JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 99, Nomor 99, Bulan 2020, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v3i4.9999

Building Synonym Set for Indonesian WordNet Using Commutative Method and Hierarchical Clustering

Valentino Rossi Fierdaus[1], Moch. Arif Bijaksana[2], Widi Astuti[3]

[123] Faculty of Informatics, Bachelor of Informatics Engineering, Telkom University, Bandung, Indonesia
Email: [1] valentinorfs@student.telkomuniversity.ac.id, [2] arifbijaksana@telkomuniversity.ac.id, [3] astutiwidi@telkomuniversity.ac.id

Abstract− WordNet is a compilation of Synonyms Set (synset), which consists of the words that have the same synonymous. The development of Indonesian WordNet has a goal to build an application that can accommodate and exhibit the relation of words. Synonym Set is a set composed of one or more words that have a similar meaning or synonym relation originated from the Indonesian Thesaurus. In previous studies, the establishment of synsets were transmitted with several approaches, one of which was the cluster ring to produce synsets and WSD (Word Sense Disambiguation). In this research, research is held up to discover the semantic similarities between words in the Indonesian Thesaurus automatically, and also to know the performance of the Agglomerative Hierarchical Clustering method for the development of Indonesian synsets. To calculate performance and evaluation, this research is using the F-measure method involving the gold standard.

Keywords: WordNet, Synset, Indonesian Thesaurus, Agglomerative Hierarchical Clustering, F-Measure.

1. INTRODUCTION

WordNet is a set of several synonyms called Synonym Set (Synset) consisting of words that have equivalent meanings or sense which are interrelated [1]. At first, WordNet is a semantic dictionary that made in English version which was first built by Princeton University, then along with development of technology, WordNet at present is one of the most widely used sources of referral information. Language dictionaries around the world in general is a dictionary that has a focus words while WordNet focuses on the meaning of words or synonyms. In WordNet, several classes of words such as nouns, verbs, adjectives, and adverbs is grouped into a synsets. Lines of words in WordNet can symbolize a meaning which is called synset.

In the process of building WordNet, the first thing to do is produce a synset or collection of synonym that have same meanings [1], that means the words are grouped into a synset according to their meaning. That is because synset is a basic concept that supports the formation of semantic relations in the lexical database [2]. Monolingual resource that used as a lexical resource is Thesaurus, because Thesaurus contains words that have an interrelated synonym relation [1]. Thesaurus that has been through the extraction process, will produce one or more synset. To combine the synset that produced from the previous process, one way to produce the best synset is using the clustering techniques. Therefore, need some further research to find out the performance of clustering techniques in the development of synset for Indonesian WordNet.

Previously, there was a development of Indonesian WordNet using Hierarchical Clustering. In that study, the data that used as input is a synset that was generated from the commutative process, then that data (synset) will be grouped and combined in the Clustering process. However, the data used as input are data generated from the results of manual commutative process. This research will focus on two main stages, the first is the stage to doing synset extraction, and the second stages is the process of combining synsets using clustering technique if in the first stages there is a word produces more than one synset. In the synset extraction process, to produce a valid synset value will use Commutative method using available monolingual resources that is Thesaurus Bahasa Indonesia, this means that if a word k1 has a synonym k2, then k2 must also be a synonym of k1. In fact, commutative relations like this do not always occur in Thesaurus Bahasa Indonesia [1]. And for the second stages, clustering technique that used in this research is Agglomerative Hierarchical Clustering.

The purpose of this research is to find out the semantic similarities between words in Thesaurus Bahasa Indonesia automatically, and also implement the Agglomerative Hierarchical Clustering method on the system to be built to determine the performance of that clustering techniques in the development of synset for Indonesian WordNet.

Submitted: **dd/mm/yyyy**; Revised: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**