Building Related Words in Indonesian and English Translation of Al-Qur'an Vocabulary Based on Distributional Similarity

Rahmad Geri Kurniawan
Bachelor of Informatics
School of Computing
Telkom University
Bandung, West Java
gerikurniawan@student.telkomuniversity.ac.id

Moch Arif Bijaksana
Bachelor of Informatics
School of Computing
Telkom University
Bandung, West Java
arifbijaksana@telkomuniversity.ac.id

Abstract*—The Qur'an is the Muslim holy book as the primary source of knowledge and guidance, consisting of 114 surahs, 30 juz, and has approximately 6200 verses in it. Searching for connections or similarities between words in the Qur'an takes a long time to find and summarize them. There is a need for a dictionary, encyclopedia, or thesaurus of the Al-Qur'an vocabulary, which contains each word entry related to other words. This study discusses the interrelations and semantic similarities between words in the Qur'an, which aims to help in searching between related words in them. The approach taken is a distributional similarity which is an important part of word embedding. Measurement of word relevance is measured by semantic similarity which is one of the lessons learned in Natural Language Processing (NLP). Semantic similarity measures the closeness of word vectors using cosine similarity. The process of changing words in vector form uses the FastText algorithm which is a development of the Word2vec algorithm. The dataset used is the translation of the word Al-Qur'an in English and Indonesian. The word becomes an input in the system and then produces a score that represents the interrelationship between words. Evaluation of system output results using the Pearson correlation method involving the gold standard. Evaluation of the use of the FastText algorithm produces a correlation value of 0.3398 for Indonesian translation corpus and 0.2326 for English translation corpus.

Keywords*— Quran, semantic similarity, Word embedding, FastText, Pearson correlation

INTRODUCTION

The Qur'an is the holy book in Islam, which was come as the primary source of knowledge, law, wisdom, and guidance for Muslims. The Qur'an consists of 114 surahs, 30 juz, and 6217 verses according to the history of Abl Medina, 6210 verses according to al-Dani's history, or 6214 verses according to Warsy's history [1]. There is a lot of information in the Qur'an that there are words with related meanings scattered about it. One way to understand the Qur'an is to try to explain the content of the verses of the Qur'an, from various aspects *of* paying attention to the sequence of the verses of the Qur'an, as stated in it [2]. Looking for similarities and linkages of words is also needed to help explain the contents of the Qur'anic verses.

Semantic similarities and similarities are related to one of the areas of discussion on Natural Language Processing (NLP), namely semantic similarity. This field discusses the measurement of the similarity of two words represented by similarities between related concepts in it. The idea of semantic similarity is to identify concepts that have the same 'characteristics'. Semantic similarity is understood as the level of taxonomic closeness between concepts (or terms, words). In other words, semantic similarity states how closely two concepts (or terms, words) are taxonomic, because they share several aspects of their meaning. Technically, the similarity measures assess numerical scores that measure this closeness as a function of the semantic evidence observed in one or several sources of knowledge [3]. In its application, for example of input systems such as the first word *"paradise"* and the input of the second word *"hereafter"* will produce a high output similarity value. As humans can be interpreted, those words have the meaning of a place of life after world life. Until now, research on semantic similarity continues to be carried out with various methods, some of which are Word2vec, Global Vector, and Support Vector Machine (SVM).

In previous studies related to distributional similarity, measurements were made of the interrelationship of words in Arabic, using a vector-based approach. The system built on this research produces a set of words that have a relationship with other words using the Word2vec model. Evaluation in the study was carried out by calculating precision based on the corrections made by linguists from the resulting system output [4]. Word2vec known ignoring morphology, these methods cannot create word vectors for new words that do not appear in the training data. Because morphological features of words are ignored, new word vectors cannot be obtained by comparing them with morphologically similar words [5].

In this study, a system was built to calculate the semantic similarity value of two input words. We use the distributional similarity approach to capture the similarity of semantic words and make groups of words that are similar.

