Intelligent Summarization: Leveraging Cohesion in Text

Arun Thundyill Saseendran , Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Professor Khurshid Ahmad

Text summarization is a task that requires the application of human intelligence in which the human shows an understanding of natural language and can process language, where human beings show creativity by presenting complex objects and events. Human beings use a collection of words in discussing specific subjects - the so-called specialist or scientific lexicon, that comprises the ontology of a domain. A computer program that can mimic these aspects of human intelligence is referred to as an information extraction system, and text summarization systems of the type discussed in this thesis are called extractive text summarization systems. In this work, a text summarization system is presented that is based on the theory of lexical cohesion, where the focus is on how a (scientist) writer repeats a specific word to convince his or her reader about the importance of the theme. The *Intelligent Text Summarization* algorithm makes use of repetition that can be visualized as a graph where the links between nodes represent the linked sentences based on the same word or its close variants. The novelty introduced is the autonomous selection of domain-specific words to produce a readable summary selecting pre-existing sentences in the text. The average summary is about 25% of the text. The system (developed using Java and Python) was tested using computer science texts and was tested using texts in bio-medicine, specifically gastroenterology. An expert compared the summaries of 15 papers generated by the application and found that in 73.6% of the cases the summary was good.