Chapter 30
Statistical Features for Extractive Automatic Text Summarization

Yogesh Kumar Meena
MNIT Jaipur, India

Dinesh Gopalani
MNIT Jaipur, India

ABSTRACT

Automatic Text Summarization (ATS) enables users to save their precious time to retrieve their relevant information need while searching voluminous big data. Text summaries are sensitive to scoring methods, as most of the methods requires to weight features for sentence scoring. In this chapter, various statistical features proposed by researchers for extractive automatic text summarization are explored. Features that perform well are termed as best features using ROUGE evaluation measures and used for creating feature combinations. After that, best performing feature combinations are identified. Performance evaluation of best performing feature combinations on short, medium and large size documents is also conducted using same ROUGE performance measures.

INTRODUCTION

Day by day the information content on World Wide Web is exponentially increasing (BIG Data) along with the increase in the number of web users. In current scenario volume of information available on World Wide Web exceeded the textual information available in the printed form in libraries. The rapid growth of online information services makes it difficult to retrieve the relevant information quickly. While searching a particular information content, many a times information system users realize that the extracted information using currently available popular tools, that is present in the form of text, is not relevant to their information need at all, even after reading the whole list of text documents. They perhaps only waste their valuable time in reading the irrelevant text document. This problem can be

DOI: 10.4018/978-1-7998-0951-7.ch030

solved if users can be provided with a summary of the given text document. However, due to a large volume of available text data, that too dynamic in nature, it is very cumbersome for human experts to summarize all the documents manually. This issue leads to the requirement of an automated system that summarize the given text document automatically. This system that condenses the text document automatically and preserving its overall information content using a computer is termed as Automatic Text Summarization (ATS) system. There are numerous applications of automatic text summarization such as a snippet in information retrieval systems, news headlines as replacement to full story of news, electronic program guide in television systems. ATS is widely used for various domains such as business, news, legal and medical domains.

The researchers have classified text summarization systems (Hovy & Lin, 1995; Jones, 1997; Mani & Maybury, 1999; Gupta & Lehal, 2010; Lloret & Palomar, 2012; Nenkova & McKeown, 2012) on the basis of three main perspectives namely input, purpose and output. The third aspect which is most popular among these (i.e. output) considers summarization systems as either extractive or abstractive, and is mostly used by researchers. In extractive automatic text summarization (EATS), a subset of sentences from the original text document is selected for the final summary. Whereas abstractive automatic text summarization (AATS), sentences are fused and regenerated using natural language resources or rules. Abstractive text summarization requires deep knowledge resources, lexical/language resources, parsers and language generators. Because of these resource requirements, it is practically infeasible to use abstractive methods for automatic text summarization. Therefore, researches mainly focused on extractive text summarization instead of abstractive text summarization. The research work carried out is also focused on extractive automatic text summarization.

A typical extractive text summarization process completes in three steps namely pre-processing, sentence scoring, and summary generation. The first step in text summarization is preprocessing the input text document. Preprocessing mainly includes the sentence segmentation, stemming, stopword removal and special symbol removal. In the step which is sentence scoring, a score is assigned to each sentence of the document based on certain specific criteria. In literature various features are defined for sentence scoring such as term frequency, numerical data inclusion, sentence location, etc. After calculating each sentence score, rank is assigned to each sentence based on these scores. The sentences with the higher score are considered more important than the sentences with lower scores and are assigned high ranks consequently. After assigning a rank to each sentence, in third step, top ranked sentences are selected to generate the summary. Total number of sentences in the summary depends on the required length of it.

This chapter is focused on statistical features used in extractive approaches. In this chapter performance of these features is analyzed individually. After that, features that perform well termed as best features using ROUGE evaluation measures are used for creating feature combinations. After that, best performing feature combinations are identified. A deep performance evaluation of best performing feature combinations on short, medium and large size documents is also conducted using same ROUGE performance measures.

The rest of the chapter is organized as follows: After discussing challenges of ATS, background methods are presented followed by features used for sentence scoring. . Next, author discusses generic summarization process used for ATS. This is subsequently followed by evaluation matrices for ATS, followed by proposed impact analysis. Features and their combinations are properly analyzed before the chapter is finally concluded.