DOI: https://doi.org/10/25047/jtit.v7i1.135

This research uses the Al-Qur'an corpus in English and Indonesian translations, as a complement to previous research. The construction of this system requires data in the form of words contained in the Qur'an. The model used in this study is FastText, which is a development of the Word2vec model. Each word in FastText is modeled by several vectors, with each n-gram vector representation. This approach is considered to be very useful for rare words and can handle out-of-vocabulary (OOV) words [6]. FastText is popularly used in many studies, some of which are text-classification, sentiment analysis and semantic similarity. These factors are the reason this approach was chosen for this research. The system built is expected to produce an excellent performance based on the calculated correlation value. System evaluation is done by calculating the correlation system with WordNet where WordNet is a combination and expansion of dictionaries and thesaurus. Then we do the factors that can influence the correlation results from Fasttext to the gold standard.

DISTRIBUTIONAL SIMILARITY

The distributional similarity approach is used to model languages and represent naturally occurring texts. This is a statistical-based model that uses the statistical distribution of words along with their context to determine the level of semantic similarity between terms. This model illustrates words with context vectors built on the distribution hypothesis, which states that similar words appear in the same context. The proposed method semantic distributions to construct word-context matrices that represent the distribution of words across contexts and to transform the text into representations of vector space models (VSM) based on semantic word similarities. The measures of equality of distribution used to capture the semantic similarities of words and to make groups of similar words [7].

SEMANTIC SIMILARITY

Semantic similarity is a method for measuring and expressing similarity between words based on the meaning of word similarity. Semantic similarity is used to estimate or calculate semantic proximity, or the relationship between various constructs in language and concepts based on numerical descriptions. In general, semantics or similarities between two-word objects can be assessed using ontologies and are used to define relationships between terms [8]. Semantically is merely taking two terms (concepts or words) as input and returning a numerical score that counts the number of similar words. Semantic similarity considers all types of semantic relations between terms [9].

WORD EMBEDDINGS

Word Embeddings are vector space models (VSM) that represent words as vectors in continuous space capturing many syntactic and semantic relationships between terms [10]. Word embeddings recently gained great popularity for modeling words in various Natural Language Processing (NLP) tasks including measurement of semantic similarities. The very well-known word embeddings represent a new branch of the corpus-based semantic distribution model that utilizes neural networks to model the context in which a word is expected to appear. The ability of word embeddings to capture syntactic and semantic information, word embedding has been successfully applied to various NLP tasks, such as Word Sense Disambiguation, Machine Translation, Similarity of Relationships, Semantic Relatedness, and Knowledge Representation [11].

FASTTEXT

FastText is a new model for word embeddings that can capture word senses, sub-word structure, and information uncertainty. FastText models words with several vectors, where vectors represent n-grams. FastText produces accurate representations of rare words, misspellings, even unknown words [6]. This model is a well-known algorithm that creates word vectors for out-of-vocabulary (OOV) words. FastText learns morphological features using subwords, and a word vector can be produced even for words that do not exist in the dictionary [5].

COSINE SIMILARITY

Cosine similarity is a measure to calculate given pairs of sentences related to one another and determine scores based on words that overlap in sentences [12]. Cosine similarity can also be defined by angle or of the angle between two vectors. This is possible documents with the same composition to be treated identically which makes this the most popular size for text documents. The vector has a unit length, then the cosine angle between two words is calculated by the dot product equation between the two vectors. The calculation of cosine similarity can be seen in the equation below [13].

$$Sim = Cos\ (\theta) = \frac{A\ .\ B}{||A||\ .\ ||B||}$$

To calculate the cosine equation between two sentences, the sentence is then converted to terms/ words. The word is transformed in the form of a vector, where each word in the text defines the dimensions in Euclidean space and the frequency of each word according to the value in the dimension [12].

GOLD STANDARD

The gold standard is a technique to evaluate the performance of a computerized system, which serves as a reference point for other things of its kind, which can be compared by calculating the correlation between the two. Gold Standard is often described as a high-quality data set that is explained by humans [14]. In this study, WordNet was used as a gold standard reference. WordNet is a lexical database and is considered the most extensive electronic dictionary. WordNet was developed by lexicographer experts whose results are made into a lexical database. WordNet is created manually by requiring a lot of resources such as language experts and time so that it has high quality [15].

PEARSON CORRELATION

The Pearson correlation is a general measure used to measure the linear relationship between two continuous variables. The Pearson correlation is defined as the ratio of

covariance of two variables to their respective standard deviations. Pearson correlation coefficients range from -1 to +1 [16]. The formula used in calculating correlations in Pearson correlation uses the equation below.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

As for its use in this study, *N* is the number of word pairs, *X* is the value of the system, then *y* is the value of the gold standard.

