

Indonesian Automated Text Summarization

Gregorius S. Budhi¹, Rolly Intan², Silvia R³., Stevanus R. R.

greg@petra.ac.id, rintan@petra.ac.id, silvia@petra.ac.id

Petra Christian University, Informatics Engineering Dept.

Siwalankerto Street 121 - 131, Surabaya 60236, Indonesia

Abstract

Automated Text Summarization (ATS) is a computer-based application to produce a summary from an article but it still keeps an accurate main point from the content of the article. In this research, we build an Indonesian ATS application using a virtual graph concept, calculation weight of sentence and weight of relation among sentences, Deductive – Inductive method in Indonesian Language and Exhaustive Shortest Path Algorithm to provide a summarization path from the first sentence to the last sentence on every paragraph in the article. The test result shows that the quality of summarization result depends on the type and the structure of the article. System will produce a good summary if the type of the article is an argumentation type and the article's structure has many paragraphs in which each paragraph has more than two sentences.

Keywords: Automated Text Summarization, Indonesian Deductive-Inductive paragraph, Dijkstra's Algorithm

1. Introduction

Automated Text Summarization (ATS), simply called Text Summarization, is a computer-based application that summarizes an article by which the results of the applications still keep the main point of the article. The main goal is to show only some main sentences of the article as a summary, and hopefully by reading the sentences, readers will save their time in understanding as much as possible the whole content of the article.

In this research, we implement text summarization using type of *informative summaries* that only concerns to provide information based on the important aspects of the article by searching and including as much as possible all relevant topics. The summaries will ignore the irrelevant topics and unimportant detail or supporting information. They only produce the important section from the article. Generally, the type of summaries is used for the overview of article.

2. Basic Theory

2.1. Graph

A graph $G = (V, E)$ consists of a set of vertices, V , and a set of edges, E . Each edge is a pair (v, w) , where $v, w \in V$. Edges are sometimes referred to as arcs. If the pair is ordered, then the graph is called directed graph or *digraph*. Vertex w is adjacent to v if and only if $(v, w) \in E$. Sometimes an edge has a third component known as a weight or a cost [8].

2.2. Weight of Sentence

Weight of sentence is a value of a sentence to determine how the sentence plays important meaning in a paragraph of an article. Clearly, higher weight of a sentence in a paragraph means existence of the sentence plays more important role

in the paragraph. Thus, in the process of summarization, higher weighted sentences should have higher priority to be chosen as a part of summary. It is necessary to ignore unimportant words in the article before starting to calculate the weight of sentences. In other words, *stemmer process* and *stopword removal* have to be already implemented in the article before calculating the weight of sentences. Here, the weight of sentences consists of the following 4 component values, $W1$, $W2$, $W3$ and $W4$ based on [7] by some modifications as follows.

$W1$ be a score concerning similar words in a sentence compared with a list of keywords in the article. Supposed an article has a keywords list. Words in the sentence are compared to the words in the keywords list of the article. If there are more words in the sentence are similar to the keywords, the value of $W1$ will be higher.

$W2$ be a value to express the frequency of words of a sentence in a article. The number of same words in the sentence and the article is calculated. The result will be divided by the total words in the article by also considering the frequency of the words in the article. If the sentence has higher score result from the calculation, it means the sentence consists of more words that have high frequency in the article.

$W3$ be a value that is determined by the position of a sentence in a paragraph. Weight of sentence is also determined by the position of the sentence in a paragraph of an article. In general, every paragraph in a good writing of an article usually only provides one main idea. Related to the *Deductive-Inductive method* in the Indonesian grammar, the first and the last sentences in a paragraph are usually considered as main