INTRODUCTION

With the rapid development in this technological era, the need to find accurate and fast information becomes even greater (Gunawan & Saputra, 2010), even the use of computers and the internet is becoming common for daily activities in the fields of research and education are not much different. The information needs of words that are used every day will usually be contained in a language dictionary but the language dictionary does not provide word synonyms (Samhith et al., 2016). Because of that English WordNet was made to assist in providing information automatically by searching for techniques or searching arranged in alphabetical order (Fellbaum & Miller, 1998). Wordnet itself was first developed by Princeton University which aims to accommodate native English speakers by lexical modelling (Gelbukh, 2007). Currently the development of WordNet (PWN) version 3 already has 117,000 synchronization and 206,941 word pairs (Miller, 1995) so that it will continue to be developed to become a perfect word in Lexical (Zhang & Hasi, 2015).

In English, a word has one or more meanings. Several words that are different but have the same meanings are called synonyms, and for different meanings they are called antonyms (Ilson, 2011). For a word that has a meaningful relationship between one word with another word, such as hyponym, hypernym, anonymous, and others (Kim & Kim, 2008). In wordnet development, words are grouped according to their meaning into a synonym set or synset (Chen et al., 2009). Synset is a part that is formed in the early stages of building a lexical database (Swain et al., 2019). This happens because synset is a basic concept that supports the formation of semantic relations in the lexical database (Zhang & Hasi, 2015). Thesaurus as a monolingual resource that is used as a lexical source because the English Thesaurus contains words that have a synonymous relationship (Priyatno & Bijaksana, 2019). Thesaurus itself is a dictionary that contains a collection of words and has interrelated meanings (Hendrik & Cahyono, 2017). The clustering process with ROCK method will help the calculation to produce a value that shows the similarity of a word, so that it helps improve the accuracy of words in the English wordnet.

Thesaurus that has been through the extraction process will produce several synset (Zhang & Hasi, 2015). To combine several synset generated in the previous process the clustering process is used. Clustering is used to combine several synset that have a similarity (Guha et al., 2001). In this study the clustering method used is ROCK. The clustering method is used because the results of the ongoing clustering process cannot be predicted before the clustering process is complete (Dembczynski et al., 2011). This test is expected to increase the accuracy of the English wordnet because in previous studies we got a value of 10.68% accuracy (Priyatno & Bijaksana, 2019) by using the hierarchical clustering method and the process does not calculate the max goodness value and directly computed the threshold using f-measure (Priyatno & Bijaksana, 2019), so the count phase is less accurate. The addition of precision in calculating the synset itself is intended for the accuracy of a synset so that when used in internal research. The development of English wordnet itself helps the construction of wordnet in other languages and is used as a benchmark (Zhang & Hasi, 2015). Advantage of the ROCK method itself is a way of removing group outliners that occurs when the clustering process is 1/3 amount of available data. And all that is intended to add the level of accuracy synset results (Guha et al., 2001). Therefore, building English wordnet using the ROCK method will be very helpful in developing wordnet itself and, will help in expanding wordnet in other languages, the more words produced the better for use of the word to utilize. Because with more references that will increase the level of accuracy and the more words will be converted into wordnet. Based on the aim to improve accuracy in the building of English wordnet, using the ROCK method is one method that is better than those that only use the hierarchical clustering method. So that there will be an increase in accuracy in English words and will help in the development of wordnet in other languages.

METHOD

In this study several methods will be used to calculate word equations in English wordnet using the ROCK method that will assist in its development, the following methods will be used

1. Similarity value
   Similarity value is a clue to determine the level of closeness between words. Then from the similarity value can be grouped based on the level of closeness to a certain threshold.

2. Synset
   The structure contains word information contained in wordnet, there is also a class of words and definitions of all word sets contained in a language that will become a single, interconnected entity. In general, the smallest unit in a language dictionary is a word, but it is different from wordnet because the smallest unit is a synset (Jain & Lobiyal, 2019). Synset is a basic concept that supports the semantic relations of lexical databases (Swain et al., 2019).

3. ROCK (Robust Clustering Using Links)
   In this study the ROCK Clustering technique is used. After calculating the similarity value in each word, then the word will determine whether the word is a neighbour or not, in this method the link is used as a reference in the clustering process of counting the number of neighbours in the word. The number of clustering algorithms whose data is numeric is one of them is hierarchical clustering. Hierarchical Clustering is a grouping of objects where each similar object will be close together (Guha et al., 2001), while the non-similar will be far apart. But there will be problems that arise in the value of attributes that are categorical (Guha et al., 2001), often objects that have a small similarity value will be grouped in one cluster regardless of whether the objects have in common or not. The solution to deal with these problems is to use the ROCK (Robust Clustering Using Links) algorithm. The clustering algorithm used in this research is:

   1) Matrix calculation for distance value. The value in the matrix is found from the calculation of the number of equations found in the two synset. This value becomes the similarity value that has been obtained and then divided by the unique words that exist in both synset.
   2) Then take the first maximum distance value obtained from the matrix. This value will be used to find the threshold value. The threshold value is obtained by multiplying the first maximum distance value by the coefficient value. The coefficient value can be changed manually.
   3) After that, combine the two synonym sets that have the same maximum distance value.
   4) Then do the recalculation to get the distance value matrix.
   5) If the maximum distance value is not lower than the threshold value, repeat the third step, stop the clustering process if not.

4. Gold Standard

   Gold Standard has the aim of knowing the magnitude of a correlation of the results of a score issued by the machine to the relevance of the word being tested. The value of the Gold Standard is produced from a collection of human opinions. The resulting value becomes the reference or measurement standard for similarity between words. The Gold Standard used in this study is the result of the validation of the synset performed by a lexical expert or called a lexicographer. Validation is done carefully so that it can be a comparison for the results of a system to measure accuracy.

5. F-Measure

   F-measure is a popular metric in terms of performance, especially in tasks with unbalanced data sets (Dembczynski et al., 2011). The method involves precision and recall. To calculate the recall (R) and precision (P) can be determined from equations 1 and 2. The role of humans is needed in determining the Gold Standard and in determining the threshold in grouping words based on the value obtained from similarity. The F-Measure method calculates the double proportion multiplied by the results of the first method (precision) and the second method (recall) divided by the sum of the two. The equation can be seen in equation 3.