**Development of Synonym sets for English Wordnet using the Commutative and Agglomerative Clustering methods**

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]widiwdu@telkomuniversity.ac.id

**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 Synonym of word, the English language thesaurus becomes the reference data for taking synonym data. To get a Synonym set it is said to be commutative if first word has a synonym of second word and second word also has a synonym of the first word . After the formation of the Synonym set the grouping process is done using Agglomerative Clustering.**

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

# 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].

 Agglomerative clustering has the advantage that it does not need to determine the number of clusters and does not take into account initial centroids (tang 2006). 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.

# Literature Review

## 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 being 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].

* 1. Synonym set

Synonym set (Synsets) are lists of terms or collocations that have the same meaning and in certain contexts the uses are interchangeable. Also noted in the Synset are the pointers to other Synset used to describe the relation between Synset [6]. Synset data sequence information is obtained from Wordnet extraction results[7]

* 1. 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 [8]. According to its function, a thesaurus is a monitoring tool (checking tool or selection tool) of words used to translate human language used by indexers or users in documents, into more system languages limited [9].

* 1. 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 [10]. 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 [11], to get distance values ​​using the following equation:

$Distance Value=\frac{Similar Words}{Unique Words}$ (1)

* 1. 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. 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.

* 1. 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.

* 1. 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 [12]. After the Recall and Precision values ​​are determined, the values ​​are processed by calculating the F-Measure value. F-measure is a relevant goal in any machine learning scenario [13], In the f-Measure calculation precision and recall are thought of as some degree of correction and completeness of results [14]. Calculation of F-Measure values ​​can use the following equation:

$Precission=\frac{(TP)}{(TP+FP)}$ (2)

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

$FMeasure=\frac{2\*Precission\* Recall}{Precission+Recall}$ (4)

# Research Method

* 1. System Overview

 In this research, hte system will be build using commutative methods and agglomerative clustering to get the final result in the form of English Synset. The following is a system design that will be made:



Fig. 1 System Overview

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.

* 1. 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:

TABEL 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 |

* 1. 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[15]:

* First word
	+ Automobile : automobile, car, auto
* Second word
	+ Car : automobile, motorcar, motorbike, machine
1. 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.

1. 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 because more and more words that have no resemblance to the gold standard will result in a low final evaluation value.

1. 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:

TABEL 2 Sysnset Candidate

|  |  |
| --- | --- |
| **Word** | **Synset candidate** |
| Automobile | Automobile,car |

* 1. 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:

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

1. 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:

$Threshold=Coefficient x distance value$ (6)

1. 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 algorithm in 1

**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**

Algoritma 1 Agglomerative Clustering

# Result and discussion

* 1. 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.

TABEL 3 Experiment Result

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | Maximum Similarity | MaximumDistance | 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.

* 1. 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.

TABEL 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 percent 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] [injeciton, 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.

1. 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.

##### References

[1] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, dan K. J. Miller, “Introduction to WordNet: An on-line lexical database,” *International journal of lexicography*, vol. 3, no. 4, hlm. 235–244, 1990.

[2] M. Iffandi Pribadi, “Pendeteksian Relasi Antar Makna Pada Wordnet Bahasa Indonesia,” Universitas Komputer Indonesia, 2017.

[3] D. Zamzami, F. Rahutomo, dan D. Puspitasari, “Aplikasi wordnet indonesia berdasarkan kamus thesaurus bahasa indonesia menggunakan algoritma rule based text parsing,” dalam *Seminar Informatika Aplikatif Polinema*, 2016.

[4] M. A. Bijaksana dan K. M. Lhaksamana, “Pembangunan Synonym Set untuk WordNet Bahasa Indonesia dengan Menggunakan Metode Komutatif,” *Indonesia Journal on Computing (Indo-JC)*, vol. 4, no. 2, hlm. 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.

[6] S. A. Nugroho, “Query Expansion Dengan Menggabungkan Metode Ruang Vektor Dan Wordnet Pada Sistem Information Retrieval,” *Jurnal Informatika*, vol. 5, no. 1, 2011.

[7] D. M. Bikel, “Automatic WordNet Mapping Using Word Sense Disambiguation,” dalam *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, China, Okt 2000, hlm. 142–147, doi: 10.3115/1117794.1117812.

[8] D. P. Nasional, “Kamus besar bahasa Indonesia,” 2008.

[9] K. Tambunan, “Tesaurus Bioteknologi: Sebagai Alat Bantu Pengindeksan Dokumen,” *JURNAL DOKUMENTASI DAN INFORMASI*, vol. 33, no. 2, 2012.

[10] 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.

[11] M. Perkowitz dan O. Etzioni, “Adaptive Web Sites: Conceptual Cluster Mining,” hlm. 6.

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

[13] D. R. Musicant, V. Kumar, dan A. Ozgur, “Optimizing F-Measure with Support Vector Machines,” hlm. 5.

[14] J. Euzenat, “Semantic Precision and Recall for Ontology Alignment Evaluation,” *A C*, hlm. 6.

[15] I. P. P. Ananda, M. A. Bijaksana, dan I. Asror, “Pembangunan Synsets untuk WordNet Bahasa Indonesia dengan Metode Komutatif,” *eProceedings of Engineering*, vol. 5, no. 3, 2018.

[16] D. Müllner, “Modern hierarchical, agglomerative clustering algorithms,” *arXiv:1109.2378 [cs, stat]*, Sep 2011, Diakses: Mei 12, 2020. [Daring]. Tersedia pada: http://arxiv.org/abs/1109.2378.

[17] K. Sasirekha dan P. Baby, “Agglomerative hierarchical clustering algorithm-a,” *International Journal of Scientific and Research Publications*, vol. 83, hlm. 83, 2013.

[18] N. Slonim dan N. Tishby, “Agglomerative Information Bottleneck,” dalam *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, dan K. Müller, Ed. MIT Press, 2000, hlm. 617–623.

[19] S. D. Kamvar, D. Klein, dan C. D. Manning, “Interpreting and Extending Classical Agglomerative Clustering Algorithms using a Model-Based Approach,” hlm. 8.

[20] A. Saputra, “Building synsets for Indonesian Wordnet with monolingual lexical resources,” dalam *2010 International Conference on Asian Language Processing*, 2010, hlm. 297–300.