STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720(online)
Volume 2, Issue 2
www.stmjournals.com

A Novel Sentence Ranking Clustering Based Algorithm for Text Summarization

Akanksha Sahu*, Chander Diwaker[2]
Department of Computer Science and Engineering, U.I.E.T., Kurukshetra University, Kurukshetra, India

Abstract

Text Summarization is the process of generating reduced text from source document by extraction or generation. A novel approach will be presented for extractive text summarization. The proposed technique performs filtering and then ranking on sentences based on NLP text mining techniques. This method uses the synonyms extracted from online thesaurus dictionary to identify similar sentences and then clusters these sentences to summarize. The sentences are ranked based on sentence ranking algorithm and then the sentences from the filtered vectors which have a rank above a certain threshold value are included. A new method will be introduced where the user is allowed to manage the summary size as per his requirements. Unlike previous techniques, this technique takes into account synonyms of the keywords found in sentences and considers them while clustering. In the end, the created summaries by using this technique will be compared with others produced by commercial and research applications to demonstrate the efficiency of our technique.

***Keywords*:** *Summarization techniques, sentence ranking, stemming, synonyms, text summary*

**Author for Correspondence* E-mail: sahuakanksha8@gmail.com

INTRODUCTION

Natural language processing is the science of converting human understandable language into machine understandable language and vice versa. World Wide Web has seen an enormous increase in the last few years and so the problem of information overload also has increased. A huge amount of information on internet is in the form of text. Hence, there is a need of a system that automates the process of retrieving, categorizing and summarizing the document as per users need.Clustering helps to find the different topics that are present in a document, and helps into feat the relevance of the different topics [1]. Clustering helps the summarization technique by categorizing the sentences according to its need.

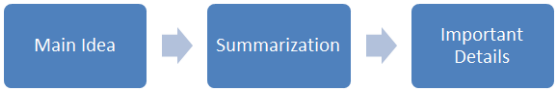

***Fig. 1:** Basic Concept of Summarization*

Summarization means to create excerpts of text documents, technical papers and magazine articles, with no informationloss, which serves the purpose of conventional abstract fully, by automatic means.

CLASSIFICATION OF SUMMARIZATION TECHNIQUES

Different types of summarization techniques can fit into different types of scenarios. So, the knowledge of various dimensions which can be a basis of summarization mechanisms. The classifications are summarized as under:

Based on Approaches

There are two strategies for summarization based on approach. Those are summarization by extraction, which consists of extracting source sentences as it is and putting them into a summary and summarization by abstraction, which creates new sentences which carry the same meaning for the summary [2]. *Summarization by extraction just extracts the selected sentences from the source document and puts them* into the summary. Extractive method is comparatively easier to implement and is based on statistical features not on semantic relation with sentences. So, the summary generated by extractive method tends to be inconsistent. Summarization by

abstraction needs to understand the original text and then generate the summary which is semantically related. It provides more generalized summary but it is not easy to formulate.

Based on Languages

On basis of language, the summarization can be classified as mono-lingual or multi-lingual. Mono lingual systems work with documents normally written in a specific language and the summarized output has to be in the same language only. On the other hand, Multi-lingual systems can work with documents written in different languages and summarize too in different languages [3].

According to Types of Details

The two ways to classify on basis of types of details are informative and indicative. An indicative summary provides only the main idea of the original text and gives a quick view of otherwise lengthy document. These are normally small and the user is encouraged to read the original document. Informative summary is almost a substitution to the original document [3]. It provides the concise information from the original document and gives user almost everything contained in the document.

Based on Limitation

There are two types of summary classifications based on limitation of input text. Genre specific systems are only meant for special inputs like newspaper articles, stories, manuals etc. [3]. So, they are limited to the type of input they can accept. Domain independent system can work with different types of text. They are independent of the domain and any type of user can use them. There are few systems that are domain dependent.

Based on Number of Input Documents

This classification is done on the basis of number of documents inputted for summarization. Single document summarization can accept only one document as input. They are usually easier to produce as it involves summarization of a single document. Multi-document summarization accepts several documents of same topic as an input. It is more difficult to implement as there are multiple documents to summarize.

