e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 4 No. 1, Juni, 2020, **Hal. 83-90**
DOI: 10.29408/edumatic.v4i1.2119

Developing Set of Word Senses of Vocabulary in Al-Qur'an

Neca Aqila[1], Mochammad Arif Bijaksana[2]
[1,2] Department of Informatics Engineering, Universitas Telkom
email: necaqila@student.telkomuniversity.ac.id[1], arifbijaksana@telkomuniversity.ac.id[2]

(Received: 12 Mei 2020/ Accepted: 1 Juni 2020 / Published Online: 20 Juni 2020)

Abstract

Al-Qur'an has become the guideline for all Muslims in the world, which makes many Muslims are eager to understand its contents. Nevertheless, Al-Qur'an consist of many words that have more than one meaning, which represent certain difficulties while understanding the meaning itself. As example, the word أَزْوَاجًا has two equivalents as it might be translated either "jodoh" ("mate") in Surah An-Nahl (16: 72: 6) and "golongan-golongan" ("groups") in Surah Al-Hijr (15: 88: 8). This case is known as a word sense, a word that has more than one meaning. This research aims to construct the word sense as a set of vocabulary, in order to simplify the vocabulary meaning in Al-Qur'an itself. The data set used in this research is nouns from Al-Qur'an which have been translated into Bahasa. In order to construct the set of word sense, this research grouped the words using Hierarchical Clustering method. The total set of the word senses found is 34 nouns, which contains diverse translation. The F measure from evaluation of this research resulted in an accuracy of 65.85%. The result was obtained based on the conformity between the results of the word senses set by the system and by the linguists. The outcome of this research is, accuracy is low, due to the type and the number data used is limited.

Keywords: Hierarchical Clustering, Qur'an, Word Senses Set

INTRODUCTION

Al-Qur'an is the holy book of Muslims which is a miracle revealed by Allah to the prophet Muhammad SAW. The Qur'an contains the words of Allah revealed in stages over a period of 22 years 2 months 22 days to Muhammad through the angel Gabriel (*Jibril*). The Qur'an consists of 30 parts (*juz*), 114 chapters (*surah*) and 6236 verses (Tim Redaksi, 2008). The Qur'an serves as the way of life for all Muslims in the world, so that many Muslims are eager to understand its contents. Nevertheless, the Qur'an contains many words that have more than one meaning, presenting certain difficulties in understanding. For example, the word أَزْوَاجًا has two equivalents as it might be translated either "*jodoh*" ("mate") as in Surah An-Nahl (16: 72: 6) or "*golongan-golongan*" ("groups") as in Surah Al-Hijr (15: 88: 8). Such case is known as word sense, a word or words that have more than one meaning(Samhith et al., 2016). For this reason, a word senses set was built which contains a collection of word senses from the Qur'an vocabulary, so that it can help facilitate understanding of vocabulary that has more than one meaning in the Qur'an. This study focuses on the vocabulary in the Qur'an which has more than one equivalent in Bahasa Indonesia due to the different context where it is used.

Computational linguistics, particularly involving Natural Language Processing (NLP) and Princeton WordNet (PWN)(Miller et al., 1991), is one of the most popular and most widely used lexical databases. It is used as the research materials in the construction of WordNet for Arabic language (Elkateb et al., 2006), set of synonyms (synset) for Bahasa Indonesia (Gunawan & Saputra, 2010), synset of Qur'an vocabulary using WordNet approach (Gupitasari, 2019), arabic corpus (Al-Thubaity et al., 2013) and so forth. But for the arabic corpus the Qur'an is still currently still lacking even though in the last few years there has been the development of arabic corpus which are free of access. These corpus can be said

not enough to encourage linguists to apply them to their corpus-based (Al-Thubaity et al., 2013). In previous research as well, namely a construction of wordnet for Arabic (Elkateb et al., 2006). Discussion about the set of word senses is only explained as a form of lexical ambiguity but was not included in the wordnet which was made as a set of word senses. Research on set of synonyms (synset) for Indonesian (Gunawan & Saputra, 2010) explains about making synonyms sets (synset) using clustering technique that can be used as a reference for this research using the same technique. However, this research does not collect synonyms from Indonesian language set but collecting the set of Arabic word senses. Another difference from this study with previous research (Gupitasari, 2019) is that this study collects Qur'an words which have different meanings but with the same lemma. Whereas the previous research discussed about vocabulary that has similar meaning.

