Development Synonym Set for the English Wordnet Using the Method of Comutative and Agglomerative Clustering

Munirsyah[1]*, Moch. Arif Bijaksana [2], Widi Astuti[3]
Bachelor of Informatics Engineering, Faculty of Informatics
Telkom University, Bandung, Indonesia
[1]munirsyahh@student.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id, [3]astutiwidi@telkomuniversity.ac.id

Abstract*— Wordnet is a collection of words that interpret or present a meaning, in its development Wordnet has an important part, the Synonym Set or Synset. In making Synonym sets, synonyms are needed and the commutative nature of words is needed. To get word synonyms, the English language thesaurus becomes the reference data for taking synonym data. Broadly speaking, the difference between Wordnet and the dictionary is that the meaning of the word is related to other words, to determine the equation requires a commutative process. The process is made easy by using commutative methods that will produce a candidate synonym set. Candidates for the synonym set cannot be used for word syntax, the grouping process of words which produces the Synonym set as the final result must be carried out. The process of grouping words can one of them use clustering techniques, in this study will use Agglomerative Clustering techniques. In the process of agglomerative clustering techniques there is a threshold value to determine the number of repetitions or as a condition to stop the iteration process. The clustering process in this study will use a threshold value of 0.1 to 1 to test the best threshold value to produce the best Synonym set and calculate its accuracy value. Accuracy calculation and evaluation will use the F-measure method to find the best results.

***Keywords*— Wordnet, Synonym Set, Agglomerative Clustering**

I. INTRODUCTION

The English Wordnet application has previously been developed by Princeton University in the United States which aims to model the lexical knowledge of native English speakers, where the results of its development take the form of desktop-based applications. Putting conventional dictionaries on line seems a simple and natural marriage of the old and the new [1].

Wordnet becomes a system that can provide information automatically because it has a dictionary concept by using a searching method rather than a dictionary in general that uses the alphabetical order [2].

The smallest unit of Wordnet is the Synonym set which represents the specific meaning of a word [3]. Each Synonym set contains the form of a Synonym set of words and a semantic pointer that explains the relationship between one Synonym set and another Synonym set [1]. Synonym set can also be called a collection of words in Wordnet that can represent one meaning, apart from the representation of meaning, a word also has a relationship between words such as hypernym, hyponym, holonym, meronym, and others [4].

In Wordnet development, synonyms of related or commutative words are needed, synonyms of words are obtained from English Thesaurus, the word has several synonym meanings and is used as data. To get commutative data, the data is processed using the commutative method, the commutative method is a data processing technique by comparing the first word with the second word, in the comparison the data checking process is related or commutative, said commutative if first word has the meaning of second word and second word has first word. The process of the commutative method produces the synonym set.

For the next process, a synonym set using Clustering, Clustering is used to group words that are similar. In this study the clustering used is Agglomerative Clustering. Agglomerative Clustering is a bottom-up Hierarchical Clustering method that combines n clusters into one single cluster [5].

According to tan Agglomerative clustering has the advantage that it does not need to determine the number of clusters and does not take into account initial centroids. these two things show the accuracy of using agglomerative clustering because to build wordnet data to be processed will be very much and not optimal if there is a cluster value in the grouping of data.

In the agglomerative clustering technique process there is a threshold value to determine the number of iterations or as a condition to stop the iteration process. The clustering process in this study will use a threshold value of 0.1 to 1 to test the best threshold value to produce the best Synonym set and calculate its accuracy value. Accuracy calculation and evaluation will use the F-measure method to find the best results [1] [5].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.855, Copyright ©2020
Submitted : 14 Mei 2020, Revised : 30 Mei 2020, Accepted : 30 Mei 2020, Published : 4 Juni 2020

II. LITERATURE REVIEW

A. *Wordnet*

WordNet is an online English lexical reference system [6]. Information in Wordnet is organized into logical groups called Synset [3]. Generally, a language dictionary is a dictionary that has a focus on the meaning of words while Wordnet be built focuses on the similarity of word meanings (synonyms). A collection of words in Wordnet can represent one meaning or can be called Synset [4].Maintaining the Integrity of the Specifications.