SYSTEM DESIGN

A. System Overview

The system built in this study is a system that can calculate the semantic similarity of the pair of words in the corpus of the English and Indonesian translation of the Qur'an. Semantic similarity values are obtained based on the implementation of the FastText method. Evaluation of the results of the system is done by calculating the correlation as a benchmark of similarity to the gold standard value. In general, the system is illustrated in the picture below.

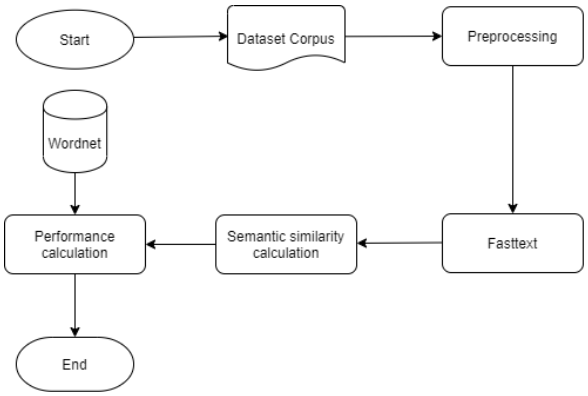

Figure 1. System Overview

B. Corpus Dataset

The corpus dataset used in this study is the corpus of the English and Indonesian translation of the Qur'an, which is obtained from an online site, *qurandatabase.org*. This site provides translations of the Qur'an in various languages. The dataset used is still in the form of a complete narrative, so it needs to be further processed.

C. Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more efficiently and effectively processed for user. The order of preprocessing can be seen in the image below.

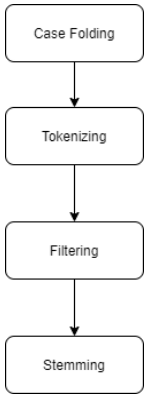

Figure 2. Preprocessing step

- Case Folding
  The initial step in processing is to do a case-folding that converts the entire text in the document into a standard form that is the lowercase.
- Tokenizing
  Tokenizing is a step that splits longer strings of text into smaller pieces or tokens. Larger chunks of text can be tokenized into sentences. Sentences can be tokenized into words.
- Filtering
  The third step is the crucial password from the token result. This stage uses the stoplist algorithm to list the words that are not descriptive, such as "which", "and", "in", "from", etc.
- Stemming
  The final step of preprocessing is stemming, which is the process of removing affixes, suffixes, and prefixes. The stemming process is not carried out in this study, because it is deemed incompatible with the corpus used based on the results of the system output.

D. FastText

FastText training stage aims to produce corpus modeling that is converted into vector form. To do this stage, the required data from preprocessing results in the previous stage, and then represented in a vector using FastText. In this study, we use the gensim library to running FastText. Gensim is billed as a Natural Language Processing package that does "Topic Modeling for Humans". But it is practically much more than that. It is a leading and a state-of-the-art package for processing texts, working with word vector models (such as Word2Vec, FastText etc) and for building topic models. The model built in the training process has parameters that affect the results of semantic similarity in words. Below are the parameters used in this study.

- Embedding size: this parameter is used to determine the dimensionality of a vector, where the dimension of the word vector must be an integer.
- Window size: this parameter determines the maximum distance between the target word and the words around the target word.

DOI: https://doi.org/10/25047/jtit.v7i1.135 ©2020 JTIT

- Minimum frequency: this parameter that determines the minimum frequency of a word in the corpus.
- Down sampling: this parameter determines the number of samples taken based on words that often appear.
- Train model: parameters that determine the model used. 1 or 0.1 ways training the Skip-gram model, and 0 means are, training the Continuous Bag of Word (CBOW) model. The training model used for this research is the Skip-gram model.
- Iteration: a number to determine how much training is done. the iteration implemented is 100