Related to the previous studies mentioned above, this research was conducted to construct a set of word senses from Arabic vocabulary used in Al-Qur'an by utilizing lexical semantic similarity on PWN. Previous research about construction set of the synonyms from vocabulary in Al-Qur'an by using a WordNet approach has successfully constructed a lexical database focusing on a set of synonyms (Gupitasari, 2019). The research for a set of word senses from Al-Qur'an vocabulary to build Al-Qur'an thesaurus is still rarely found, while the thesaurus is not only contains synonyms, but also includes the word senses and it is deemed incomplete enough. Therefore this research becomes important to build a collection of word senses that can be used as a prototype to build Al-Qur'anic thesaurus. The construction of this set of word senses from Al-Qur'an vocabulary is using the same techniques as those used in the construction of synonym sets for Indonesian (Gunawan & Saputra, 2010)and the construction of synonym sets for Arabic vocabulary (Gupitasari, 2019), namely clustering techniques. The clustering technique that is used in this research is hirerarchical clustering. The primary reasons of choosing hierarchical clustering are because the result of the clusters can not be predicted before the clustering process is done (therefore, the partitional clustering is inappropriate for this need) and the simplicity concept of hierarchical clustering itself. Furthermore, the method of hierarchical clustering used is the agglomerative method. Then all members of the set of word senses that have been formed would be evaluated by comparing them with the gold standard, and the accuracy would be calculated using the F measure method.

The purpose of this research is to help linguists to collect the word senses set of vocabulary Quran, it is expected that the construction of this set of word senses can be used to facilitate understanding of the vocabulary in the Qur'an and the result of word senses set from this research can be used as a prototype to create a corpus of the Qur'an and tesaurus in order to increase the resources of Qur'an research.

METHOD

Constructing the set of word senses involved several processes in managing the data. The data of this research were the lemmas in the Qur'an (available at openburhan.net), and their translation in Bahasa Indonesia was written manually by the authors. The data were 28,253 data lines corresponding to the total number of words in the Qur'an from Surah Al-Fatihah to Surah Yunus. The process carried out in the system can be seen in Figure 1.

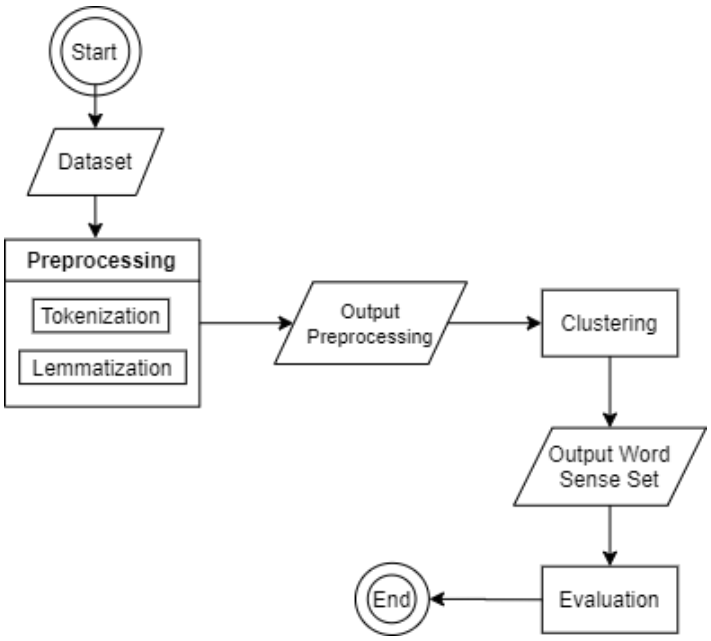

Figure 1 System Overview

1. *Preprocessing*

Data preprocessing is a process for preparing raw data before other processes are carried out (Mujilahwati, 2016). In general, data preprocessing is done by eliminating inappropriate data or transforming data into a form that is more easily processed by the system. Not all data from openburhan.net were used; therefore, data preprocessing was conducted to sort out the important data that would be used in this research. The data preprocessing in this research was conducted in two stages: Tokenization and Lemmatization. Which can be seen in the Table 1 above.