B. *Synonym Set*

Synonym set (Synsets) are lists of terms or collocations that have the same meaning and in certain contexts the uses are interchangeable. Every member in the synonym set can replace one member and another member without changing the meaning [6].

C. *Thesaurus*

A thesaurus is a reference book in the form of a word list with its synonyms, a reference book in the form of information about various sets of concepts or terms in various fields of life or knowledge [7]. This allows the user to use various alternative search words compared only with indexes and catalogs [8].

D. *Clustering*

Agglomerative Clustering is one of the techniques of grouping objects based on their characteristics, which starts with individual objects until they combine into a single group [9]. In the Agglomerative Clustering algorithm, the process starts from the cluster that has the lowest point and merges a cluster with other clusters, the clusters are close together or have the highest level of similarity. Data is grouped based on distance values [tan, 2006], the Clustering algorithm is used to produce small groups and describe these groups using expressions [10], to get distance values using the following equation:

$$\text{Distance Value} = \frac{\text{Similar Words}}{\text{Unique Words}} \tag{1}$$

E. *Commutative Method*

The commutative method is a data processing technique by comparing the first word with the second word. This method is used because building Wordnet requires correlated or commutative data [11]. Data can be called related if the first word has the meaning of the second word and the second word has the meaning of the first word [12].

F. *Gold Standard*

Gold Standard has the aim to find out the magnitude of the correlation of the results of a score issued by the machine to the relevance of the words tested, the value of the Gold Standard is produced from a collection of human opinions.

G. Preprocessing

Preprocessing is the process of transforming text into collections of words. Text is unstructured data, which is quite difficult to process with a computer [13].

H. *F-Measure*

F-Measure is an accuracy test that uses Precision and Recall as a benchmark, Precision is a true positive prediction that is compared with the overall positive predicted result. Whereas Recall is a true positive prediction that is compared with an overall positive true result [14]. F-measure measures the effectiveness of a model taking into account every distribution class [15]. Calculation of F-Measure values can use the following equation:

$$\text{Precision} = \frac{(TP)}{(TP+FP)} \tag{2}$$

$$\text{Recall} = \frac{(TP)}{(TP+FN)} \tag{3}$$

$$\text{FMeasure} = \frac{2*\text{Precission}*\text{ Recall}}{\text{Precission}+\text{Recall}} \tag{4}$$

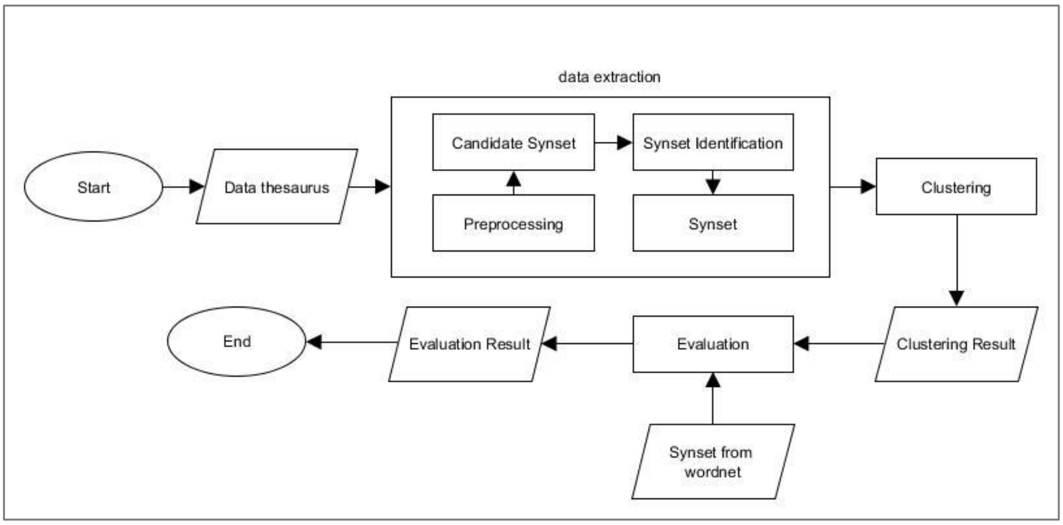


Fig. 1 System Overview

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.855, Copyright ©2020
Submitted : 14 Mei 2020, Revised : 30 Mei 2020, Accepted : 30 Mei 2020, Published : 4 Juni 2020

III. RESEARCH METHOD

A. System Overview

In this research, the system will be build using commutative methods and agglomerative clustering to get the final result in the form of English Synset. The system design that will be made can be seen in figure 1.

