Automatic Pashto Text Summarization

The ever-growing volume of documents available online have forced the necessity of automatic text summarization. Summarization is the process of extracting summary of a given text. A good summary contains important aspects of the document. The summary needs to be informative and provide the most important information in a document. In addition, being non-repetitive and brief are the characteristics of a good summary. Sometimes, in a text, the same information is repeated to represent the significance of information. However, some other words which appears too frequently and convey no information, which are called stop words. But a summary should give u much precise information as possible. It is very difficult, if not impossible, for human beings to manually summarize large documents of text and process them. In this thesis we intend to conceptualize and implement automatic text summarizer for Pashto text. We apply statistical features measurement and machine learning techniques comparatively to achieve batter accuracy of summarization. Dataset of 200 documents, hybrid approach to identify and eliminate stop words from Pashto text, dictionary of 445 Pashto stop words, automatic text summarizer models, and a Pashto Natural Language Processing (NLP) web repository are our contributions in this thesis.