OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

IND. JOURNAL ON COMPUTING
VOL. 4, ISSUE. 2, SEPT 2019. PP. 137-146
DOI:10.21108/INDOJC.2019.4.2.332

Analysis of the Commutative Method Approach on English Thesaurus for Developing Synonym Sets

Arini Rohmawati [#1], Moch. Arif Bijaksana [*2], Kemas Muslim Lhaksmana [#3]

[#] School of Computing, Telkom University Bandung
Indonesia
[1] aringituloh@student.telkomuniversity.ac.id
[3] arifbijaksana@telkomuniversity.ac.id
[2] kemasmuslim@telkomuniversity.ac.id

Abstract

WordNet is a lexical database for languages, the difference between WordNet and dictionaries in general is that WordNet focuses on the synonyms. The main unit of WordNet is synonym set (synset), synset is a set of one or more words that have the same meaning and certainly can be replaced in certain contexts. Synset is a very important element in implementing WordNet. In this paper, an analysis of the synonym extraction process is carried out by using commutative approach, the data test obtained from the *Oxford Paperback Thesaurus* by taking 50 word entries. Commutative method has similar characters with synonym set, synonym set can replace each other in certain contexts. The data test extraction process is carried out until the performance measurement evaluation process using *F1Score*. The system generates synonym sets that matched with the manual extraction, the result of *F1Score* between the program and Princeton synonym sets are worth 10%. The *F1Score* is very low because of the data test sources factor, the gap in the year of publishing sources is quite far and English is always evolving to adjust time.

Keywords: commutative, thesaurus, wordnet, synonym.

I. INTRODUCTION

DICTIONARY originally came from the Latin word dictionarius which means ”a manual or book of words.” A dictionary is most commonly used to look up the definitions of particular words, other information such as etymology and usage guidelines, and so on [1]. In a dictionary there is also information about word usage information, the origin of words, decapitation, and various other information. Each word can have one or more meanings, and several words that have different writing and can have the same meaning. For example, the word imitation has the same meaning as artificial, fake, artificial, and synthetic [2].

WordNet is a study in the computational field of linguistics, and *freeware* which can be downloaded for free. WordNet is a database of English dictionaries that have been developed by Princeton University located in New Jersey, United States [3]. The difference between WordNet and language dictionaries in general is that language dictionaries generally focus on words, which WordNet is a lexical reference system that contains word information, word classes, and definitions of all sets of words contained in a language. WordNet focuses on synonyms of the word, it has been developed since 1985 with the help of English experts lexicographer from Princeton at that time made manually by experts. This is considered

RECEIVED ON AUGUST 2019. ACCEPTED ON SEPT 2019

to require longer time and higher costs, so the automatic and semi-automatic methods are used as a method of developing WordNet.

WordNet users can generally be divided into two groups, they are common users and developers. Common users are people who want to use WordNet to add their knowledge or search for information. Another group that often use WordNet are the researchers and developer which move mainly in the field of computational linguistic and NLP (Natural Language Processing).

Synonym set or synset is the main unit used in WordNet, it is a set of one or more words that have the same meaning and of course can replace each other in certain contexts. The synonym character that can replace each other in certain contexts is very familiar with the commutative law. There are several processes in developing synonym set such as extraction processes and clustering processes.

In this paper, the analysis focuses on extraction process by implementing the commutative method and an evaluation process using *F1Score* to measuring the performance on the system which will be compared with Princeton WordNet. Departing from an existing dissertation regarding the process of synonym set extraction by implementing commutative method [4].

II. Literature Review

WordNet has played an important role and has greatly contributed to text mining and web mining. The smallest unit in WordNet is not a word, but a synset or synonym set whose members are words that have the same meaning. The latest version of WordNet English from Princeton University has 117,659 synset [4].

Synonym set extraction implements a commutative method because synonym relations should be commutative, meaning that if a word $word_1$ has a synonym of $word_2$, then $word_1$ must be a synonym of $word_2$. Valid synonyms only occur in word pairs that have commutative relationships in the thesaurus. For a word in a thesaurus that does not have a commutative relations with other words, whether it has or does not have an item, then the two words do not have a synonymous relation or the synonym relation between the two words is invalid. [4].