Figure 1 shows the system process that will be made. In the initial stage of the system, the data is obtained from an English thesaurus then the extraction process is carried out to obtain valid data. After that, the grouping was done using Agglomerative Clustering. After that, the evaluation process to get the value of recall and precision.

B. Data test

In table 1 there are 40 data from Oxford thesaurus to build Wordnet, the data is taken randomly and only nouns can be used as data. The following test data:

TABLE 1 Data Test

Word			
anthem	broker	eyebrow	knife
apparel	bulb	eyelash	leftover
ash	calendar	eyelid	moustache
autocrat	cargo	gamble	naturist
autograph	combustion	gate	robot
automobile	diocese	glasshouse	sausage
axe	dose	greenhouse	soldier
beetle	entrepreneur	habitat	spire
bishopric	equestrian	injection	steeple
bride	expert	innovator	valley

C. Commutative Method

This stage the data extraction process is carried out using the commutative method, at the process there are several stages that must be carried out. Using the Synset [automobile] example, the commutative process is as follows:

1) Determine the candidate word

This stage a Synset candidate search is performed on the dataset, for details such as the following explanation [11]:

a) First word
   Automobile : automobile, car, auto

b) Second word
   Car : automobile, motorcar, motorbike, machine

2) Check the relationship between words

This stage take the results of words from point one predetermined word candidates, the word is checked whether it has a link or relation between words or not. For example the word 'automobile', the word 'automobile' has the meaning of the word 'automobile, car, auto' for the meaning of the word to do another search that is incorporated into the meaning of 'automobile'. The word 'car' has the meaning 'automobile, car, motorcar, motor, machine' while the word 'auto' has no meaning. In this case it shows that the word 'automobile' with the word 'car' has a commutative relation and can be used to be a Synset.

3) Checking candidates for Synset

This stage, the process of checking the prospective Synset is done so that the subset of the other Synset is eliminated or removed to get maximum results

4) Final output Commutative Method

The final output is in the form of a Synset candidate that has been extracted using the commutative method, for example as follows:

TABLE 2 Synset Candidate

Word	Synonym set Candidate
Automobile	Automobile, car

D. Agglomerative Clustering

Agglomerative Clustering is one of the techniques in unsupervised machine learning. Agglomerative schemes are obtained from several data into a single node and then combined with several stages by checking the closest resemblance and becoming new data [16]. Agglomerative Clustering is a bottom-up approach [17], bottom-up means that each process starts from the cluster itself and the cluster pairs are merged which moves up like a hierarchy. This provides several advantages over the top-down method. this provides higher mutual information per cluster [18].

In agglomerative grouping, it starts from a partition of data where the sample is in a single cluster [19]. For the grouping process, several stages are carried out. Here are a few points raised:

1. Calculate the distance value to get the value of distance must get the value of the same word and unique word, the same word is obtained by comparing two Synset, these two Synset check the word similarity. to get a unique word to count the number of the second Synset words compared earlier. The difference in distance values can be seen when viewed:

$$\text{Distance Value} = \frac{\text{Similar Word}}{\text{Unique Word}} \quad (5)$$