The results of the training produced by the system produce a vector of each word from the corpus data. FastText produces every word that can have more than one vector, because every word in a sentence has a different context. Examples of the results of system output with the input *"verse"* can be seen in figure 3, then for the Indonesian translation corpus with the input word *"ayat"* can be seen in figure 4. For example, we use an embedding size of 60, the system will display 60 vectors words according to the embedding size specified in the training model.

```
[ 0.17395614  0.41536897 -0.45998493  0.47009876 -0.6628106  -0.48038128
  0.04055812  0.3897398  -0.64992595 -0.49506357 -0.0683428  -0.18011461
  1.0589808  -0.45836365  0.57261336  0.24008612  0.9942866   0.98182297
  0.37293917  0.73189145 -0.50056493 -0.06233291 -0.08219557 -0.90214884
  0.640637   -0.3089008   0.81220824 -0.24856307 -0.34277466  0.67015845
 -0.8952421  -0.5458824   0.37099937  0.7095037  -0.5635748  -0.79176676
  0.57667047  0.34072214  0.99566996  0.2970866  -0.78841305  0.03290058
  1.1427369  -0.07600334 -0.46750852 -0.3474398   0.6180059   0.8350583
 -0.29526553 -0.50421345 -0.33583328 -0.5779925  -0.18977326 -0.4374255
 -0.16101952  0.6383303  -0.5101725   0.47073033  1.1498312  -0.3604628 ]
```

Figure 3. System output of the word vector with English input *"verse."*

```
[ 0.38972422  0.17614502  0.42351964 -0.099403    0.6022625  -0.22197805
  0.5637281  -0.12799373 -0.05638991 -0.8038889  -0.37783825  0.39310646
 -0.33812183 -1.2966657  -0.7660862   0.3430934  -0.12245457 -0.44380635
 -0.84767514  0.6241475  -0.27278626  0.04162245  0.3323735  -0.2189648
 -0.36617643 -0.06907339  0.38844192 -0.10143476  0.19108443  0.23305917
 -0.52003455  0.20668812  0.3618938  -0.19457711 -0.6372332  -0.62346417
  0.24228764 -0.38931492  0.47132176  0.66142    -0.23669472  0.04717317
  0.15402065  1.2949777   0.45634902  0.18077339  0.03500935  0.11103446
 -0.3331516   0.05895928 -0.07838535 -0.31051868  0.23729515  0.0197126
 -0.0757603   0.15009095  0.40844336  0.8859281   0.6751713  -0.11666941]
```

Figure 4. System output of the word vector with Indonesian input *"ayat."*

D. Semantic Similarity Calculation

After doing the vector calculation process of each word in the translation of the verses of the Qur'an, the calculation of semantic similarities is done using Cosine similarity. The system can produce several words that are judged to be the most similar to the input word. Examples of system output can be seen in Figure 5 and Figure 6.

```
verse:['recited', 'reciting', 'wrath', 'book', 'recognize', 'took']
merciful:['forgiving', 'pardoning', 'appreciative', 'pardon', 'forbearing', 'a
truth:['story', 'statement', 'falsehood', 'lead', 'bringing', 'true']
creature:['moving', 'ascribe', 'trees', 'colors', 'staff', 'decrees']
knowledge:['thought', 'shout', 'proofs', 'proof', 'mislead', 'without']
```

Figure 5. The Related Word output with English word input

```
ayat:['dibacakan', 'membacakan', 'ayatnya', 'membacakannya','ingkar', 'quran']
penyayang:['pengasih', 'pengampun', 'pemaaf', 'penyantun', 'maha', 'penerima']
kebenaran:['kebenarannya', 'jauhnya', 'risalah', 'sejauh', 'bukti', 'membawa']
makhluk:['memulai', 'melata', 'disembah', 'pencipta', 'baru', 'mengurus']
pengetahuan:['ilmu', 'singgasana', 'menduga', 'berselisih', 'wahyu', 'kitab']
```

Figure 6. The Related Word output with Indonesian word input

The system that was built can also measure the value of similarity to test the interrelationships between a pair of words. The higher the similarity produced by the system, then the pair of input words tested were judged to be more similar. The testing can be seen in table 1.

Table 1. Similarity score of a pair of words

English Corpus			Indonesian Corpus		
Word 1	**Word 2**	**Similarity Value**	**Word 1**	**Word 2**	**Similarity Value**
verse	reciting	0.5443	ayat	dibacakan	0.7244
merciful	forgiving	0.6967	penyayang	pengasih	0.8013
truth	story	0.5710	kebenaran	cerita	0.4317
creature	moving	0.7824	makhluk	melata	0.6019
knowledge	proofs	0.4974	pengetahuan	ilmu	0.6632

In Figures 7 and 8 there is a visualization of related words from input words that have been represented in two dimensions. The collection of dots in Figure 7 and 8 represents the distribution of word relatedness in the Qur'an based on the vector value generated from the FastText process.

