Intelligent Text Summarization: a strategy to reduce information misrepresentation

Shreya Jacob, Master of Science in Computer Science

University of Dublin, Trinity College, 2021

Supervisor: Prof. Khurshid Ahmad

The year 2019 would remain in history forever due to the outbreak of Covid-19. Even though it has been two years after the discovery of coronavirus, the situation in many nations remains uncontrollable. According to a few experts, one of the causes is public obliviousness due to distortion of the truth. In an unfamiliar environment, the requirement for precise information is paramount. The information gets diffused on a text cline from scientific papers to science magazine articles, then to newspapers, and then to the general public via social media, where half of it gets lost or corrupted. The accurate findings of the researchers get suppressed. This project work designs an automatic text summarization system based on the theory of lexical cohesion that can efficiently extract the pertinent information from the research papers to reduce the misrepresentation of text in the first level of text cline. The notion of lexical cohesion is that the repetition of words in the sentences creates a bond between them and brings the text closer. Identifying the highly bonded sentences would thus aid in creating a summary that is concise and meaningful. The concept of using an external keyword list that consists of top terms present in the domain for keyword identification was a significant contribution of this work. The system efficiency was statistically evaluated using various metrics like average sentence length, readability, sentiment similarity, and syntactic similarity. The evaluation results of the summaries generated for ten research papers confirmed the efficiency of the algorithm when compared to their abstracts (human-generated summaries). Although the summaries were less readable than the abstracts, they were highly similar to the original text on sentiment and syntactic similarity.