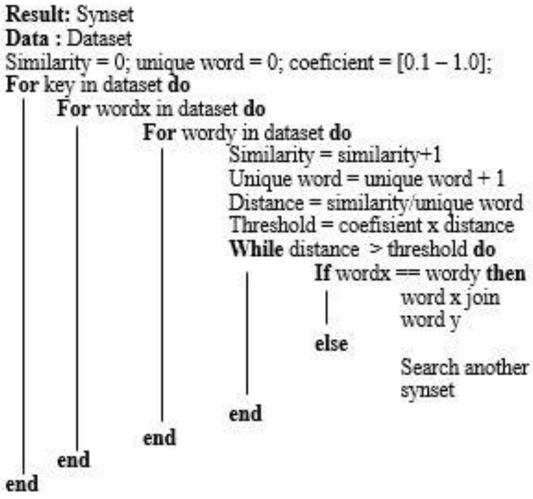
2. Calculate the threshold in this section calculates the threshold based on the coefficient and the first maximum distance value, the coefficient is obtained from some of the author's experiments. for this study the authors provide a coefficient value [0.1 - 1.0] to see the final result with the maximum number of Synset. The threshold of the equation can be seen in the following equation:

$$\text{Threshold} = \text{Coefficient x distance value} \quad (6)$$

3. The iteration process the grouping process will run if the distance value is greater than the threshold value, and the process will stop if the distance value is greater than the threshold value [20].

In the grouping, process is intended to group data based on the largest value and the largest distance value. For the algorithm grouping figure 2.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.855, Copyright ©2020
Submitted : 14 Mei 2020, Revised : 30 Mei 2020, Accepted : 30 Mei 2020, Published : 4 Juni 2020



```
Result: Synset
Data : Dataset
Similarity = 0; unique word = 0; coeficient = [0.1 – 1.0];
For key in dataset do
    For wordx in dataset do
        For wordy in dataset do
            Similarity = similarity+1
            Unique word = unique word + 1
            Distance = similarity/unique word
            Threshold = coefisient x distance
            While distance > threshold do
                If wordx == wordy then
                    word x join
                    word y
                else
                    Search another
                    synset
            end
        end
    end
end
```

Fig. 2 Agglomerative Clustering Algorithm

IV. RESULT AND DISCUSSION

A. Experiment Result

This stage, the authors conducted an experiment by changing the coefficient values from 0.1 to 1.0. the coefficient value is very influential with the amount of Synset produced. The results of the experiment are explained in table 3.

TABLE 3 Experiment Result

Coefficient	Maximum Similarity	Maximum Distance	Loop	Synset Result
0.1	1	0.25	22	31
0.2	1	0.25	22	31
0.3	1	0.25	19	32
0.4	1	0.33	17	33
0.5	1	0.33	17	33
0.6	2	0.5	15	34
0.7	2	0.5	15	34
0.8	2	0.5	15	34
0.9	2	0.5	15	34
1.0	2	0.5	15	34

In table 3 shows the greater the coefficient value, the greater the number of Synset produced, this is influenced by the amount of closeness between words that affect the final result in the form of grouping Synset. This is supported by the decreasing number of loops if the greater the coefficient value. but there is a maximum coefficient limit that affects the final number of Synset, in table 3 can be seen the coefficient values 0.6 - 1.0 get the same value, this shows there is a maximum value in the grouping process, this can also be influenced by the amount of data.

B. Evaluation Result

After doing the commutative process and grouping using Agglomerative Clustering which produces a synonym set that has been grouped. The results are evaluated using the F-Measure, to get the F-Measure value first to find the Precision and Recall values. The values of Precision, Recall, and F-Measure can be seen in table 4.

TABLE 4 Evaluation Result

Precission	Recall	F-measure
75.56	77.27	76.4

Based on the analysis conducted using the Commutative and Agglomerative Clustering methods, the test results can be seen in table 4. In the table shows a Precision value of 75.56 percent, a Recall value of 77.27 percesnt and F-Measure value of 76.4 percent. This value can be said to be not too large for 40 data, it is due to several factors namely the final results of the Synonym set produced by the Commutative and Agglomerative Clustering methods have not gotten the maximum results as the Synonym set made by experts. For example, the word [entrepreneur] which has several synonyms, after the word is processed commutatively and Clustering gets the Synonym set [entrepreneur, businessman, tycoon] while the Synonym set that can be from the expert is [entrepreneur, enterpriser] so that it greatly influences the test value.

Another factor is the difference in the number of results of the Synset produced by the system and the Synonym set of experts, such as the word [injection] only has one Synonym set of system results namely [injection]. While the Synonym set of experts has three Synonym sets namely [injection] [injection, injectant] [injection, shot] This leads to inequality between the Synonym set of the system and the Synonym set of experts that affect the final value of the F-Measure.

