A SURVEY ON SENTENCE SIMILARITY BASED AUTOMATIC TEXT SUMMARIZATION TECHNIQUES

[1]Renjith S R
[1]Dept. Of Computer Science, College Of Engineering, Cherthala, Alappuzha, Kerala, India
Email:[1]renjisaras@yahoo.com

Abstract—Text Summarization is the process of generating a short summary for the document, that contains the significant portion of information. In an automatic text summarization process, a text is given to the computer and the computer returns a shorter less redundant extract of the original text. It has been an area of interest since many years. The similarity of a sentence to other sentences in the document plays a key role in ,Summarization methods have been developed based on this aspect. This paper performs a survey on various automatic text summarization techniques based on sentence similarity feature. Performance of different methods in terms of precision, recall and f-measure values are compared.

Keywords—Text summarization, Sentence Extraction, Sentence similarity, Summary generation, Precision, Recall, F-measure

I. INTRODUCTION

With enormous growth of information on cyberspace, conventional Information Retrieval techniques have become inefficient for finding relevant information effectively. When we give a keyword to be searched on the internet, it returns thousands of documents overwhelming the user. It becomes a time consuming and difficult task to recall the precise documents. Meanwhile, due to the massive increase in the text information that we receive every day, text summarization systems are helpful in finding the most important content of the text in short time. Research in automatic text summarization has received considerable attention due to the exponential growth in the quantity and complexity of the information sources on the internet. Text summarization approaches are used as a solution to this problem which reduces time required to find the web document having relevant and useful data. Text summarization is the process of automatically creating a compressed version of the text containing significant information.

The summaries can help the reader to get a quick overview of an entire document. Another important issue related to the information retrieval from the internet is the existence of many documents with the same or similar topics, known as duplication. This kind of data duplication problem increases the necessity for effective document summarization. The advantages of automatic text summarization are saving in reading time, facilitating document selection and literature searches, improvement of document indexing efficiency, free from bias, and they are useful in question-answering systems where they provide personalized information.

Input to a summarization process can be a single document or multiple text documents of related topic. When only onedocument is the input, it is called single document text summarization and when the input is group of documents, it is called multi document summarization. We can also categorize the text summarization based on the type of users the summary is intended for: i) User focused summaries are intended to satisfy the requirements of a particular user or group of users and ii) generic summaries are aimed at a broad community. Depending on the nature of summary, it can be categorized as an abstract or an extract. An abstract is a summary, which represents the subject matter of an article by understanding the whole meaning, which are generated by reformulating the salient units selected from the input sentences. It may contain some text units which are not present in the original input text. An extract is a summary consisting of a number of sentences selected from the input text. Sentence extraction methods are easier to perform and have been studied extensively over the past decade.

Based on information content of the summary, it can be categorized as informative and indicative summary. The indicative summary represents an indication about an articles purpose and it prompt the user for selecting the article for in-depth reading for detailed understanding. On the other hand, informative summary covers all significant information in the document at an abstract level, that is, it will contain information about all the different aspects such as articles purpose, scope, approach, content, domain, results and conclusions. For example, an abstract of a research article is more informative than its headline.

The similarity of a sentence to other sentences in the document plays a key role in summarization procedure. Several summarization methods have been developed based on this aspect. This paper performs a survey on various automatic text summarization techniques based on sentence similarity feature. The rest of this paper is organised as follows: Section II is a survey on various sentence similarity based summarization techniques, Section III performs a comparison of different techniques in terms of precision, recall etc, and Section IV concludes the paper.

II. TEXT SUMMARIZATION TECHNIQUES BASED ON SENTENCE SIMILARITY

Several methods have already been developed for text summarization. Six automatic text summarization techniques based on sentence similarity have been specified here and their performances are compared in terms of precision, recall, f-score etc

A. Information content based Sentence Extraction

Daniel Mallett et al. [1], proposes a FULL-COVERAGE algorithm that extracts sentences that fully covers the concept-space of a document by iteratively measuring the similarity of each sentence to the whole document and striking-out words that have already been covered. It is based on the fact that the relevance of a sentence is proportional to its similarity to the whole document. The first phase of this algorithm is to parse a document into sentences. During this phase they also perform tokenization, removal of stop-words and apply the Porter stemming algorithm. The second step of the algorithm is to calculate FC, the subset of the sentences that cover the entire concept space of a document. The method for determining FC is to treat each individual sentence S_i ($i = 1, ..., N$) of D as a document within the overall collection of D itself. The next step is to use the entire document as a query against each individual sentence, adding the highest ranked sentence to the full coverage set. Once the ranked FULL COVERAGE set of sentences has been determined, the third step is to actually generate and return a summary. Given a percentage p, a CR(p) summary consists of the first n sentences from FC where $n \leq p*|FC|$.