Table 1. Preprocessing

Data raw	Tokenization Process	Lemmatization Process
الْمِحْرَابَ	‘ الْ ‘ ; ‘ مِحْرَابَ ‘	مِحْرَابَ
يَدَيَّ	‘ يَدَ ‘ ; ‘ يَّ ‘	يَدَ
أَهْلَهَا	‘ أَهْلُ ‘ ; ‘ هَا ‘	أَهْلُ
أَمْرُنَا	‘ أَمْرُ‘ ; ‘ نَا ‘	أَمْرُ
الْمَسْكَنَةَ	‘ الْ ‘ ; ‘ مَسْكَنَةَ ‘	مَسْكَنَةَ
الدِّينَ	‘ ال ‘ ; ‘ دِّينَ ‘	دِّينَ
الْكِبَرُ	‘ الْ ‘ ; ‘ كِبَرُ ‘	كِبَرُ
الْغَيْظِ	‘ الْ ‘ ; ‘ غَيْظِ ‘	غَيْظِ
زَوْجَهَا	‘ زَوْجَ ‘ ; ‘ هَا ‘	زَوْجَ
الْأَنفسُ	‘ الْ ‘ ; ‘ أَنفُسُ ‘	أَنفُسُ

Tokenization is a process of splitting or cutting a text in the form of words, sentences, paragraphs or documents into certain tokens or parts (Nur, 2017). As this research only focuses on nouns, separate the lemma from words such as: conjuction, preprosition, adjective and pronouns (Attia, 2007). Lemma in the Qur’an is often added with personal pronouns, so a tokenization process is needed to separate the pronouns from the nouns. The word الْكِبَرُ for example, was carried out by dividing the noun into its first part and second part, resulting in كِبَرُ and الْ.

Lemmatization is the process of finding the base from (or lemma) of a word by considering its inflected from (Mubarak, 2019). After separated between tokens from the lemma, the lemmatization process is carried out which will take only the basic words.

2. *Clustering*

The data that have gone through preprocessing were processed using agglomerative hierarchical clustering. The purpose of this clustering was to group the data based on their similarity value and comparison with the threshold value (Putri, 2019). The similarity value was obtained from the same number of words between one set of sense and another set of sense (Pedersen et al., 2004) using the following equation.

$$PATH(S1, S2) = \frac{1}{path\_length(S1,S2)} \quad (1)$$

S1 was the first lemma and S2 was the second lemma that would be compared. Once the similarity value was found, it was compared with the threshold value. The threshold value was obtained from the following equation:

$$Threshold = coefisient.first\ maximum\ distancevalue \quad (2)$$

The evaluation carried out in this study was twice, first by comparing the results of the word senses set of the system with the results of the word senses set of the gold standard. After that the accuracy value is calculated to the value of F measure, recall and precision using the formula below:

a. *Recall:* Calculation of Recall is the success rate of the system in rediscovering infomration, used to measure the ratio of the number of predictions that are correctly expected to toal predictions(Rohmawati et al., 2019). Generally formulated as follows:

$$Recall = \frac{tp}{tp+fn} \quad (3)$$

Where,

TP = True Positive refers to words that match the results of both the system and the gold standard.

FN = False Negative refers to words that are not found in the system results but are found in the gold standard.

b. *Precision:* The calculation of Pecision is the level of accuracy between the information requested by used and the answers given by the system (Rohmawati et al., 2019). Generally formulated as follows:

$$Precision = \frac{tp}{tp+fp} \quad (4)$$

Where,

TP = True Positive refers to words that match the results of both the system and the gold standard.

FP (False Positive) refers to words that are found in the system results but are not found in the gold standard.

c. *F-measure* is used to measuring the accuracy of the system. F-measure is a combination of precision and recall, taking the weight harmonic average of precision and recall (Rohmawati et al., 2019). With teh following formulas:

$$F\ measure = 2\ \frac{Precision\ .\ Recall}{Precision\ +\ Recall} \quad (5)$$

RESULT AND DISCUSSION

Result

This is the result of preprocessing where the data will be used later for the next process, which is in clustering process.

Table 2. Result of Preprocessing

Result of Preprocessing
مِحْرَابَ
يَدَ
أَهْلُ
أَمْرُ
مَسْكَنَة
دِّينَ
كِبَرُ
غَيْظِ
زَوْجَ
أنفسُ

The result from the preprocessing data in Table 2 using clustering is stated in Table 3. From the total of 28,253 words inserted into the system the total number of sets of word senses found is 5621 nouns, with 5587 nouns containing the same translation and 34 nouns containing different translation. For the other word categories are ignored. Table 4 below shows some examples of the results of the sets of word senses obtained. Lemma آلِ have 2 different translation is "Keluarga" and "Kaum", and so on.