Many studies related to WordNet or synonym set development. Building Synsets for Indonesian WordNet with Monolingual Lexical Resources by Gunawan and Andy Saputra is one of them. The paper presents an approach to build synsets for Indonesian WordNet semi-automatically using monolingual lexical resources available freely in Bahasa Indonesia. Explored an approach to build synsets for lexical database similar to Princeton WordNet. Monolingual lexical resources refer to Kamus Besar Bahasa Indoensia or KBBI (monolingual dictionary of Bahasa Indonesia) and Tesaurus Bahasa Indonesia (Indonesian thesaurus). Monolingual resources played an important role in synsets building, because it provide more accurate senses specifically for Bahasa. Besides, resources that have been used are produced by Bahasa Indonesia Language Center, which is a government institution that manages Bahasa Indonesia development [5], [6].

In a synonym set there are synonyms that can replace each other in certain contexts, it means one synonym and the other have commutative relations. To get a set of synonyms that have commutative relations, synonym extraction needs to be done. The commutative method is very suitable for synonym characters, and one of suitable method for obtaining synonym results is by implementing commutative methods using matrix tables because the matrix has a symmetrical character. Commutativity implies certain symmetries in the structure of products of commutative matrices [7], [8].

III. Research Method

A. System Overview

In this section, the process stages needed to get the appropriate results, developing synonym set that will be reevaluated with *F1-Score* will be explained. There are preprocessing the data set from the Oxford Paperback Thesaurus, synonym set extraction implementing commutative method and the final step is

evaluation the system performance using *F1-Score*. The following in Figure 1 is an overview of the stages on the system:

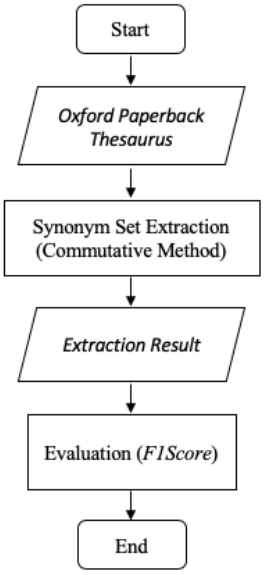

Figure 1: Flowchart of the system

B. WordNet

WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. The difference between WordNet and language dictionaries in general is that language dictionaries focus on the meaning of word while WordNet focuses on the synonyms of the word. The smallest unit of WordNet is synonym set (synset), synset members are words with the same concept [4] , [9].

The first WordNet developed by Princeton University is one of the most popular and the most widely used lexical database in building WordNet in several countries, the latest version of Princeton WordNet has 117,000 sets of synonyms. [10]

C. Synonym Set

Synonym set is a set consisting of one or more words that have a similarity of meaning or synonym. Synonym set is an important part of WordNet in any language. Each member of the set can replace each other in most uses of the word in contexts without changing the sense or meaning of the sentence [4]. Synonym sets are grouped into four word classes: nouns, verbs, adjectives, and adverbs that have different concepts. There is also a single synonym set because the word does not have synonyms or does not have interconnected words so that the word becomes a single word [10]. In table I shows the synonym set of the word **`cabinet`**, it has 4 sense of synonyms:

Table I: The example of synonym sets

Synonym set of **`Cabinet`**	
(n) [cabinet]	(a piece of furniture resembling a cupboard with doors and shelves and drawers; for storage or display)
(n) [cabinet]	(persons appointed by a head of state to head executive departments of government and act as official advisers)
(n) [cabinet, locker, storage locker]	(a storage compartment for clothes and valuables; usually it has a lock)
(n) [cabinet, console]	(housing for electronic instruments, as radio or television)

D. Thesaurus

Thesaurus dictionary is a dictionary focusing on one word that has several meanings and sub-language. The difference between dictionaries in general is one of indexing, the dictionary is organised alphabetically and thesaurus by word grouping [11] [1]. The thesaurus focuses more on synonyms than words and even adds antonyms to words. Thesauruses produced (manually) to support information retrieval systems. Characteristics of the thesaurus are as follows:

1) Reference source that contains information on a word arranged alphabetically (A-Z) or based on a theme
2) The information in thesaurus are class words, meanings, usage examples and synonyms or antonyms of the word
3) Thesaurus gives information on differences between one word and another word

A good thesaurus provides additional information to distinguish each word, for example; die has similar meaning with pass on, expire, death and so on. The example of Oxford Paperback Thesaurus entry as follows below:

Attitude **n**. view, viewpoint, outlook, perspective, stance, standpoint, position, opinion, ideas, conviction, feeling, thinking.

