Ekstraksi Sinonim Set dengan Agglomerative Hierarchical Clustering pada Wordnet Bahasa Indonesia

Gilang Rananda Tama, Mochammad Arif Bijaksana

Fakultas Informatika, Universitas Telkom, Bandung
gilangrananda@students.telkomuniversity.ac.id, arifbijaksana@telkomuniversity.ac.id,

Abstract

Wordnet Indonesia is a lexical dictionary developed by the Retrieval Information Lab, Faculty of Computer Science, University of Indonesia. Currently the Indonesian wordnet can be said to be far from perfect and still requires more development and also the vocabulary contained in it is relatively small. Because the development of the Indonesian Wordnet is considered not that interesting, the amount of research on natural language processing still not that much . The development that will be carried out by the author here is by implementing the Agglomerative Hierarchical Clustering method as a means to process data and determine performance in the development of the Indonesian Wordnet. The way this method works is by grouping words that have similarities in meaning and then put them into one cluster, after that we need to checked again whether there are other words that have the same meaning, if not then it will be a cluster in which there are words with the same meaning or can also be called a synonym set. For the dataset, the author takes data from the Indonesian Standard Thesaurus. The data from this study were obtained manually and randomly. For each step, first calculate the distance matrix between the data to find out which words are eligible to be included in the cluster, then combine the 2 closest groups based on the proximity parameters that have been searched previously and then update the proximity matrix to represent the proximity of the new group to the remaining groups. Repeat continuously until only 1 group is left. And later on, the results will be in the form of several clusters that can be used as a reference for the development of the Indonesian Wordnet.

Keywords: WordNet, Agglomerative Hierarchical Clustering, Indonesian Thesaurus