Abstract

Due to the increasing amount of available data online, researchers in the field of Natural Language Processing have focused heavily on the problem of automatically summarizing texts (NLP). Reducing the source text into a shorter form to contain the most important points and content of original text. Punjabi Shahmukhi script has been used by approximately three-fourth of Punjabi speakers around the globe, but it was never the focus of research activities, Gurmukhi has always been the focus. Gurmukhi Punjabi script summarizer is available but when it comes to the Punjabi Shahmukhi script, there is neither a publicly accessible dataset nor any kind of framework that can be used. There are two types of text summaries (1) Extractive Summaries (2) Abstractive summaries Extractive summaries are the of those sentences which are selected form the original given text based on most important line and Abstractive summaries are those one which don't have exact same sentences from the original text abstractive summaries include not just new vocabulary but also the synonyms of the terms that are used in the original material that they are based on. In our research on Punjabi Shahmukhi text summarization we have worked on extractive text summarization of Punjabi Shahmukhi Language and as per our knowledge we are the first one to purpose Punjabi Shahmukhi text summarization system. A wide variety of algorithms exist for summarization purposes. Term frequency inverse document frequency (TF-IDF) is one of these methods. It was the purpose of this study to develop a TF-IDF-based automatic text summarizer for Punjabi Shahmukhi text and to evaluate its performance in comparison to summaries written by humans. Sentence weight algorithm is also used for Punjabi Shahmukhi text summarization in our research which gives a percentage value to sentences and then selected in final summary on highest percentage basis. There is a very limited amount of data available of Punjabi Shahmukhi text so experimented our works on five documents of text We have used human generated summaries as our base line and compared those summaries with our TD-IDF automated generated summaries using Rouge-1, Rouge-2 and Rouge-L evaluation methods for text summarization and have generated some handful results accuracy of 80%, 75% and 79% respectively.