TABLE 5 Comparison Table

Synset output Commutative	Synset output Clustering	Validation
automobile, car	automobile, car	automobile, car, auto, machine, motorcar
apparel	apparel	apparel, wearing apparel, dress, clothes

In table 5 shows the comparison between the results of the Commutative, the results of Clustering, and validation. This shows a significant difference in word similarity, it results in a reduction in the value of the percentage of evaluation. In the grouping process after the commutative process, there are no differences that can be concluded that the use of Agglomerative Clustering is appropriate but in this study, the final results have not been maximized because the dataset obtained from the English thesaurus there are visible differences with the validation data. The following mapping clustering in the form of dendogram:

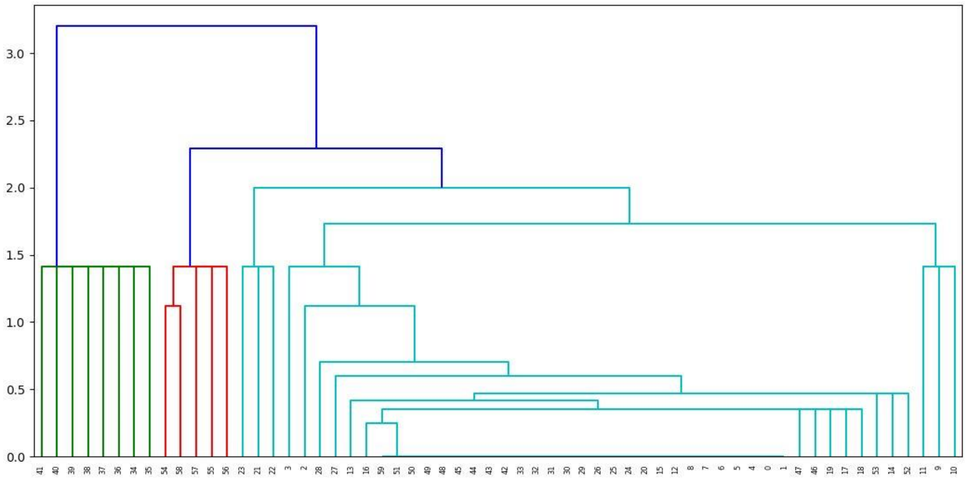


Fig. 3 Dendogram

V. CONCLUSIONS AND SUGESSTION

A. Conclusions

Wordnet development uses a combination of commutative methods and agglomerative clustering. in the commutative part of the test data obtained from the thesaurus the extraction process is carried out and then gets the final result of the Synset candidate in the process. Then proceed with the clustering process or grouping, in this process the writer uses agglomerative clustering which is a bottom-up approach. The grouping process is based on repetition if the distance value is greater than the threshold value.

In the commutative process of data equations as a comparison between the first word and the second word to produce prospective data has a weakness in the word reduction process, this is because the word thesaurus has so many meanings in one word and also causes the data extraction process to be a little longer with only 40 data only. How is a thesaurus extracted? According to the authors will be very time-consuming in the extraction process. This also causes the clustering process to be not optimal because the output of the extraction process has data that is far from the comparative data obtained from Princeton. In this study the authors get the best value at the coefficient value 0.6, this will change with increasing test data. In this study gave an evaluation score of f1 of 76.4%, the value was sufficient but not too large for 40 test data This is certainly very influential on the selection of data in the thesaurus based on the amount of meaning of the test data.

B. Sugesstions

Weaknesses in this study are the data used are not optimal so the final results in this study are also not maximal. In the process of forming the dataset of suggestions from the author when creating a dataset, input a lot of datasets so that maximum results can be assessed with a lot of data. Suggestions from the authors also try other clustering methods with similar methods such as the commutative method, because the commutative method has not produced a good candidate for the synset. Suggestion writer tries to use a combination of Latent Semantic Analysis and Rock Clustering.

REFERENCES

[1] G. A. Miller, "Introduction to WordNet: An on-line lexical database," *International journal of lexicography,* vol. 3, pp. 235-244, 1990.