Table 3. Number of Word Sense Set Retrieved

Category	Word Sense Set with single member	Word Sense Set with multiple members	Total
Noun	5587	34	5621

Table 4. Some Results of Sets of Word Senses

Lemma entries	Set of Word Senses
آلِ	Keluarga , Kaum
سَيِّئَة	Keburukan ,Bencana
سُلْطَانَا	Kekuasaan , Keterangan
أَوْلِيَاءَ	Pemimpin , Penolong
قَرْنِ	Kurun , Generasi
فِتْنَةً	Fitnah , Cobaan
مُّحْكَمَاتٌ	Jelas, Tegas
يَدَ	Tanganku, Sebelumku
مَسْكَنَة	Kelemahan, Kemiskinan
حَسْرَةَ	Kerugian, Penyesalan
نِحْلَة	Ikhlas, Wajib
أَنفَسُ	Jiwa, Manusia
آيَةٍ	Ayat-ayat, Tanda-tanda
دِّينَ	Agama, Ketaatan

أُمَّةً	Umat, Waktu
أَمْرٌ	Perintah, Azab Kami
كَلِمَتُ	Kalimat, Hukuman
مِحْرَابَ	Mihrab, Mimbar

The evaluation in this research also by comparing the result of the word senses set by the system and word senses set by expert (gold standard), so that it can be used to calculate the next accuracy using f measure. The result are as show in table 5.

Table 5. Comparison of the Results of Set of Word Senses by Gold Standard

Lemma	Set of Word Senses by the system	Set of Word Senses by the expert (Gold Standard)
مِحْرَابَ	mihrab , mimbar	mihrab , mimbar
سُلْطَانًا	kekuasaan , keterangan	kekuasaan , alasan , bukti
أَوْلِيَاءَ	pemimpin ,penolong	pemimpin , pelindung , penguasa
دِّينَ	agama , ketaatan	agama , ketaatan , pembalasan
يَدَ	tanganku , sebelumku	tanganku , hadapanku
سَيِّئَة	keburukan , bencana	keburukan , kejahatan , jelek
قَرْنٍ	kurun , generasi	kurun , generasi, ummah
آيَةً	ayat-ayat , tanda-tanda	ayat-ayat , tanda-tanda , bukti-bukti
فِتْنَة	fitnah , cobaan	fitnah siksaan , cobaan
أَمْرُ	perintah , azab	perintah , urusan
آلِ	Keluarga, Kaum	Keluarga
مَسْكَنَة	Kelemahan, Kemiskinan	Kelemahan, Menenangkan, Tempatkan
حَسْرَةَ	Kerugian, Penyesalan	Kesedihan, Penyesalan
نِحْلَة	Ikhlas, Wajib	Ikhlas
أَمْرُنَا	Perintah, Azab Kami	Perintah, Urusan
أُمَّة	Umat, Waktu	Umat
أَهْلُ	Penduduknya, Pemiliknya	Penduduk, Ahli

The evaluation calculations in this research use the F measure, which aims at measuring the accuracy of the results of the clustering that has been done. The F measure uses Precision and Recall to do calculations. Table 6 presents the results of Precision, Recall, and F measure.

Table 6. Testing Result

Testing	Precision	Recall	F measure
Program Results	71%	61%	65.85%

Discussion

The result of preprocessing which listed in table 2 are used as input in clustering process and as an evidence in grouping a set of word sense.The results of this preprocessing are obtained through two other processes, which are tokenization process and lemmatization process as described above. The result of preprocessing are used in the next process, which is the clustering process, using formulas 1 and 2. In this research, the threshold values is 0.5. If the similarity value were equal to or greater than the threshold value, then the lemma would be grouped into the same cluster. On the other hand, if the similarity value were smaller than the threshold value, the lemma would be placed in a new cluster. This process would stop when the maximum distance value had been smaller than the threshold value. The

accumulation of word senses set according to Table 3. Some example of the generated set of word senses are in table 4, which is one lemma vocabulary that contains 2 different meaning. It can be seen in Table 4 above.

