e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 4 No. 1, Juni, 2020, **Hal. 57-62**
DOI: 10.29408/edumatic.v4i1.2063

Building Synonym Sets for English WordNet with Robust Clustering using Links Method

Sarah Suryaningsih[1], Moch Arif Bijaksana[2], Widi Astuti[3].
[1,2,3] Department of Informatics Engineering, Universitas Telkom
email: sarahsy@student.telkomuniversity.ac.id[1], arifbijaksana@telkomuniversity.ac.id[2], widiwdu@telkomuniversity.ac.id[3]

(Received: 26 April 2020/ Accepted: 26 Mei 2020 / Published Online: 20 Juni 2020)

Abstract

English WordNet is an important synonym set to present the similarity of meanings between words. Synonym Set is built using Oxford Thesaurus which is accessed through lexico.com, which is a part of the lexical database that will be used. After using the extraction process through Oxford Thesaurus it will produce a synonym set with the same meaning between words. The difference between WordNet and ordinary dictionaries is that the word is interconnected with other words. One method employed for this approach is Robust Clustering Using Links method, which is similarity values and synonym sets that have been created to be used to build a lexical database. Therefore the main purpose of the development of the English WordNet is to produce an accurate synonym set using clustering techniques. The evaluation calculation will use the F-measure method and will use the gold standard for the calculation method. With the ROCK method, there is an increase in accuracy output from dataset input. Building the English wordnet is to improve words that can be used to help research and development of other language wordnets with role models using more accurate English wordnets. And the use of ROCK method there is an increase in the accuracy upon results of the development of English wordnet compared to the previous method, which is using hierarchical clustering. The outcome of this study resulted in improved accuracy so that the ROCK method is one of the good methods used in the development of the English wordnet.

Keywords: F-measure, Gold Standard, Robust Clustering Using Links, WordNet

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:
   a. 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.
   b. 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.
   c. After that, combine the two synonym sets that have the same maximum distance value.
   d. Then do the recalculation to get the distance value matrix.
   e. 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.

RESULTS AND DISCUSSION

In this section the results of the test will do the clustering process with a threshold of a range of 0.1 to 0.9 can be seen in the following table 1. Which from previous method did not occur the results of calculations using word comparisons between max goodness values that produce a synset result. Then the value will be processed in the calculation using f-measure.

Table 1. Comparison of Threshold values

Threshold	Max Goodness Measure	Synset Results
0.1	0.08	6
0.3	0.21	5
0.5	0.54	5
0.7	1.41	4
0.9	6.41	4

In the results of the above table calculation, it can be seen that the greater the threshold value, the less the resulting synset. Threshold 0.1 is the smallest of the value taken to do the calculation and will get a max goodness measure that will affect the number of synset results and calculations on the f-measure. In the calculation of max goodness value, resulting from determination of the range threshold and then compared with similarity value from each word. And the process will take place as long as the range of threshold value is greater than the similarity value. Which why the greater of threshold value used, the greater accuracy in determining the number of synset results. The threshold value itself is like a filter to see which words are next to the inputted words, and helps in the next calculation which will add the number of links and the level of accuracy. So in the table 1 will produce English words that have been determined similarity value to be calculated toward the next stage, which helps in improving the English wordnet.

Table 2. Testing Results using f-measure

Dataset	Precision	Recall	F-measure
Before doing Clustering	12.82%	11.36%	12.05%
0.5	13.89%	11.36%	12.5%
0.7	13.51%	11.36%	12.35%
0.9	13.16%	11.36%	12.2%

In calculating the Rock method, the value of the threshold is very important. In table 2 it can be seen when clustering is greatly affected by the value of the threshold and in the search for a max goodness measure it can determine whether two synsets will be merged or not in the clustering process, and will affect the number of links that will be generated. In this test, the threshold values range from 0.1 to 0.9 and use a threshold value of 0.5 until 0.9 to see the results of recall, precision in calculations using the f-measure. Because the threshold range 0.5 to 0.9 produces a smaller synset than the values 0.1 and 0.3. By using a range of threshold, the comprising words 1 & 2 will be calculated whether the grade is greater than max goodness value, if the grade of a word is smaller than max goodness value, deleting the word will occur. And the process will repeat until the words being compared run out. Therefore, in the calculation of accuracy will be influenced by the amount of the threshold value used, then the clustering process will determine whether the two synsets are neighbors or not, this result determines the number of links that will be generated. The precision value changes because this value is influenced by the previous threshold and the largest value is at 0.5 for 13.89%, this shows the increased accuracy of the value before clustering. At values 0.7 and 0.9 there is a decrease in value and this shows a decrease in the accuracy of the threshold value. The calculation of the word value will pass the f-measure process, which is must be searching for recall and precision values, and then divide them. All of these values will produce the last of the f-measure, which will be the final value of determining the amount of accuracy of the synset calculation. The results from this stage help to improve the accuracy of synset processing, so that the words produced are better than before.