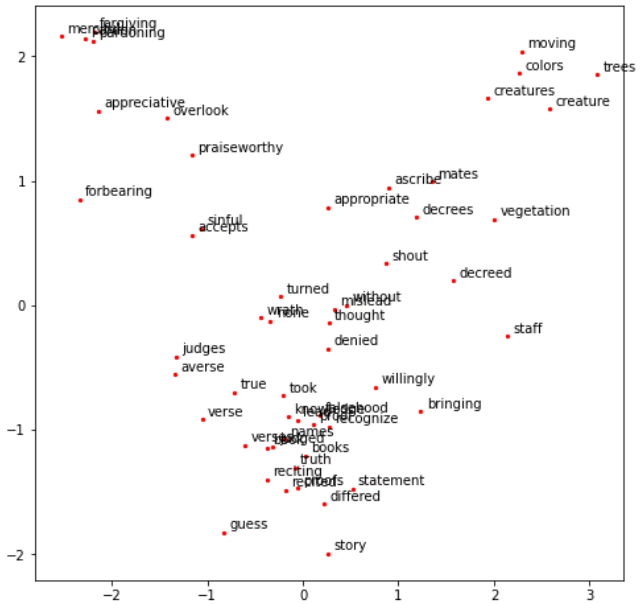

Figure 7. visualization of word linkage distribution vectors with English input

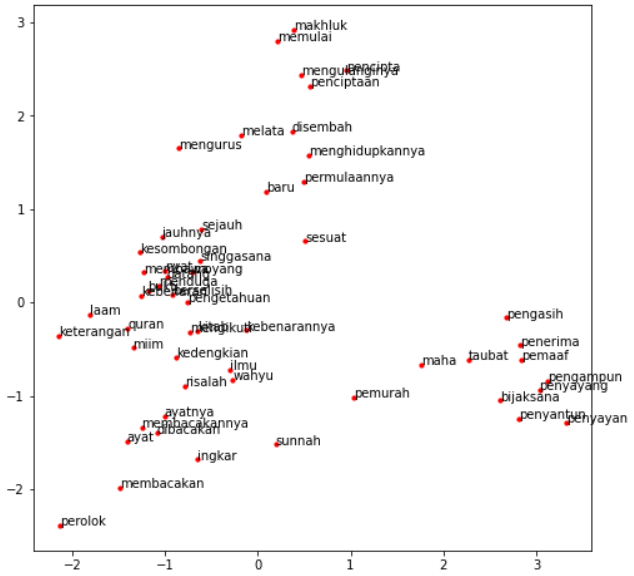

Figure 8. visualization of word linkage distribution vectors with Indonesian input

E. Performance Calculation

Performance calculations are performed to compare the semantic similarity values of system outputs to the gold standard values using the Pearson correlation. The gold standard value is obtained from the calculation of Wu-palmer similarity, where the relation considers the depth of two sunsets in WordNet's taxonomy and the extent of LCS (Least Common Subsumer).

EVALUATION

System testing in this study was conducted to see the performance of system output results against the gold standard. Measurement of system performance is measured using the Pearson correlation. Correlation measurement done by comparing ten words related to system output for each word input with a gold standard value.

A. Testing Scenario

Tests carried out are analyzing the semantic similarity relationship between input word pairs. The output of the system will then be compared with the synsets of the gold standard. Testing is done is to enter each set of words contained in the corpus of English and Indonesian translations, then every ten words related to system output, the correlation is calculated with the gold standard. Scenarios The parameters used in this test are 100 embedding sizes with window sizes 5, 7, and 10 for each corpus. Examples of testing for one-word input can be seen in Figures 9 and 10, then for the resulting correlation can be seen in table 2.