E. Commutative Method

Commutative law or commutative method can be interpreted that we can exchange numbers that will be counted in multiplication and addition and the answer generated will remain the same as in the illustration examples are in formula (1) and (2) below:

$$a + b = b + a \tag{1}$$

$$(a + b) + c = a + (b + c) \tag{2}$$

In this paper commutative method is implemented, that mean in each word can exchange and not change the values contained therein. Commutative law itself has properties that are in accordance with the nature of synonym set, where the members of the set in it can replace each other. For words in thesaurus that do not have commutative relations in other words, whether they have or do not have an item, then the two words do not have a synonym relation, both words are invalid [4].

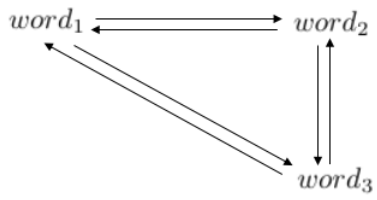

Figure 2: Commutative Method Illustration

In Figure 2 there is a synonym set which has 3 members, there are $word_1$, $word_2$ and $word_3$ which are connected with arrows means each other word has a commutative relations. $word_1$ is a synonym of $word_2$ and $word_3$, $word_2$ is synonyms of $word_1$ and so on.

Valid synonyms only occur in word pairs that have commutative relationships in the Thesaurus. For a word in a Thesaurus that does not have a commutative relations with another word, whether it has or does not have an item, then the two words do not have a synonymous relations or the synonym relation of the two words is invalid [4].

F. Evaluation

1) Recall: Calculation of *Recall* is the success rate of the system in rediscovering information, used to measure the ratio of the number of predictions that are correctly expected to total predictions. The number of correctly classified positive examples divided by the number of positive examples in the data [12], generally formulated as follows:

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

Where,
TP = true positive or the amount of data that is predicted to be positive and it is true, positive.
FN = false negative or the amount of data predicted to be negative but it is positive.

2) Precision: The calculation of *Precision* is the level of accuracy between the information requested by the used and the answers given by the system. The number of correctly classified positive examples divided by the number of examples labeled by the system as positive [12], generally formulated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Where,
TP = true positive or the amount of data that is predicted to be positive and it is true positive
FP = false positive or amount of data that is predicted to be positive but it is negative

G. F1-Score

F1-Score is used to measuring the accuracy of the system. F1-Score is a combination of precision and recall, taking the weighted harmonic average of precision and recall [13] [12], with the following formulas:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

IV. RESULTS AND DISCUSSION

A. Data

Oxford Paperback Thesaurus Fourth Edition 2012 is the source for getting data tests in developing synonym set [14]. It provides thousands of synonyms for English, it has been taken manually as many as 50 words are used as data test. The following below in table II are 50 word entries data test:

Table II: 50 Entries Oxford Paperback Thesaurus Data Test

Oxford Paperbak Thesaurus Data Test				
Attitude	Apartment	Achievement	Anthem	Answer
Apparel	Bed	Blanket	Boat	Broker
Buyer	Cabinet	Cafe	Calendar	Canteen
Cargo	Carnival	Chair	City	Clothes
Coast	Coat	Diplomat	Discoverer	Door
Dose	Entrepreneur	Expert	Fur	Future
Gamble	Gate	Habitat	Harm	Hardware
Impact	Injection	Innovator	Leftover	Logo
Mail	Medicine	Nutrition	Purse	Slogan
Soldier	Toy	Trait	Tummy	Valley

B. Synonym Set Extraction (Commutative Method)

There are several processes carried out at the extraction process. The following below in Figure 3 is the flowchart of the extraction process:

