Abstract

The rate of information growth due to the World Wide Web has called for a need to develop efficient and accurate summarization systems. With huge amount of information available on the World Wide Web, there is a pressing need to have Information Access systems that would help users, in providing the relevant information in a concise, pertinent format. There are various modes of Information Access including Information Retrieval, Text Mining, Machine Translation, Text Categorization, Text Summarization, etc. In this thesis, we study certain aspects of Text Summarization as a technology for Information Access.

Internet has provided easily any information, so people acquire more information in work and life, so automatic text summarization gained a lot of importance. People keep abreast of world affairs by listening to news bites. They take investment decisions based on the stock market updates. They go to movies largely on the basis of reviews. With summaries, they can make effective decisions in less time. A simple reason to prepare a summary would be to gain access to and control the flood of information, as everyone needs to know in brief what is worth reading and what is useful for a particular purpose.
Nobody wants to waste time reading what is useless. Summaries save readers time by giving an overview or outline of the content.

Summary should be informative, providing the condensed representation of a document’s contents. A summary should not be repetitive and it should outline the important aspects of the document in a precise way. Extractive text summarization aims to create a condensed version of one or more source documents by selecting the most informative sentences.

Most research on summary generation techniques still relies on extraction of important sentences from the original document to form a summary. Although many summarizing tools are available, it is becoming very difficult to generate meaningful and timely summaries, with the increasing volume of online information.

We perceived many writers to be having a common inclination in describing their texts. Most of the writers were observed to be having a tendency to use longer sentences, whenever they are at the crux of explaining a critical theme, perhaps due to an inability to elucidate the complex feature in simpler sentences. It is also observed that the words that are considered more important are observed to have repeating more frequently than those that are considered less important. Suppose, the writer of the text is explaining the ‘black hole phenomenon’, he naturally tends to use the word ‘black hole’, ‘universe’, many a time.
Having knowledge of these observations, a novel approach is proposed in this thesis for 'Automatic English Summarization' with the following postulates:

- The longest sentence in the text file will be an important sentence in the text.
- The frequently repeated words in the text are certainly the important words in the text.
- The association of the frequently repeated words with long sentences is a key factor.

With these three factors in mind, a well-defined procedure is followed to prepare a ‘summary’ of the given text. Any duplicate sentences generated, if at all, are removed by a Similarity-Search algorithm, and the resultant summary is then compared with the following, for the purpose of calculating efficiency:

- Summary generated by two jurors, who went through the whole text and prepared their own summary, extracting only the original sentences from the source text.
- Summary generated over the net by the 'Digital Library of India’s (DLI) Text Summarization Utility.

It is concluded that the procedure not only gave good summaries, but also is efficient in summarizing ‘technical’ content and the performance is satisfactory in ‘general’ content also.
The second approach is ‘Ontology Based Automatic Text Summarization’, in which we present an approach to sentence extraction that maps sentences to nodes of a hierarchical ontology. By considering ontology attributes we are able to improve the semantic representation of a sentence’s information content. Our experimental results show that the ontology-based extraction of sentences outperforms baseline classifiers, leading to higher accuracy of summary extracts.

As an analogy, one can describe the ontology of a domain as the relational schema of a database. Relational schema represents both the entities (concepts) and the dependency relations between the entities whereas the ontology consists of concepts and semantic relations between the concepts.

Each ontology node is populated by a bag-of-words constructed from a web search. Sentences are represented by sub trees in the ontology space, which allows to apply similarity measures in the ontology-space and to compute relations between sentences based on graph properties of the sub trees. Furthermore, node confidence weights computed by the classifier enable us to identify the main topics of a document. The classifier maps a sentence to the taxonomy by choosing the sub trees which best represent the sentence. If the maximum similarity of a child is lower than the current node’s similarity to the sentence, or if a leaf node has been reached, the algorithm stops. A sentence is therefore not necessarily classified to a leaf node, but may be assigned to an internal
node. For ontology-based summarizer we compute a set of features for each sentence based on the output of the hierarchical classifier. If a sentence is mapped to multiple sub trees in the taxonomy, we include all nodes from every sub tree. If a sentence is classified as a leaf node of a certain depth, it is assumed to contain more specific information than a sentence that is classified to a higher-ranked internal node. We create a bag-of-words for each sentence by removing stop words and applying stemming.

In this thesis we also explored the problems facing Indian language information access and quantified the extent of the problem. Our experiments on text summarization tasks using state-of-the-art algorithms used for English like languages showed low accuracy when applied to Indian language like ‘Telugu’.

So we presented our approach towards ‘Single and Multi-document Telugu Automated Text Summarization’. Our approach attempts to generate a text summary from the article of Telugu news papers, while avoiding the repetition of identical or similar information and presenting the information in such a way that makes sense to the reader. The proposed algorithm work as follows:

- Summarize each document
  - The first step of the process of generating a multi-document summary is to create individual single-document summaries for all documents in the set.
➢ Group the summaries in clusters

  o The second step of the multi-document (MD) summarization process is the grouping of the individual summaries into clusters. Documents are ‘clustered’ on the basis of the contents of their summaries, where a cluster consists of summaries that describe a similar topic.

  o For those documents that seem to discuss a similar topic, representative segments are eventually chosen for the MD summary while the other ones are ‘hidden’, i.e., not shown but still accessible to the user.

➢ For each cluster select representative passage(s) that will contribute to the final summary

  o The third step of the multi-document summarization process involves selecting a member of a cluster as a representative summary for the cluster. When the user wants a topical summary, the topic description is used to pick the summary that has most similarity to the topic.

  o In the case of a generic summary, the representative summary chosen is one that best represents the cluster. In this case, the summary that has most occurrences of the common terms across documents in a cluster is chosen. Since clusters can be overlapping it is possible that the same segment(s) is chosen to represent several clusters.

➢ Organize these passages in a logical way.
Finally, the last step of the multi-document summarization process involves organizing the selected passages in an order that makes sense to the reader. Currently, we organize the selected passages based upon topic similarity.

Finally it is concluded that the proposed procedure not only gave good summaries, but also with higher accuracy.