	Kata input	Kata terkait	Synset Kata input	Synset kata terkait	Nilai keluaran sistem	Nilai gold standard
0	faith	faithful	Synset('religion.n.02')	Synset('congregation.n.01')	0.665482	0.666667
1	faith	jealousy	Synset('religion.n.01')	Synset('jealousy.n.02')	0.527113	0.500000
2	faith	youth	Synset('religion.n.02')	Synset('young.n.09')	0.508796	0.461538
3	faith	rancor	Synset('religion.n.01')	Synset('resentment.n.01')	0.508759	0.333333
4	faith	loser	Synset('religion.n.01')	Synset('loser.n.01')	0.506184	0.153846
5	faith	belief	Synset('religion.n.01')	Synset('belief.n.01')	0.505900	0.923077
6	faith	fails	Synset('religion.n.01')	Synset('fail.v.01')	0.496456	0.000000
7	faith	increased	Synset('religion.n.01')	Synset('increase.v.01')	0.479868	0.000000
8	faith	content	Synset('religion.n.01')	Synset('content.n.05')	0.477129	0.833333
9	faith	tranquility	Synset('religion.n.01')	Synset('tranquillity.n.01')	0.466495	0.400000

Figure 9. Comparison of system output values against the gold standard with English input

	Kata input	Kata terkait	Synset Kata input	Synset kata terkait	Nilai keluaran sistem	Nilai gold standard
0	iman	imannya	Synset('creed.n.01')	Synset('creed.n.01')	0.495183	1.000000
1	iman	mengucapkan	Synset('religion.n.01')	Synset('state.v.01')	0.431458	0.000000
2	iman	quran	Synset('religion.n.01')	Synset('koran.n.01')	0.426860	0.285714
3	iman	menyiapkan	Synset('religion.n.01')	Synset('ready.n.01')	0.423812	0.400000
4	iman	perbaikan	Synset('religion.n.01')	Synset('sir.n.01')	0.422331	0.142857
5	iman	menyeru	Synset('religion.n.01')	Synset('entreaty.n.01')	0.416624	0.400000
6	iman	turun	Synset('religion.n.01')	Synset('descent.n.03')	0.411477	0.375000
7	iman	keimanan	Synset('creed.n.01')	Synset('creed.n.01')	0.411074	1.000000
8	iman	nafsunya	Synset('religion.n.01')	Synset('frailty.n.02')	0.408193	0.470588
9	iman	seruanmu	Synset('religion.n.01')	Synset('entreaty.n.01')	0.405292	0.400000

Figure 10. Comparison of system output values against the gold standard with Indonesian input

Table 2. The correlation of related words is based on the word input

Input Word	Pearson correlation with ten related words
faith	0.2720
iman	0.4049

The primary test is carried out to calculate the correlation for each set of words contained in the corpus of the Qur'an in English and Indonesian translations, which are filtered based on the terms listed on the gold standard (WordNet). The registered words consist of 4505 for the English translation corpus, and 4821 for the Indonesian translation corpus. Then for the final test, Pearson's correlation is calculated for all words with the gold standard.

B. Testing Result

Tests carried out are analyzing the semantic similarity relationship between word pairs with FastText based on window size. Three window sizes used in this test are window size 5, window size 7, and window size 10, and the corpus used is the English and Indonesian translation of the Qur'an. The value of the system output from each window size will be compared with the value of the gold standard with the Pearson correlation calculation. The test results can be seen in Figure 11.

DOI: https://doi.org/10/25047/jtit.v7i1.135 ©2020 JTIT

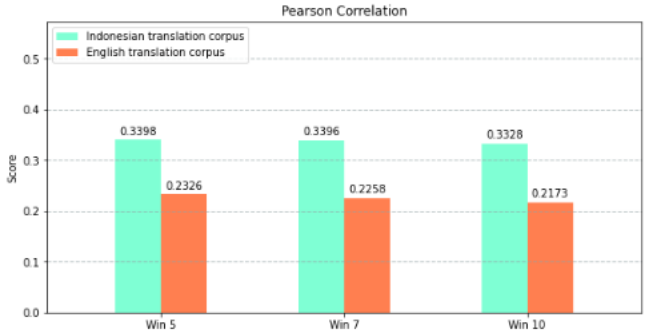

Figure 11. Correlation of system output with the gold standard

The test results produced in Figure 10 show that the test results using the Indonesian translation Al-Qur'an corpus showed the highest correlation value compared to the English translation corpus. The results of each correlation on the Indonesian translation corpus are, window size5 is 0.3398, window size7 is 0.3396, and the smallest correlation value is found in window size 10 of 0.3327.

