the "final aggregated heuristic" function described in that satisfies the admissible heuristic constraints. The input to the heuristic is the set of sentences sorted according to their scores.

The heuristic first adds the highest scoring sentence into the summary. After adding a sentence, the summary length is updated. If the length limit of the summary is not violated then the next highest scoring sentence is added. When the next high scoring sentence is too long to be added to the summary the heuristic skips this and continues with the next one until it finds the best scoring sentence that does fit into the summary.
3. Feature Extraction

In order to apply the search techniques described above, it is necessary to define a set of features relevant to the summarization task that can be determined at each step of the search process.

Fig. 2 Rating Predictive Models (RESTAURANT).

Fig. 2 shows the architecture of our rating prediction model system. In this configuration, each review document is associated with a set of predefined aspects that have been assessed by the reviewers with star-rating evaluations. During training, text features such as n-grams, parts of speech, shallow parsing chunks and others are used together with the reviewer-assigned ratings to create a discriminative model classifier.

4. Experimental Evaluation

To evaluate our STARLET approach, Authors compared it with two summarizers:

i. a baseline summarization system that randomly selects sentences with no repetition till it reach the desired length of 100 words;

ii. The open source MEAD system with the same output length. The resulting summaries were assessed using the automatic metric ROUGE and manual evaluation.

STARLET is a novel approach to multi-document summarization for evaluative text that considers the rating distribution as summarization feature to consistently preserve overall opinion distribution expressed in the original review. In STARLET approach A* Search Algorithm is used. STARLET is a multi document summarization of Service and Product Reviews with balanced rating distribution. STARLET summaries are more efficient than RANDOM based and MEAD based techniques. Finally, the 'coverage' score for STARLET is decidedly better than for the other approaches, showing that STARLET correctly selects information relevant to the users.

III. GMDS SUMMARIZATION MODEL

Generic Multi Document Summarization based on STARLET approach is graph based multi document summarization algorithm.

Algorithm consists of following steps as shown in Fig. 3. The input to the model is a set of related documents. Firstly, the set of documents is pre-processed. The directed acyclic graph is constructed for each document with sentences as nodes and similarities as edges. Thereafter, weighted ranking algorithm MEAD * [7] is performed on the graph to generate salient score for each sentence in the document. The sentences are ranked according to their salient scores. The top-ranking sentences are selected to form the summary for each document using A * search algorithm and Maximal Marginal Relevance also is used to filter out redundant information. Secondly, all the single summary of each document are assembled into one document. Finally, the described above process is applied to this combining document to form the final extractive summary.

Fig. 3 Main Process Generic Multi Document Summarization.

A. Preprocessing

Before constructing graph, the input set of related documents needs to be preprocessed. In the first step, input documents are parsed to extract all sentences. Those
sentences, which are too short or almost contain no information, are eliminated.

B. Graph Construction

The directed acyclic graph \( G = (V \times E) \) represent each document is constructed as follow. Each sentence appearing in the document becomes a node in the graph representing that document. The edges of the graph represent similarity between the sentences.

C. Sentence Ranker

Once document graph is built, the sentences in a document will be ranked through random walk on \( G \). We compute a salient score for each node using the MEAD *(star) algorithm.

D. High Scoring Sentence Selection

In this step, high scoring sentences are selected using A * search algorithm by traversing graph for generation of summery.

E. Summary Generation

In this step, final summary is generated using sentences selected by using A * search algorithm [5]. Simply, sentences with high ranking scores may be chosen as the final ones in the summary. However, there may be much redundancy among the top ranking sentences, since similar sentences tend to get similar ranking scores during the ranking process. The modified version of Maximal Marginal Relevance (MMR) is applied to re-rank and select sentences to add into summary. A sentence is added if it is high ranked and not too similar to any sentence existing in the summary.

IV. CONCLUSION

STARLET is a novel approach to multi-document summarization for evaluative text that considers the rating distribution as summarization feature to consistently preserve overall opinion distribution expressed in the original review. In STARLET approach A* Search Algorithm is used. STARLET is a multi document summarization of Service and Product Reviews with balanced rating distribution. In STARLET multi document summarization technique, authors have compared with other multi document summarization techniques such as RANDOM and MEAD and showed that STARLET is more efficient for product and service reviews than these techniques.

STARLET is an opinion summarization. (i.e. review based). A Generic Multi Document Summarization is a graph based multi document summarization technique. In this technique, directed acyclic directed graph will be constructed and A * search algorithm is used for selecting high scoring sentences and summary is generated. Finally, we will be using Generic Multi Document summarization to show that how it efficient than other available multi document summarization techniques such as RANDOM and MEAD.

V. REFERENCES