Table 3. Example The Results of Synonym Set

Input Words	Dataset (input)	The output
wisp	[wisp, piece, scrap, shred, thread]	[wisp, piece, scrap, shred, thread]
hosiery	[hosiery, socks, knee socks, ankle socks]	[hosiery, socks, hose]
roadblock	[roadblock, barrier, barricade, blockade, obstruction, checkpoint]	[roadblock, barrier, barricade]

In table 3 we can see the output of the previous word and its synonyms will be processed and the results of the process produce a more accurate synset using a threshold value of 0.5 whose calculation results are greater than the other values. At the output, we see the removal of several words from the dataset, after doing the clustering process and counting using the f-measure, the accuracy of a word with other words occurs, whether a neighbour or not. So, table 3 is the output of a few words that have passed those calculation or filtering process, also in order to make the synonyms more accurate and in the calculation, process has been resolved better than before. With a comparison of several threshold values that become the first key and continued with a calculation of measure which produces the best final value of 12.5%. And has passed checking using the gold standard, before the final results are made into example for table 3. This is the result of all the previous calculation stages, and examples with some through the advantages of using ROCK method for building of the English wordnet.

CONCLUSION

Based on the results of tests that have been done, it can be concluded that in the search for a good synonym set is to use the clustering process and determine the best threshold value by looking at the results of the closest synset data with the validation value. In this test the best threshold value is 0.5 with an accuracy value of 12.5% of the total data set of 50 words with calculations using the f-measure. From the results of testing the previous method with the best results 10.68% using the hierarchical clustering method (Priyatno & Bijaksana, 2019).Therefore, it can be said that testing using the Rock method is better than the previous test to determine the synonym set in the development of the English wordnet because with an increase of about 2% of the value of the previous measure. With a comparison of each f-measure that occurs in words that are input by appeal the similarity of each word's meaning and there is an output with a higher level of accuracy than before and determine the number of links contained in each of these words.

REFERENCES

Chen, D., Jianzhuo, Y., Liying, F., & Bin, S. (2009). Measure Semantic Distance in WordNet Based on Directed Graph Search. *International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government*, 57–60. https://doi.org/10.1109/EEEE.2009.16

Dembczynski, K. J., Waegeman, W., Cheng, W., & Hüllermeier, E. (2011). An Exact Algorithm for F-Measure Maximization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24* (pp. 1404–1412). Curran Associates, Inc. http://papers.nips.cc/paper/4389-an-exact-algorithm-for-f-measure-maximization.pdf

Fellbaum, C., & Miller, G. (1998). The Lexical Database. In *WordNet: An Electronic Lexical Database* (p. 22). MITP. http://ieeexplore.ieee.org/document/6285385

Gelbukh, A. (2007). *Computational Linguistics and Intelligent Text Processing: 8th*

International Conference. Springer Science & Business Media.

Guha, S., Rastogi, R., & Shim, K. (2001). ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*, *25*, 345–366. https://doi.org/10.1016/S0306-4379(00)00022-3

Gunawan, & Saputra, A. (2010). Building Synsets for Indonesian WordNet with Monolingual Lexical Resources. *International Conference on Asian Language Processing*, 297–300. https://doi.org/10.1109/IALP.2010.69

Hendrik, & Cahyono, A. (2017). Model WordNet Bahasa Indonesia berbasis Linked Data. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *6*(1), 8–14. https://doi.org/10.22146/jnteti.v6i1.288

Ilson, R. (2011). On the Historical Thesaurus of the Oxford English Dictionary. *International Journal of Lexicography*, *24*(3), 241–260. https://doi.org/10.1093/ijl/ecq032

Jain, G., & Lobiyal, D. K. (2019). Word Sense Disambiguation of Hindi Text using Fuzzified Semantic Relations and Fuzzy Hindi WordNet. *9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 494–497. https://doi.org/10.1109/CONFLUENCE.2019.8776967

Kim, Y. B., & Kim, Y. S. (2008). Latent Semantic Kernels for WordNet: Transforming a Tree-Like Structure into a Matrix. *International Conference on Advanced Language Processing and Web Information Technology*, 76–80. https://doi.org/10.1109/ALPIT.2008.40

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Priyatno, J., & Bijaksana, M. A. (2019). Clustering synonym sets in english wordNet. *7th International Conference on Information and Communication Technology, ICoICT 2019*. https://doi.org/10.1109/ICoICT.2019.8835313

Samhith, K., Tilak, S. A., & Panda, G. (2016). Word sense disambiguation using WordNet Lexical Categories. *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES),* 1664–1666. https://doi.org/10.1109/SCOPES.2016.7955725

Swain, D., Tambe, M., Ballal, P., Dolase, V., Agrawal, K., & Rajmane, Y. (2019). *Lexical Text Simplification Using WordNet* (pp. 114–122). https://doi.org/10.1007/978-981-13-9942-8_11

Zhang, Y., & Hasi. (2015). A Constructing Method of Mongolia-Chinese-English Multilingual Semantic Net Based on WordNet. *International Conference on Computer Science and Applications (CSA)*, 196–198. https://doi.org/10.1109/CSA.2015.47