Correlation results describe the level of accuracy of the system results. We did not compare the accuracy results with previous studies. That's because the datasets that we use are different, previous studies used the Qur'anic corpus with Arabic language and writing. As for our research which is a continuation of previous research, we use English and Indonesian translation corpus which is a complement to previous research. Then our evaluation method is different, namely in previous studies using precision in the range of 0% -100%, based on corrections from linguists. As for this research, we use Pearson correlation with range (-1) - (+1) based on an electronic dictionary/thesaurus namely WordNet. In the previous research, the accuracy that was produced was a precision value of 98% by choosing only 10 sample input words and then corrected by a linguist. Whereas for this research the test was conducted for all words contained in the corpus that we used.

Table 3. correlation of each word output system with gold standard for English Translation

Size of Correlation	Parameter		
	Window size 5	Window size 7	Window size 10
0.80 to 1.00 (−.80 to−1.00)	456	456	436
0.60 to 0.80 (−.60 to−.80)	801	779	775
0.40 to 0.60 (−.40 to−.60)	889	909	900
0.20 to 0.40 (−.20 to−.40)	973	938	982
0.00 to 0.20 (.00 to−.20)	1386	1423	1412

Table 4. correlation of each word output system with gold standard for Indonesian Translation

Size of Correlation	Parameter		
	Window size 5	Window size 7	Window size 10
0.80 to 1.00 (−.80 to−1.00)	610	594	570
0.60 to 0.80 (−.60 to−.80)	1159	1167	1157
0.40 to 0.60 (−.40 to−.60)	1073	1069	1053
0.20 to 0.40 (−.20 to−.40)	1030	948	968
0.00 to 0.20 (.00 to−.20)	949	1030	1065

The test results in tables 3 and 4 are the output of the correlation calculation between the ten words related to the system output with the gold standard, then the level of correlation is determined based on the criteria. Based on table 3 and 4, the Indonesian translation corpus produces more words with strong correlations, while the English translation corpus produces more words with low correlation. Based on table 4 for each window size, it can be seen that the larger the window size does not increase the strong correlation criteria, while the low correlation always increases.

Table 5. Some words not found in the gold standard (WordNet)

English Translation Corpus	Indonesian Translation Corpus
ababil, zulkifli, ruhul, fidyah, nazar, yusuf, zakarya, qudus, ibnusabil, taufik, mushrik, mudharat, baitul, qarun, jihad, mahfudz, zarrah, qabil, habil, sunnatullah, lim, khamar, sakaratul, kaffarat, tabiat, sidratul, ihram, mukmin, hawariyyin	luqman, hunain, aliif, ahqaf, tayammum, kaaf, imran, iblis, talut, zihar, umat, jizyah, aqsa, umrah, ramadhaan, yajuj, yusuf, shaitaan, surah, masajid, injeel, ihram, salih, zaboor, lahab, taurat, marwah, yathrib,badr, kauthar

The test results for the words in Table 5 are not correlated because of the unavailability of these words in the gold standard. That is because the word concerned is still in Arabic, so no correlation calculations are made in these words.

C. Analysis of Testing Results

The process of calculating the value of semantic similarity between words using FastText is done with a corpus that has a different size. After preprocessing data, the number of words processed in FastText consists of 96971 words in the Indonesian translation corpus and 53804 in the English translation corpus. Tests are also carried out using a different window size for each corpus used. This aims to find out what are the factors that can influence the results of the value of the use of FastText for the calculation of semantic similarities between words.

The test results show that several factors can influence the performance value of the system. Here are the things that affect the amount of semantic similarity calculations using FastText:

- Use of parameter values. The results of FastText correlation show the window size 5 parameter has the highest correlation value, while the lowest correlation value is generated by window size 10. The use of window size can determine the number of possible words in pairs with other words. So that the similarity value produced by FastText can be increased based on determining the amount of window sizes
- Size of the corpus of data. The size of the corpus is very influential on the value of the performance of the system output. The larger the corpus size, the more vocabulary the corpus has, so the better the semantic similarity values produced by the system.
- The vocabulary owned by the corpus greatly influences the output of semantic similarity in words. That is because the training process captures pairs of words that often appear in the corpus used.
- Performance values depend on WordNet as the gold standard. WordNet cannot yet recognize a name from an entity such as "mushrik", "mudharat", and a few words with other Islamic contexts.