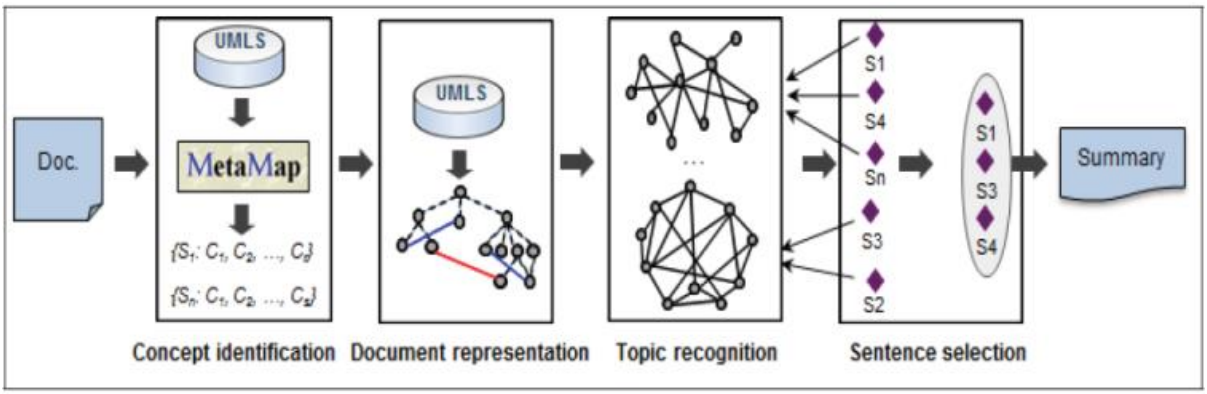

Fig. 2: *Architecture of Graph Based Summarization System [1].*

REVIEW OF LITERATURE

Alguliev *et al.* stated that text summarization was the process of automatically creating a compressed version of a given document preserving its information content [4]. There were two types of summarization: extractive and abstractive. Extractive summarization methods simplified the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. Abstractive summarization composed novel sentences, unseen in the original sources. In this study the author focussed on sentence based extractive document summarization. The extractive summarization systems were typically based on techniques for sentence extraction and aimed to cover the set of sentences that were most important for the overall understanding of a given document. In this paper, the author proposed unsupervised document summarization method that created the

summary by clustering and extracting sentences from the original document. For this purpose new criterion functions for sentence clustering had been proposed. Similarity measures played an increasingly important role in document clustering. The author developed a discrete differential evolution algorithm to optimize the criterion functions. The experimental results showed that the suggested approach could improve the performance compared to state-of-the-art summarization approaches. Nenkova *et al.* stated that the topic representation of approached work will be started by creating an intermediate representation of the text where the topics discussed in the input were captured [5]. The sentences in the input document were rated for relevance according to the ways for representing the topics. The text was shown as a diverse set of possible indicators of relevance which did not aim at discovering topicality in indicator representation approaches. These indicators were used together, very often using machine learning techniques, to rate the relevance of each sentence. Finally, a summary was created by selecting sentences based on a greedy approach, selecting sentences one at a time, that were included in the summary, or optimizing the selection globally, choosing the set of sentences, which are best rated, in the summary. In this research, the author had given a detailed overview of existing approaches based on these distinctions, with particular emphasis on how sentence scoring, representation or strategies for summary selection altered the overall performance of the summarizer. A.N. Pimpalshende presented an analysis of extractive text summarization techniques [6]. The author detailed different types of summarization techniques, their features, evaluation strategies along with metrics and problems faced in these text summarization methods. The research concentrated mainly on extractive summarization methods. The summary resulted from these methods contained important sentences selected from the original document. The sentences were ranked based on their statistical and linguistic features. Tian Xia worked on vector space model for text classification [7]. According to author, extracting the semantics from natural language text was not easy. Before applying any NLP techniques, many problems had to be considered for a large collection of texts. The Vector Space Model conceptually stored the document as a vector of terms chosen from the original document. The weights attached to these terms represented the importance of each term in the document(s). The unclassified documents were also similarly modeled. In the proposed method, the document title, although it contained much semantic information, was not taken into special consideration, this paper proposed Title Vector method to summarize multiple documents. Li *et al.* proposed a graph-ranking-based method [2]. It created sentence graphs and then performed constrained reinforcements on them with the previous and current documents to analysis the importance of the sentences. The constraints confirmed that the updates to previous documents formed the most prominent sentences in current documents. This method was an NP-hard, approximate method and had polynomial time complexity, was proposed by author. The effectiveness and efficiency of the proposed method was experimentally confirmed on the benchmark data like TAC 2008 – 09 datasets. The author concluded that the first sentence in a document was extremely important and its up-weighting could hugely improve the performance of the proposed method. Yulia *et al.* suggested a new method based on graph-based ranking algorithms for extractive text summarization [8]. This paper worked by ranking Maximal Frequent Sequences (MFS) in order to get the most relevant information in a text. These sentences were set as nodes of a graph in term selection step, and then, in term weighting step, they were ranked using a graph based algorithm. It was experimentally confirmed that the proposed method delivered better results than other contemporary methods. In addition, it found the best sentences for summarisation which represented the context better than other terms. It was concluded that the longer MFS, yielded the better results. The exclusion of stop-words, basically destroyed the sense of MFS, and the results were worse. The proposed method was highly portable to other contexts in a way that it did not need deep linguistic knowledge or knowledge about particular domain. Gross *et al.* proposed a