In Table 5 above there are some words that are not a set of word senses from lemma but are found in the system and some word senses are not found in the system. As example, in lemma مِحْرَابَ we found that word senses in the system which are "mihrab" and "mimbar" and by expert (gold standard) we found that the words "mihrab" and "mimbar" as well so the result of the word senses is TP (true positive). Whereas in lemma سُلْطَانًا word senses found in the system is "kekuasaan" and "keterangan", by the expert (gold standard) found "kekuasaan", "alasan" and "bukti". So "keterangan" is FP (false positive) and "alasan" and "bukti" is FN(false negative).

From the result of the above comparison, we can get variables to calculate the value of recall, precision and F measure. With the formula (3), (4) and (5). And the result can be seen in the table 6. In table 6 the precision value was 71%, recall value was 61% and the F measure was 65.85%. Precision is a large number of elements in the set of word sense that are formed where it is the correct element. Recall is the number of correct elements in the formed set of word sense. The number are obtained from comparison the result of word senses set by the system and from gold standard contained in table 5, which was explained earlier above. F measure is the result of the calculation of precision and recall that have been obtained previously.

In previous research (Gupitasari, 2019), the recall value was 81%, precision value was 86% and the F measure was 83%, factors that affect the value of recall and precision in this study (Gupitasari, 2019) is because there is a difference between the amount of synset by the system and the amount of synset by the Godl standard so it affects the accuracy of the relevant elements. Results from this previous study (Gupitasari, 2019) higher than the result of this study because the amount of data used was greater. The data used is all vocabulary in the Qur'an.

CONCLUSION

Based on the result of the research by using the agglomerative hierarchical clustering method, the total number of sets of word senses found is 5621 nouns, with 5587 nouns containing the same translation and 34 nouns containing different translation constructs. The use of F measure for evaluation results in an accuracy of 65.85% with the recall value is 61% and the precision value is 71%. This value is obtained from the word senses comparison result by the system and by experts (gold standard). The accuracy is low due to the limitations of the type and the number of data used, i.e. 28,253 nouns used in Al-Qur'an ranging from Surah Al-Fatihah to Surah Yunus .

REFERENCES

Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). New language resources for arabic: Corpus containing more than two million words and a corpus processing tool. *Proceedings - 2013 International Conference on Asian Language Processing, IALP 2013*, 67–70.

Attia, M. A. (2007). Arabic tokenization system. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources* (pp. 65-72).

Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). Building a Wordnet for Turkish. *Romanian Journal of Information Science and Technology*, *7*(1–2), 163–172.

Elkateb, S., Black, W., Vossen, P., Rodríguez, H., Pease, A., Alkhalifa, M., & Fellbaum, C. (2006). Building a WordNet for Arabic. *Proceedings of the 5th International*

Conference on Language Resources and Evaluation, LREC 2006, 29–34.

Gunawan, & Saputra, A. (2010). Building synsets for Indonesia Wordnet with monolingual lexical resources. In *2010 International Conference on Asian Language Processing* (pp.297-300).

Gupitasari, L. (2019). Pembangunan Synonym Set Kosa Kata Al-Quran dengan Pendekatan WordNet. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *6*(2), 163–170. https://doi.org/10.35957/jatisi.v6i2.188

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography,* 3(4), 235-244.

Mubarak, H. (2018). Build fast and accurate lemmatization for Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (*LREC 2018),* 1128–1132.

Mujilahwati, S. (2016). Pre-Processing Text Mining Pada Data Twitter. *Seminar Nasional Teknologi Informasi Dan Komunikasi*,(Sentika), 2089–9815.

Nur, I. (2017). Analisis dan Implementasi Tokenisasi Bahasa Arab pada Al-Quran. *Skripsi.* Program Sarjana. Universitas Telkom.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet : Similarity - Measuring the Relatedness of Concepts Measures of Relatedness. In *AAAI Conference, (4), pp.* 25–29.

Putri, K. N. (2019). Clustering Ekstraksi Synonym Set Bahasa Indonesia Menggunakan Agglomerative Hierarchical Clustering. *Skripsi.* Progam Sarjana. Universitas Telkom.

Rohmawati, A., Bijaksana, M. A., & Lhaksmana, K. M. (2019). Analysis of the Commutative Method Approach on English Thesaurus for Developing Synonym Sets. *Indonesia Journal on Computing (Indo-JC)*, *4*(2), 137–146.

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

Tim Redaksi. (2008). *Pengertian Al-quran*. Wikipedia Ensiklopedia Al-Quran. https://id.wikipedia.org/wiki/Al-Qur%27an