CONCLUSION AND SUGGESTION

Conclusion

Based on the results of tests and analyzes that have been done, the correlation score produced by the system is relatively low, in the Indonesian translation corpus produces a correlation of 0.3398 and an English Translation corpus of 0.2326. In this study, we analyze the use of parameters in FastText which aims to see the best correlation results produced by the system. The best correlation is obtained using a windows size of 5 with vector dimensions of 100. The use of window size used must be adjusted to the size of the corpus. The large size of the corpus with the adjustment of the window size affects the suitability of the word similarity value produced in the training process.

Another Factors that affect the value of semantic similarity between words using FastText is that a pair of words is generated influenced by the number of words appearing in the corpus. This is based on the corpus used, where after preprocessing data, the number of words processed in FastText consists of 96971 words in the Indonesian translation corpus and 53804 in the English translation corpus.

The gold standard used as an evaluation reference is also a factor influencing the correlation results. The use of WordNet as the gold standard in this study, cannot be used as the main reference for testing the system for corpus translation of the Qur'an. WordNet has not been able to capture several words related to the Islamic context.

Suggestion

The following recommendations might be useful as an extension to develop this study, or simply to avoid errors when conducting such similar research:

- Test with other approaches, so that more optimal performance values can be generated.
- Developing FastText performance measurements involves selecting complex parameters such as window size, embedding size, minimum frequency, down sampling, and training models.

BIBLIOGRAPHY

[1] M. Zahid, "Perbedaan Pendapat para Ulama Tentang Jumlah Ayat Al-Qur'an dan Implikasinya Terhadap Penerbitan Mushaf Al-Qur'an di Indonesia," *NUANSA: Jurnal Penelitian Ilmu Sosial dan Keagamaan Islam*, 2012.

[2] M. R. Daulay, "Studi Pendekatan Alquran," *Thariqah Ilmiah*, vol. 1, 2014.

[3] Y. Jiang , X. Zhang, Y. Tang and R. Nie, "Feature-based approaches to semantic similarity assessment of concepts using Wikipedia," *Information Processing \& Management*, vol. 51, pp. 215-234, 2015.

[4] F. F. Guntara, "Pembangunan Daftar Kata Terkait pada Kosa Kata Al-Qur'an Berdasarkan Kesamaan Distribusiaonal," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 6, pp. 139-146, 2020.

[5] J. Choi and S.-W. Lee, "Improving FastText with inverse document frequency of subwords," *Pattern Recognition Letters*, vol. 133, pp. 165-172, 2020.

[6] B. Athiwaratkun, A. G. Wilson and A. Anandkumar, "Probabilistic fasttext for multi-sense word embeddings," *arXiv preprint arXiv:1806.02901*, 2018.

[7] A. A. wajan, "Semantic similarity based approach for reducing Arabic texts," *International Journal of Speech Technology*, vol. 19, pp. 191-201, 2016.

[8] C. Yang, Y. Zhu, M. Zhong and R. Li, "Semantic Similarity Computation in Knowledge Graphs: Comparisons and Improvements," in *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, 2019.

[9] M. Kathuria, C. Nagpal and N. Duhan, "Semantic similarity between terms for query suggestion," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2016.

[10] A. B. Soliman, K. Eissa and S. R. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," *Procedia Computer Science*, vol. 117, pp. 256-265, 2017.

[11] I. Iacobacci, P. M. Taher and R. Navigli, "Sensembed: Learning sense embeddings for word and relational similarity," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015.

[12] P. Sravanthi and B. Srinivasu, "Semantic similarity between sentences," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, pp. 156-161, 2017.

[13] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," *Journal of King Saud*

DOI: https://doi.org/10/25047/jtit.v7i1.135 ©2020 JTIT

University-Computer and Information Sciences, vol. 29, pp. 189-195, 2017.

[14] D. A. Wiranata, M. A. Bijaksana and M. S. Mubarok, "Quranic Concepts Similarity Based on Lexical Database," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, 2018.

[15] I. P. P. Ananda, M. A. Bijaksana and I. Asror, "Pembangunan Synsets untuk WordNet Bahasa Indonesia dengan Metode Komutatif," *eProceedings of Engineering,* vol. 5, 2018.

[16] H. Zhou, Z. Deng, Y. Xia and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing,* vol. 216, pp. 208 - 215, 2016.