novel, unsupervised technique to summarize (multi-)documents [9–12]. The approach summarized documents in an unsupervised and language-independent fashion and evaluated the relevance of word associations from the source document(s). The sentences which covered the most relevant word associations from source document(s) were contained in summaries. The proposed technique was tested on the DUC 2007 dataset. It was experimentally confirmed that the proposed methodology had the best results as compared to other unsupervised summarization methods which were not using human created knowledge bases. This approach had a specialization that it explored relevant associations rather than words. The approach was generalized to multiple documents. It was mainly language-independent and used unsupervised and simple resources. In 2014, Menendez *et al.* aimed to improve a graph-based summarization process by creating a hybrid of genetic clustering graph connectivity technique [1]. So while the different topics which were dealt with in a document were identified by genetic clustering, the relevance of the different topics was assessed by connectivity information (in particular, degree centrality). It was shown by the new algorithm, called Genetic Text Clustering (GTC) that the hybrid of continuity-based measures and degree centrality produced better results as compared to original techniques. The automatic summaries were compared with others produced by commercial and research applications. This demonstrated the effectiveness of using this combination of techniques for automatic summarization by experimenting on 150 bio medical documents. Munot *et al.* 2014 made a comparative study of various text summarization methods based on their types of application [3]. The paper detailed extractive and abstractive summarization methods. The paper described taxonomy of summarization systems. It also explained statistical and linguistic approaches for summarization.

PROBLEM FORMULATED

Any text document, which is read, technical paper or magazine article, is not merely made up of sets of unrelated sentences. These sentences are connected to each other in their context. They contain rich information stored in them. Human brain can easily read it and easily conclude regarding the topic details of respective document and the context behind it. It is quite easy to understand the information content. But for machine, to do the same job, with similar expertise, is challenging. The main aspect of summarization is to study the relation between keywords and find the keywords which co-occur together called as bigram. The more cohesive the terms are with each other, the more is the probability of information content in between them. The proposed model is mainly based on the research by Hector Menendez *et al.* [1]. The graph based process is used to summarize the document and used genetic clustering to cluster the sentences. The graph based approach is always complicated and has higher time complexity. The fitness function used in clustering does consider features like density, synonyms based clustering etc.

Table 1: *Comparison of Results of Other State of the Art Methods [13].*

Method	F-Measure	Significance
Baseline Random	38.817	0%
Text Rank	44.320	26.47%
Graph anking(Without Pre-Processing)	48.065	44.47%

PROPOSED METHODOLOGY

Abstractive summarization may contain novel sentences, which carry similar meaning to the original document but which were not a part of the original document. However, abstractive approaches are difficult to achieve and there is a need of deep analysis of semantic representation, understanding, inference and natural language generation, which is yet to find the best outcomes from contemporary techniques. Extractive summarization systems are the most prominent techniques in automatic summarization to summarise documents. Here present an extractive system which makes the result close to abstractive techniques by evaluating sentences based on synonyms of keywords in those sentences. The summary created also alters the sentences to include synonyms. The technique is based on sentence ranking algorithm to form a summary of the document under consideration. The steps involved in methodology are listed as under:

