International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 09 , September 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

TEXT SUMMARIZATION USING ENHANCED MMR TECHNIQUE

Akshit Shah[1],Ashish Naik[2], Vaibahvi Dharashivkar[3]

[1,2,3]*Information Technology,St.Francis Institute of technology Mumbai University(India)*

ABSTRACT

Automatic text summarization aims to address the information overload problem. It is a way to give abstract form of large document so that the moral of the document can be communicatedeasily. In this paper we propose a method of personalized text summarization which improves the conventional automatic text summarization methods by taking into account the differences in readers' characteristics. We use annotations added by readers as one of The sources of personalization. We have experimentally evaluated the proposed method in the domain of learning, obtaining better summaries capable of extracting important concepts explained in the document when considering the relevant domain terms in the process of summarization. This will reduce the computational cost, storage and time.

Keywords: Automatic Text Summarization, Personalization, Annotations;

I. INTRODUCTION

To find protuberant points for summarization in a collection of documents. We here recommend a system to detect points for summarization from a huge or diversiform paragraphs. We use a virtuous method to discover important points from the provided content. The content is split into two parts namely Summarized Content and Summarized Point. One would predict peculiar words to appear in the content more or less frequently: "screen" and "battery" will appear more repeatedly in documents about a laptop, "tires" and "headlight" will appear in documents about cars, and "the" and "is" will appear equally in both. A credential typically concerns various topics in different proportions; thus, in a document that is 10% of car and 90% of the laptop, there would seemingly be about 9 times more laptop words than car words. Our intended system captures this intuition in a framework made up from mathematics and will research the content of a peculiar set of documents.Keywords will be extracted by the system and the topics will be discovered from the particular set of documents with the help of cluster

algorithm .Keywords which occur a number of times are extracted by the system by using clustering algorithm and will detect the point of summarization from a collection of documents.Co-occurrences of terms are taken into account by the system which gives the best results.

In this paper [1] using mmr for diversity- based reranking and (2) evaluating summaries] develops a method for combining query relevance with information-novelty in the context of text retrieval and summarization. The Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

in re-ranking retrieved documents and in selecting appropriate passages for text summarization. Preliminary results indicate some benefits for MMR diversity ranking in ad-hoc query and in single document summarization. The latter are borne out by the trial-run (unofficial) TREC-style evaluation of summarization systems. However, the clearest advantage is demonstrated in the automated construction of large document and non-redundant multi-document summaries, where MMR results are clearly superior to non-MMR passage selection. This paper also discusses our preliminary evaluation of summarization methods for single documents.

In this paper [Automatic Summarization] It has now been 50 years since the publication of Luhn's seminal paper on automatic summarization. During these years the practical need for automatic summarization has become increasingly urgent and numerous papers have been published on the topic. As a result, it has become harder to find a single reference that gives an overview of past efforts or a complete view of summarization tasks and necessary system components. This article attempts to fill this void by providing a comprehensive overview of research in summarization, including the more traditional efforts in sentence extraction as well as the most novel recent approaches for determining important content, for domain and genre specific summarization and for evaluation of summarization. We also discuss the challenges that remain open, in particular the need for language generation and deeper semantic understanding of language that would be necessary for future advances in the field.

In this paper [1], they develop a method for combining query-relevance with information-novelty in the context of text retrieval and summarization. The Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance in reranking retrieved documents and in selecting appropriate passages for text summarization. Preliminary results indicate some benefits for MMR diversity ranking in ad-hoc query and in single document summarization. The latter are borne out by the trial-run (unofficial) TREC-style evaluation of summarization systems. However, the clearest advantage is demonstrated in the automated construction of large document and non-redundant multi-document summaries, where MMR results are clearly superior to non-MMR passage selection. This paper also discusses our preliminary evaluation of summarization methods for single documents.

In this paper [2], As the amount of online information increases, systems that can automatically summarize one or more documents become increasingly desirable. Recent research has investigated types of summaries, methods to create them, and methods to evaluate them. Several evaluation competitions (in the style of NIST's TREC1) have helped determine baseline performance levels and provide a limited set of training material. Frequent workshops and symposia reflect the ongoing interest of researchers around the world. The volume of papers edited by (Mani and Maybury, 1999) and a book (Mani, 2001) provide good introductions to the state of the art in this rapidly evolving subfield. A summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that. "Text" here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc. The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective as any reduction in the size of a document would carry a proportional decrease in its in formativeness. Luckily, information content in a document appears in bursts and one can therefore distinguish between more and less informative segments. Identifying the informative segments at the expense of the rest is

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

the main challenge in summarization. Of the many types of summary that have been identified (Borko and Bernier, 1975; Cremmins, 1996; Jones, 1999; Hovy and Lin, 1999), indicative summaries provide an idea of what the text is about without conveying specific content and informative ones provide some shortened version of the content. Topic-oriented summaries concentrate on the reader's desired topic(s) of interest, while generic summaries reflect the author's point of view. Extracts are summaries created by re-using portions (words, sentences, etc.) of the input text verbatim, while abstracts are created by re-generating the extracted content. Extraction is the process of identifying important material in the text; abstraction the process of reformulating it in novel terms, fusion the process of combining extracted portions, and compression the process of squeezing out unimportant material. The need to maintain some degree of grammaticality and coherence plays a role in all four processes.