B. Fuzzy logic based method for text summarization

Ladda Suanmali et al.[2],propose text summarization based on fuzzy logic method to extract important sentences as a summary. In the pre-processing step, Sentence Segmentation, Tokenization, Removing Stop Word, and Word Stemming are done. Then they extracted the important features for each sentence of the document consisting of the following elements:

title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data. The numerical vectors corresponding to the features are calculated to obtain the sentence score base to be used on fuzzy logic method. A set of highest score sentences are extracted as document summary based on the compression ratio. The fuzzy logic system consists of four components namely fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. Fuzzifier translates inputs into linguistic values using a membership function and the results are used as the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IF- THEN rules. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score.

The input membership function for each feature is divided into five fuzzy set which are composed of unimportant values (low (L) and very low (VL)), Median (M) and important values (high (H) and very high (VH)).Then example of a rule is, IF (No Word In Title is VH) and (Sentence Length is H) and (Term Freq is VH) and (Sentence Position is H) and (Sentence Similarity is VH) and (No ProperNoun is H) and (No Thematic-Word is VH) and (Numberical Data is H) THEN(Sentence is important).

C. Query-based summarizer based on similarity of sentences

A. P. Sivakumar et al.[3], proposes a query based document summarizer based on similarity of sentences and word frequency. The summarizer uses Vector Space Model for finding similar sentences to the query and Sum Focus to find word frequency. In this paper they propose a query based summarizer which is based on grouping similar sentences. In the proposed system first the query is processed and the summarizer collects required documents and finally produces summary. After Pre-processing, producing the summary involves the following steps: i)Calculating similarity of sentences present in documents with user query, ii) After calculating similarity, group sentences based on their similarity values, iii) Calculating sentence score using word frequency and sentence location feature, iv) Picking the best scored sentences from each group and putting it in summary, v) Reducing summary length to exact 100 words.

D. Language Independent Sentence Extraction Based Text Summarization

Krish Perumal et al. [4], proposed a language independent sentence extraction based text summarization technique which uses a structural characteristics based sentence scoring along with a PageRank based sentence ranking. The effectiveness of the proposed approach had been confirmed for English and Tamil documents by applying the ROUGE evaluation. The method was carried out in four different phases namely i)Pre processing, where stop word removal and stemming are performed in order to prepare the source data for summary generation, ii)Scoring, where the sentences were given scores based on their position, length, topic similarity and TF-IDF feature such that longer sentences similar to the title of the document and appearing at the beginning of the document are getting high scores, iii)Ranking, where the sentences are ranked according to Google’s PageRank formula and finally, iv)summary generation, where the final summary comprises of the top ranked sentences displayed in the same order as they appear in the source document text. The number of top ranked sentences selected for the summary may be user-defined in terms of the number sentences or compression ratio with respect to the length of the source document text.

E. Text summarization using clustering technique

Anjali R. Deshpande et al.[5], proposes a new approach to multi-document summarization which ensures good coverage and avoids redundancy. The user selected collection of documents and query is the input to the summarizer. They have maintained a list of maps where each term from document collection is stored in a map with its number of occurrences. Query modification technique is used as follows:

Split a query into tokens and find the synonym for each token and if the token or synonym exists in a document collection then append the most frequent synonym of the query term to query. The most frequently occurred words from corpus are selected and those words are appended to the query. So the query is strengthened.

The features used to calculate sentence scores are listed as follows : Noun Feature, Cue Phrase Feature, Sentence Length Feature, Numerical data Feature, Sentence Position, Sentence centrality (similarity with other sentences), Upper Case word feature, Sentence similarity with user query, Term frequency and Inverse Document Frequency. The documents are clustered by using, cosine similarity as a similarity measure to generate the appropriate document clusters. Then from every documentcluster, sentences are clustered based on their similarity values. Calculate the score of each group (sentence cluster). Sort sentence clusters, in reverse order of group score. Pick the best scored sentences from each sentence cluster and add it to the summary.

F. Extractive multi-document summarizer algorithm

Amit S. Zore et al.[6], proposes an Extractive Multi Document Summarization algorithm which is a graph based multi document summarization consists of following steps: The input to the model is a set of related documents. Firstly, the set of documents is pre-processed. The undirected acyclic graph is constructed for each document with sentences as nodes and similarities as edges. Thereafter, weighted ranking algorithm 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. Secondly, all the single summary of each document assembled into one document. Finally, the described above process is applied to the combining document to form the final extractive summary.

III. COMPARATIVE STUDY OF SENTENCE SIMILARITY BASED SUMMARIZATION TECHNIQUES

The information content based method [1] proposes the FULL-COVERAGE algorithm using the concept that the relevance of a sentence is proportional to its similarity to the whole document. They experiment the system with data from SMARTs TIME Magazine Collection as well as the TREC documents used for 2002 edition of DUC.A summarized version of the TIME Magazine collection is only 40% the size of the original text and for DUC, the algorithm produces summaries 22% the size of the original texts. Two other baseline techniques, namely, a random and a lead-based summarizer were used for comparison with the proposed system. The precision of the system was found to be 0.347 using ROUGE-I evaluation.

The fuzzy logic based method [2] extracted 8 important features and calculated their score for each sentence. It proposes a text summarization based on fuzzy logic to improve the quality of the summary created by the general statistic method. The system was tested using DUC 2002 dataset. They compared their results with the baseline summarizer and Microsoft Word 2007 summarizers. ROUGE-I evaluation tool shows a precision of 0.49769, recall of 0.45706 and f-measure of 0.47181.