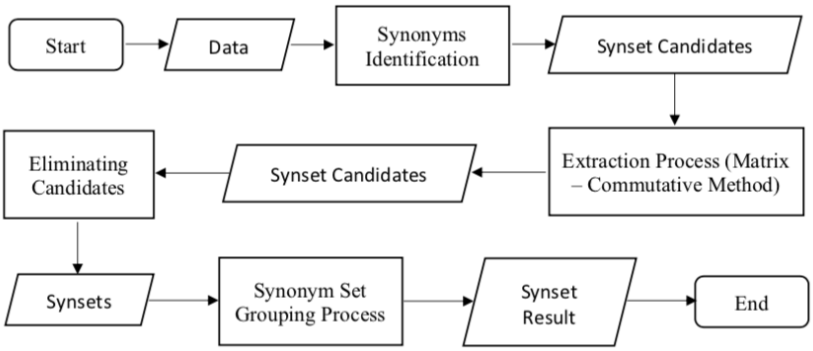

Figure 3: Flowchart of the Synonym Set Extraction

The data set that has been prepared will then be processed by implementing the commutative method. The synonym set extraction process at this stage refers to the algorithm in the synonym set extraction section base on monolingual resources [4].

1) Synonyms Identification: At this stage the process of identifying the set of synonym set is carried out on the word that is being searched for. The example is the word **`buyer`** has a set of synonyms which are **`purchaser`**, **`costumer`**, **`consumer`**, **`shopper`**, and **`investor`**. The members of the synonym called candidates, then every candidates of the set is identified again in each word contained in the sense.

In table III shows the results of the identification of each candidate of the synonym set has been identified by each member. It shows which candidates have the word **`buyer`** in each synonyms and which ones do not have it.

Table III: Result of Identification of synonym set (**`buyer`**)

Synonyms of: **`buyer`**	
purchaser	**1** acquisition, investment, order, bargain. **2** grip, grasp, hold, foothold, toehold, anchorage, attachment, support
costumer	consumer, **buyer**, purchaser, patron, client
consumer	purchaser, **`buyer`**, customer, shopper
shopper	**`buyer`**, purchaser, customer, consumer, client, patron
investor	originator, creator, innovator.

2) Extraction Process Using Matrix by Implementing the Commutative Method: This stage is formed matrix by implementing the commutative method with the data test that has been obtained in the previous process, commutative relations are sought in each word. If the word has a commutative relations in other words then it given the **`true`** value and if it does not have a commutative relations, the word is given the **`false`** value.

Table IV: The Matrix of Each Member of **`Beast`**

	buyer	**purchaser**	**costumer**	**consumer**	**shopper**	**investor**
buyer	***true***	***true***	***true***	***true***	***true***	***true***
purchaser	*false*	***true***	*false*	*false*	*false*	*false*
costumer	***true***	*false*	**true**	***true***	*false*	*false*
consumer	***true***	*false*	***true***	***true***	***true***	*false*
shopper	***true***	*false*	***true***	***true***	***true***	*false*
investor	*false*	*false*	*false*	*false*	*false*	***true***

In table IV shows which synonyms have commutative relations with the word **buyer** and given **true** values. The following in Algorithm 1 below is the algorithm of matrix extraction:

Algorithm 1: Matrix extraction algorithms

```
Result: Synonym set candidate
initialization;
for synonym candidate in the search word do
    check synonyms of the synonym candidate;
    if search word in the synonyms of the synonym candidate then
        true;
    else
        false;
    end
end
```

3) Eliminating the Candidate of Synonym Set: From the results of extracting process using matrix by implementing the commutative method, the synonyms set that do not have commutative relation with the word **`buyer`** is given the **`false`** value and automatically eliminated from the candidate. For examples, the word **`purchaser`** and **`investor`** have **`false`** value in the matrix table because both words do not have a commutative relation with the word **`buyer`**, because there is no **`buyer`** in the synonyms of **`purchaser`** and **`investor`**. The result of the eliminating process are words **`costumer`**, **`consumer`** and **`shopper`**. The following in Algorithm 2 below is the algorithm of candidate elimination process:

Algorithm 2: Eliminating the candidate algorithm

```
Result: Eliminating candidates
initialization;
for synonym in candidate list do
    check synonym of the candidate in the search word;
    if true then
        add to the candidate;
    else
        delete;
    end
end
```

4) Synonym Set Grouping Process: The results of the extraction process using matrix table by implementing the commutative method that have been done in the previous process are then grouped into sets of synonyms based on the value of **`true`** that has been given in each of them. Each synonyms that have **`false`** value cannot be related to each other and grouped on different synonym set. The word **`shopper`** in entered into different synonym set with the word **`costumer`**, because the word **`costumer`** does not have a commutative relation with the word **`shopper`**. The following is the result of synonym set grouping process for the word **`buyer`**.