In this paper[A Survey of Unstructured Text Summarization Techniques] Due to the explosive amounts of text data being created and organizations increased desire to leverage their data corpora, especially with the availability of Big Data platforms, there is not usually enough time to read and understand each document and make decisions based on document contents. Hence, there is a great demand for summarizing text documents to provide a representative substitute for the original documents. By improving summarizing techniques, precision of document retrieval through search queries against summarized documents is expected to improve in comparison to querying against the full spectrum of original documents

In this paper[A Bayesian Method to Incorporate Background Knowledge during Automatic Text Summarization A Louis – Proceedings of ACL] In order to summarize a document, it is often useful to have a background set of documents from the domain to serve as a reference for determining new and important information in the input document. We present a model based on Bayesian surprise which provides an intuitive way to identify surprising information from a summarization input with respect to a background corpus. Specifically, the method quantifies the degree to which pieces of information in the input change one's beliefs' about the world represented in the background. We develop systems for generic and update summarization based on this idea. Our method provides competitive content selection performance with particular advantages in the update task where systems are given a small and topical background corpus.

In this paper []

A summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Due to the problem of information overload, access to sound and correctly-developed summaries is necessary. Text summarization is the most challenging task in information retrieval. Data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. This paper presents a combined approach to document and sentence clustering as an extractive technique of summarization.

B. Construction of a personalized terms-sentences matrix

We have identified the construction of a terms-sentencematrix representing the document as a step suitable forpersonalization of the summarization. In this step termsextracted from the document are assigned their respectiveweights. Our proposed weighting scheme extends theconventional weighting scheme based on tf-idf

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

method bya linear combination of the multiple raters, which positivelyor negatively affect the weight of each term (see Fig. 1).

We formulate the weighting scheme as follows:

$$w(t_{ij}) = \sum_k \alpha_k R_k(t_{ij}), \qquad (1)$$

where*w(tij)* is a weight of a term *tij*in the matrix and *αk* is alinear coefficient of a rater R*k*. Both the weights *w(tij)* and thelinear coefficients *αk* can be any real number. The rater R*k*isa function, which assigns each term from the extractedkeywords set *T* its weight:

$$R_k : T \to \Re. \qquad (2)$$

C: THE raters have been divided into two sub groups

1. *Generic raters:* terms frequency rater, terms location

rater and relevant domain terms rater

Personalized raters*:* knowledge rater and annotations

Rater

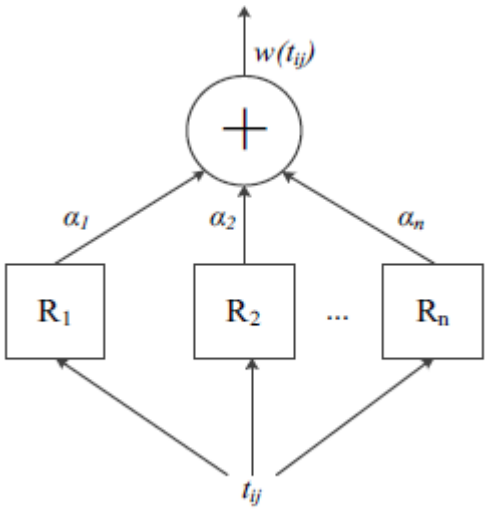

Figure 1. Term weighting by a combination of raters.

Features

- Summarized content and summarized point from the provided content will be provided by the system.
- The system will use clustering algorithm to extract keyword so that from the particular set of keywords the topic summarization will be discovered.
- Co-occurrence of terms is taken into account which gives best result.

Advantages

- User can specify how much % the content should be summarized.
- The algorithm provides quick result with the summarized data.
- Selects the best suitable points for summarization.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Disadvantages

- This system extracts words rather than phrases.
- The provided content must be more than 100-150 characters.

Applications

- This application can be used by many web users

II. CONCLUSION

MMR ranking user to minimize the redundancy by providing information in a useful and beneficial manner. Especially in the subject of query-relevant multi document summarization. Studies are currently performed to extend this into additional document collection also we will able to investigate handling of co-reference and analyzing the system as well as different parameters and clustering for output result.

Text summarization is still at a beginner stage in the world of evaluationMany different techniques can be applied to text summarization. But the evaluation of this technique is not considerable.

REFERENCES

[1] Ani Nenkova and Kathleen McKeown" Autmatic Summarization", Foundations and Trends R! inInformation RetrievalVol. 5, Nos. 2–3 (2011) 103–233c!2011 A. Nenkova and K. McKeownDOI: 10.1561/1500000015

[2] Dragomir R. Radev ,Eduard Hovy, Kathleen McKeown “Introduction to the Special Issue onSummarization”,

[3] Manjula.K.S ,Sarvar Begum , D. VenkataSwethaRamanna” Extracting Summary from Documents Using K-Mean Clustering Algorithm”

[4] [4] Annie Louis,” A Bayesian Method to Incorporate Background Knowledge duringAutomatic Text Summarization”, ILCC, School of Informatics,University of Edinburgh,Edinburgh EH8 9AB, UK

[5] Anjali R. Deshpande , Lobo L. M. R. J.,”Text Summarization using Clustering Technique”,International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- August 2013

[6] Sherif Elfayoumy,Jenny Thoppil,”A Survey of Unstructured Text Summarization Techniques”,(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 5, No. 4, 2014