ABSTRACT

Text summarization is machine-based generation of a shortened version of a text. The summary should be a non-redundant extract from the original text. Most researches of text summarization use sentence extraction instead of abstraction to produce a summary. Extraction is depending mainly on sentences that already contained in the original input, which makes it more accurate and more concise. When all input articles are surrounding a particular event, extracting similar sentences would result in producing a highly repetitive summary. In this paper, a novel model for text summarization is proposed based on removing the non-effective sentences in producing an extract from the text. The model utilizes semantic analysis by evaluating sentences similarity. This similarity is provided by evaluating individual words similarity as well as syntactic relationships between neighboring words. These relationships addressed throughout the model as syntactic patterns. Word senses and the correlating part of speech for the word within context are provided in the semantic processing of matched patterns. The introduction of syntactic patterns knowledge supports text reduction by mapping the matched patterns into summarized ones. In addition, syntactic patterns make use of sentence relatedness evaluation in defining which sentences to keep and which to drop. Experiments proved that the model presented throughout the paper is well performing in results evaluation of compression rate, accuracy, recall and other human criteria like correctness, novelty, fluency and usefulness.

Keywords: Semantic Analysis, Syntactic Analysis, Syntactic Patterns, Text Summarization, Word Sense Disambiguation

1. INTRODUCTION

Text Summarization is considered to be one of the most widely used sub-fields of the Natural Language Processing (NLP) field. A summary is the text generated from one or more text that convey most important information, highly repeated concepts and core ideas in the original texts. The summary text is at most half in length of the original text. It is usually significantly less than half length. As the amount of 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. Text summarization has been developed and improved in order to help users manage all the information available these days. 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 (Barzilay, Elhadad & McKeown, 2001). The main goal of a summary is to present the main ideas in a document in a reduced size. 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 contained information (Lloret, Ferrández, Muñoz, Palomar, 2008). There are many categories for summarization: single document, multi-document, extractive, abstractive, informative, indicative, user-focused, generic, statistical, linguistic, and machine learning approach based.

Dingding et al. (Wang, Li, Zhu & Ding, 2008) proposed a new multi-document summarization framework. First -given a set of documents which need to be summarized- they clean these documents by removing formatting characters and decompose the set of documents into sentences. Second, they calculate sentence-sentence similarities using semantic analysis and construct the similarity matrix (through parsing each sentence into frame(s) using a semantic role parser. Pair-wise sentence semantic similarity is calculated based on both the semantic role analysis and word relation discovery). Third, symmetric matrix factorization is used to group sentences into clusters. Finally, in each cluster, they identify the most semantically important sentence - using a measure combining the internal information (e.g., the computed similarity between sentences) and the external information (e.g., the given topic information) - the most informative sentences are selected from each group to form the summary. While the system provided by Sarkar, K. (2009) consists of three primary components: first, document preprocessing (deals with formatting the input document, segmentation and stop removal), then, sentence ranking (assigns scores to the sentences based on the domain knowledge, word level and sentence level features) and then, summary generation (selects top n sentences based on scores). Finally, the sentences included in to the summary are reordered to increase the readability.

The model presented by Kowsalya et al. produced an extractive summary for given set of documents based on word sequence models by extracting Maximal Frequent Sequences MFS from the given text (Kowsalya, Priya & Nithiya, 2011).. They have employed the word sequence information from the self-text for detecting the candidate text fragments for composing the summary. To compose the effective summarization they have used the MFS technique to extract and detect the most important terms in the source document and Normalized Google Dissimilarity NGD distance for sentence clustering.

Nastase (2008) introduced an approach to allow the users to specify their request for information as a query/topic. To “understand” the query it expands it using encyclopedic knowledge in Wikipedia. The expanded query is linked with its associated documents through spreading activation in a graph that represents words and their grammatical connections in these documents. The topic expanded words and activated nodes in the graph are used to produce an extractive summary.

Hennig (2009) proposed a new method to the problem of query-focused multi-document summarization based on probabilistic latent semantic analysis to represent sentences and queries as probability distributions over latent topics. The approach combines query-focused and thematic features computed in the latent topic space to estimate the summary-relevance of sentences. Yih, Goodman, Vanderwende and Suzuki (2007) used a simple procedure based on maximizing the number of informative content-words. He first assigned a score to each term in the document cluster, using only frequency and position information, and then he found the set of sentences in the document cluster that maximizes the sum of these scores, subject to length constraints. This system has two components: one component uses machine learning to compute scores for each word in the set of documents to be summarized; the other
XML Documents Normalization Using GN-DTD
Zurinahni Zainol and Bing Wang (2013). Information Retrieval Methods for Multidisciplinary Applications (pp. 54-77).
www.igi-global.com/chapter/xml-documents-normalization-using-dtd/75901?camid=4v1a