- The first step is preprocessing where the title will be separated, abstract and the body part of the document.
- Then, break the entire document into a Vector of Sentences where the position in vector is an id for each sentence.
- Next create a map which will contain each sentence id and its score which is initialized to 0.
- Now, iterate the vector from step 1 and perform these steps.
- Tokenize each sentence into a vector which contains words from the sentence after filtering the support words (is, am, are, was, were, will, shall, can …) and then stemming of each of the words.
- If the document has a title, then match the tokens with keywords from title. The successful match adds to the score of sentence. The match is done for the words and also for their synonyms fetched from an online dictionary dynamically.
- Find the most frequent keywords from the document and identify the sentences containing those keywords. Increase the sentence score for each match after stemming of words.
- Next cluster the sentences using K-Means clustering algorithm where sentences are clustered on basis of their similarity and cluster importance is measured by number of sentences in the cluster.
- Increase the score of each sentence in the vector by the cluster value.
- Input the percentage summary from the user.
- Now extract the top scoring sentences from the vector till desired percentage.
- This is the resultant summary.

CONCLUSION

Summarization is a very commonly practiced task which is implemented very commonly. No matter what industry require to work for, thus often need to go through lengthy documents and it is always convenient to go through summary than the document itself. To experiment this concept the use of text documents from news domain will be there. The number of documents of different categories will be selected and the algorithm is used to evaluate the results. The extractive method proposed in this research uses sentence ranking algorithm and will sure give much optimized results. The summary should contain only the sentences more focused on the matter and ranked highly above a certain threshold value set by the user. The threshold value can also be maintained flexible so that the optimum summary size will be generated. The final results will be compare with others produced by commercial and research applications, to demonstrate the appropriateness of using this combination of techniques for automatic summarization. The work can be extended by using the same technique for doc. and docx. Files and including images and tables in the summary generated.

REFERENCES

1. Menendez HD, Plaza L, Camacho D. Combining graph connectivity and genetic clustering to improve biomedical summarization. *In Proceedings of the IEEE Congress on Evolutionary Computation*. 2014; 2740–2747p.
2. Li X, Du L, Shen YD. Update summarization via graph-based sentence ranking. *In Proceedings of IEEE Transactions on Knowledge and Data Engineering*. 2013; 1162–1174p.
3. Munot N, Govilkar SS. Comparative study of text summarization methods. *IJCA* (0975 – 8887). 2014; 102(12): 33–35p.
4. Alguliev R, Aliguliyev R. Evolutionary algorithm for extractive text summarization. *IIM*. 2009; 1(2): 128–138p.
5. Nenkova A, McKeown K. Automatic summarization foundation & trends in info. *Retrieval*. 2011; 5(2): 103–133p. DOI: 10.1561/ 1500000015.
6. Pimpalshende AN. Overview of text summarization extractive techniques. *IJECS*. 2013; 2(4): 1205–1210p.
7. Xia T. Improved VSM text classification by title vector based document representation method. *In Proceedings of 6th International Conference on Computer Science & Education (ICCSE)*. 2011; 210–213p.

8. Ledeneva Y, García-Hernández RA, Gelbukh A. Graph ranking on maximal frequent sequences for single extractive text summarization. *Proceedings of the 15th International Conference of Intelligent Text Processing and Computational Linguistics*; CICLing. 2014; 8404: 466–480p.
9. Gross O, Doucet A, Toivonen H. Document summarization based on word associations. *SIGIR.* 2014; 1: 1023–1026p.
10. Lloret E, Palomar M. Challenging issues of automatic summarization: relevance detection and quality-based evaluation. *Informatica (Slovenia).* 2010; 34(1): 29–35p.
11. Padma Priya G, Duraiswamy K. An approach for concept-based automatic multi-document summarization using machine learning. *IJAIS.* 2012; 3(3): 49–55p.
12. Garg, Sunil Chhillar. Review of text reduction algorithms and text reduction using sentence vectorization. *IJCA.* 2014; 107(12): 39–44p.
13. Kamble P, Dharmadhikari SC. Context score based term weighting model for text summarization. *IJCA.* 2014; 98(12): 41–46p.

Cite this Article
Akanksha Sahu, Chander Diwaker. A Novel Sentence Ranking Clustering Based Algorithm for Text Summarization. *Journal of Artificial Intelligence Research & Advances*. 2015; 2(2): 1–6p.