buyer	[buyer, customer, consumer]
	[buyer, consumer, shopper]

The following in Algorithm 3 below is the algorithm of synonym set grouping process:

Algorithm 3: Synonym set grouping algorithms

```
Result: Synonym set
initialization;
for synonym in the candidate list do
    check commutative relation in each candidate in the candidate list;
    if true then
        group;
    else
        separate;
    end
end
```

C. Measuring Evaluation

The last step that has been done is validation by measuring the system performance using *F1Score*. With the provision of *recall* is the value of the predicated synonym set data and *precision* is the actual prediction data from the commutative results, the results of the system performance is as follows:

$Precision$	0.990
$Recall$	0.990
$F1Score$	0.990

After getting the results from the measuring evaluation of system performance by implementing *F1-Score*, comparison measurements were done with Princeton WordNet. The following in table V is 6 data test sample from 50 data test used from manual retrieval and Princeton WordNet as follows:

Table V: Data Test Comparison Between Manual Retrieval and Princeton WordNet

Words	Manual Data Test	Princeton WordNet Data Test
Attitude	[attitude, outlook, perspective, stance, standpoint] [attitude, perspective, stance, standpoint, position] [attitude, stance, standpoint, position, opinion]	[attitude, mental attitude] [position, posture, attitude] [attitude]
Apartment	[apartment, flat] [apartment, suite]	[apartment, suit]
Boat	[vessel, ship]	[boat] [gravy boat, gravy holder, sauceboat, boat]
Bed	[bed, berth] [bed, bottom]	[bed] [bed][bed, bottom] [seam, bed]
Cafe	[cafe, cafeteria, buffet	[cafe, coffeehouse, coffee shop, coffee bar]

The results of validation measurements from the program and Princeton WordNet is as follows:

$Precision$	0.108
$Recall$	0.106
$F1Score$	0.107

V. Conclusion

By implementing the commutative method for extracting synonym, the *F1Score* result in the program is 0.990 and it is almost 100% it means that implementing the commutative method is suitable for synonym extraction for developing synonym sets because commutative method basically has the same character as synonym sets, can replace each other. The synonym set results of commutative method in the program have been compared to the Princeton WordNet and produced an *F1Score* value of 0.107 it is so low because the data test that has been used is from *Oxford Paperback Thesaurus* is the fourth edition published in 2012, while the Princeton WordNet that has been used for validation is version 2.1 which was released in 2005. From the publication of data sources, shows the gap that is quite far in

the year the two sources are published. English continues to grow as the times develop, there are many words that are rarely used anymore and many new words appear.

Therefore, the suggestion for further research is to combine more than one source to compile test data, it could be the *Oxford Paperback Thesaurus* and *Oxford Dictionary* combine into one or the other sources. And use the latest sources to get more complete data.

REFERENCES

[1] Oxford. *The Oxford Thesaurus*. Springer Science & Business Media, 2013.
[2] Creandivity. *Penjelasan WordNet, Sistem Lexical Database*. creandivity, 2010.
[3] Encyclopedia of language and linguistics, second edition, oxford: Elsevier. 2005.
[4] Gunawan. Synonym sets extraction based gloss acquisition using supervised learning. *Disertation*, 2016.
[5] Andy Saputra et al. Building synsets for indonesian wordnet with monolingual lexical resources. 2010.
[6] KBBI Tim Penyusun. Kamus besar bahasa indonesia. *Balai Pustaka: Jakarta*, 2008.
[7] Tomaž Košir. On the structure of commutative matrices. *Linear algebra and its applications*, 187:163–182, 1993.
[8] Yukio Kobayashi. Commutative law for the multiplication of matrices as viewed in terms of hankel's principle. *European Journal of Pure and Applied Mathematics*, 7(4):405–411, 2014.
[9] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
[10] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.
[11] Adam Kilgarriff and Colin Yallop. What's in a thesaurus? In *LREC*, pages 1371–1379, 2000.
[12] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.
[13] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer, 2005.
[14] Maurice Waite. In *Oxford Paperback Thesaurus*. Clays Ltd, St Ives plc, 2012.