[2] M. I. Pribadi, "Pendeteksian Relasi Antar Makna Pada Wordnet Bahasa Indonesia," *Universitas Komputer Indonesia,* 2017.

[3] F. R. d. D. P. D. Zamzami, "Apliasi wordne indonesia berdasarkan kamus thesaurus bahasa indonesia berdasarkan kamus thesaurus bahasa indonesia menggunakan algoritma rule based text parsing," *Seminar Informatika Aplikatif Polinema,* 2016.

[4] M. A. B. d. K. M. Lhaksamana, "Pembangunan Synonym Set untuk WordNet Bahasa Indonesia dengan Menggunakan Metode Komutatif,," *Indonesia Journal on Computing (Indo-JC),* vol. 4, pp. 147-156, 2019.

[5] G. A. Pradnyana, "Perancangan dan Implementasi Automated Document Integration dengan Menggunakan Algoritma Complete Linkage Agglomerative Hierarchical Clustering," Jurnal Ilmu Komputer, vol. 5, no. 2, 2012.," *Jurnal Ilmu Komputer,* vol. 5, 2012.

[6] L. D. Anggaraini, "Analisis Pembangunan Word Sense pada WordNet Bahasa Indonesia Menggunakan Metode Hierarchical Clustering," Telkom University, Bandung, 2019.

[7] D. P. Nasional, Kamus besar bahasa Indonesia, 2008.

[8] w. a. c. l. suprafti, "KONSTRUKSI TESAURUS NASKAH KUNO DENGAN PENDEKATAN LITERARY DAN USER WARRANT," *jurnla ilmu perpustakaan,* vol. 7, pp. 221-230, 2018.

[9] A. Fadliana, "Penerapan metode Agglomerative Hierarchical Clustering untuk klasifikasi Kabupaten/Kota di Provinsi Jawa Timur berdasarkan kualitas pelayanan keluarga berencana," *Universitas Islam Negeri Maulana Malik Ibrahim,* 2015.

[10] P. Etzioni, "Adaptive Web Sites: Conceptual Cluster Mining," *IJCAI,* p. 6, 1997.

[11] M. A. B. I. P. P. Ananda, "Pembangunan Synsets untuk WordNet

Bahasa Indonesia dengan Metode Komutatif," *eProceedings of Engineering,* vol. 5, 2018.

[12] D. J. Restina, "Pembangunan Synonym Set untuk WordNet Bahasa Indonesia dengan Menggunakan Metode Komutati," *Indo-JC,* vol. 4, no. 2, 2019.

[13] a. sabrina, "KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA MENGGUNAKAN MULTINOMIAL NA{\"I}VE BAYES," *publikasi tugas akhir s-1 PSTI FT-UNRAM,* 2020.

[14] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

[15] R. a. I. P. P. A. Cahyani, "Analisis Sentimen terhadap Ulasan Hotel menggunakan Boosting Weighted Extreme Learning Machine," *Jurnal Pengembangan Teknologi Indormasi dan Ilmu Komputer,* p. 2548, 2019.

[16] B. Sasirekha K, "Agglomerative hierarchical clustering algorithm-a," *International Journal of Scientific and Research Publications,* vol. 83, p. 83, 2013.

[17] D. Mullner, "Modern hierarchical, agglomerative clustering algorithms," 2011.

[18] A. Saputra, "Building synsets for Indonesian Wordnet with monolingual lexical resources," *International Conference on Asian Language Processing,* pp. 297-300, 2010.

[19] rahayu, "Analisis Dan Implementasi Algoritma Agglomerative Hierarchical Clustering Untuk Deteksi Komunitas Pada Media Sosial Facebook," *eProceedings of Engineering,* vol. 5, 2018.

[20] S. T. a. F. M. Z. Cristina Bosco, "Somewhere between Valency Frames and Synsets. Comparing Latin Vallex and Latin WordNet," in *Proceesings of the Second Italian Conference on Computational Linguitics,* trento, Accademia University Press, 2015.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.855, Copyright ©2020
Submitted : 14 Mei 2020, Revised : 30 Mei 2020, Accepted : 30 Mei 2020, Published : 4 Juni 2020