The Query based summarizer [3] present a query based document summarizer based on similarity of sentences and word frequency. They used AQUAINT-2 Information-Retrieval Text Research Collections as the test corpus and the obtained summary sentences are evaluated using ROUGE metrics. The summarizer does not use any expensive linguistic data. The Summarizer uses Vector Space Model for finding similar sentences to the query and Sum Focus to find word frequency. The accuracy achieved using the proposed method was compared to the, TAC (Text Analysis Conference) system. To evaluate the proposed system they used TAC2009 datasets proposed by NIST for update summarization task. It consists of 48
topics, each having 20 documents divided into two clusters ”A” and ”B” based on their

chronological coverage of topic. For evaluation of the system they used cluster 'A' documents of TAC2009 data. The precision, recall and f-score values were 0.29034, 0.30127 and 0.29961 respectively.

The Language Independent Sentence Extraction Based approach [4] involves the use of a structural characteristics based sentence scoring along with a PageRank based sentence ranking. The effectiveness of the proposed approach has been tested using English and Tamil documents by applying the ROUGE evaluation. The results for English were compiled using DUC 2002 data on single document summarization, whereas those for Tamil were compiled using a set of 100 human written summaries. The proposed system was compared with baseline summarizer. The recall values for English and Tamil were found to be 0.5200 and 0.4877 respectively.

The clustering based approach [5] presents a combined approach to document and sentence clustering as an extractive technique of summarization. It is the clustering based approach that groups first, the similar documents into clusters and then sentences from every document cluster are clustered into sentence clusters and best scoring sentences from sentence clusters are selected in to the final summary. The method was compared against techniques based on statistical features and those based on document clustering only. The precision, recall and f-measure values were found to be 0.57, 0.48 and 0.52 respectively.

The Extractive Multi-Document Summarizer algorithm [6] proposes an approach where extractive summary of multiple relevant documents is produced using various sentence features such as word class, sentence length and sentence similarity. To evaluate Extractive Multi Document Summarization, it has compared with two summarizers: Random and LEAD. The resulting summaries are assessed using the automatic metric ROUGE and manual evaluation. The precision, recall and f-measure values were found to be 0.54850, 0.57328 and 0.56062 respectively.

A comparison of precision, recall and f-measure values of different techniques are given in table 1, which shows that the EMDS (Extractive Multi-Document Summarizer) algorithm gives the best results in summarization. Another table (2), gives a comparison among the summarization techniques in terms of test corpus and the existing summarizers used for comparison. Almost all methods used the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit to evaluate the summaries generated by the method

TABLE I. Precision, Recall and F-measure values comparison of sentence similarity based summarization techniques

Summarizer	**Precision**	**Recall**	**F-measure**
FUZZY	0.49769	0.45706	0.47181
QUERY	0.29034	0.30127	0.29961
CLUSTERING	0.57	0.48	0.52
EMDS	0.5485	0.57328	0.56062

TABLE II. Summarizers-A comparative study

Summarizer	Test corpus	Compared with
Information content based	SMART Time magazine, TREC documents	Random, LEAD based
Fuzzy based	DUC2002 dataset	Baseline, MS Word 2007
Query based	AQUAINT-2 IR Text research collections	TAC 2009
Language independent	English-DUC 2002,Tamil-human written	Baseline Summarizer
EMDS	DUC2002 DATASET	Random, LEAD based

IV. CONCLUSION

Automatic text summarization is an area of interest since many years. Sentence similarity plays a key role in summarization. This paper focuses on a survey of summarization techniques using the sentence similarity aspect. Six different methods are compared here in terms of their test corpus, precision, recall, f-measure etc. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) tool was used by almost all methods for evaluating the summaries generated and comparing them with existing techniques. It was found that the EMDS (Extractive multi-document summarizer) algorithm gives the best results in terms of precision and recall values.

REFERENCES

[1] Daniel Mallett, James Elding, Mario A. Nascimento Information content based Sentence Extraction for Text Summarization, 2004.

[2] Ladda suanmali, Naomie salim, Mohammed salem Binwahlan Fuzzy logic based method for improving Text Summarization, International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

[3] A. P. Siva kumar, Dr. P. Premchand and Dr. A. Govardhan, Query- Based Summarizer Based on Similarity of Sentences and Word Frequency International Journal of Data Mining and Knowledge Management Process, vol.1, no.3, May 2011.

[4] Krish Perumal, Bidyut baran Chaudhuri Language Independent Sentence Extraction based Text Summarization, In Proceedings of ICON 2011, 9th International Conference on Natural Language Processing.

[5] Anjali R Deshpande, Lobo L M R J Text summarization using Clustering technique, International Journal of Engineering Trends and Technology, Volume 4, Issue 8 (August 2013).

[6] Amit S. Zore, Aarati Deshpande Extractive Multi-Document summarizer algorithm International Journal of Computer Science and Information Technologies, Vol. 5, 5245-5248, 2014.