Cross-Lingual Semantic Similarity in Pieces of Al-Quran Verses Translation Using Word Alignment and Semantic Vector Approach

Reza Amelia[1], and Moch. Arif Bijaksana, Ph.D[5]

[1,2] Telkom University, Bandung, Indonesia
rezaamelia@student.telkomuniversity.ac.id
arifbijaksana@telkomuniversity.ac.id

Abstract. Al-Quran is a holy book of Muslims, that serves as a guide of life. Therefore, it is very important to understand the meaning of the Qur'an verse. The understanding is not only in the Qur'an in a single translation language but also in different translation languages. This research proposed the solutions to these problems with the analysis and implementation of Cross-Lingual Semantic Similarity on the translation piece of Al-Qur'an verse using Word alignment and Semantic Vector approach. The experimental result shows that the correlation value using combination between Word Alignment and Semantic Vector method is 0,62007 while the correlation value using Word Alignment method is 0,63231, and Semantic Vector method is 0,34841. This value was obtained by comparative value in the form of gold standard, namely the semantic similarity on the pair of translation pieces on the verses of Al-Quran in Indonesian and English based on human assessment manually. The correlation value is quite low due to the translation of Indonesian Al-Quran verses by using google translate feature before the semantic similarity calculation process produces many results that do not match the sentence of origin that is translated. Therefore the semantic value is much different from the gold standard.

Keywords: component, formatting, style, styling, insert (key words)

1 Introduction

Al-Quran is the holy book of Muslims, which contains a collection of divine revelation that was revealed to the Prophet Muhammad for about 23 years, which is called Al-Quran [12]. Al-Quran consists of 30 juz, 114 surah, and 6236 verses with a majority in the form of short sentences. On the 6236 verses, there are verses that have the same meaning and are interrelated but are not in the same surah or juz, so it is difficult to relate the information between the verses for layman. Moreover such verses have a different transalation language because in this modern era, Al-Quran has been translated into many languages in the world, and the most widely found translation is

Proceedings of the International Seminar on Information and Communication Technologies 2019 (ISICT 2019)
2 March 2019, Surabaya, Indonesia.
Institut Teknologi Telkom Surabaya, Indonesia
ISBN: 978-623-94143-0-6

in English. At the same time if viewed from its main function as a guide to life, it is very important for Muslims to understand the entire contents of Al-Quran.

To be able to understand Al-Quran thoroughly, one way is by knowing the Al-Quran that have similarities and linkages of meaning. This underlies the research of Cross-Lingual Semantic Similarity on the pairs of Al-Quran translation with the aim to simplify the calculation and search of similarities in meaning to different Al-Quran verses and in different languages. In this study, the translation of Al-Quran verses to be campared is only the Indonesian and English translation.

The approach used is a combination of Word Alignment and Semantic Vector. This approach was chosen because in the competition of SemEval on task STS in 2014, Word Alignment method produced a fairly high correlation and ranked first. Because there are still deficiencies in Word Alignment method, then in 2015 and 2016, Word Alignment method was combined with Semantic Vector and resulted in higher correlation than in 2014. Ridge Regression Model was used to combine both of these methods. In addition, a gold standard was also made. As a comparison of final result, it requires a gold standard to determine the similarity between systems with human intuition.

2 Related Work

Several studies related to semantic similarities among sentences have been conducted in several previous studies. The research [1] describes model architectures for computing continuous vector representations of words from very large data sets, namely CBOW and also Continuous Skip-gram Models. In this study, five types of semantic questions and nine types of syntactic questions to measure the quality of word vectors were used.

The research [5] conducted Semantic Textual Similarity (STS) research with new datasets in English and Spanish including image captions, news headlines, Wikipedia articles, news, and new genres like answers from a tutorial dialogue systems, answers from Q & A websites, and commited belief. The most widely used tool in this study is Stanford's NLP parser and the OpenNLP framework. Furthermore, named-entity recognition and acronym repositories, ConceptNet, NLTK, and time and date resolution or PPDB were also used. Other than that, WordNet or Mikolov embeddings were used to compute word similarity.

The research [3] introduced the first probabilistic approach to modeling cross-lingual semantic similarity (CLSS). The data set used in this study has been constructed in Spanish (ES), Italian (IT), and Dutch (NL) with the aim of obtaining a

third translation of the three languages into English. This study used latent cross-lingual concepts.

In the SemEval competition in the 2014 STS task, the research was conducted [4] by using the Word Alignment method to calculate semantic similarities between two sentences. This study uses 6 datasets where each test data has a number of sentence pairs, and each pair of sentences has a human-assigned similarity score in the range [0, 5] which increases with similarity. With the method used in this research, it produces a fairly high correlation and ranks first among the 38 other systems.

In 2015 and 2016 the Word Alignment method was combined with Semantic Vector and produced higher correlation results than in 2014. In study [6], the Word Alignment method was combined with Semantic Vector with the correlation results obtained reaching 80.15%.

The research related to Cross-Lingual Semantic Similarity was started on SemEval 2016. In the research [7] the approach used is based on supervised regression with an ensemble decision tree using cross-lingual data set where one of the sentences is in English and the other is in Spanish. The system built produced Pearson correlation values of 0.39533 and ranked 7th of all SemEval 2016 participants.

The other related research is [9]. In this study the data sets used were English sentence pairs whose semantic similarities would be calculated. The data was used in SemEval from 2012-2015. This aimed to train the regression model built. The system created in this study obtained a correlation value of 73.6%.

3 Research Method

3.1 Design

The researchers wake system aims to find common ground between two pieces of semantic translation of the verses of the Quran in two languages, i.e. the English Language and Indonesian one. Prior to the measurement value equality semantics, the couple pieces of verse translations of the Koran were separated in advance between the United Kingdom and verse-speaking verse speaking in Indonesia. Then the piece of the translation of the text in the language of Indonesia is translated into the language of the United Kingdom automatically by the system. After the results obtained in English Verse translation of the United Kingdom, the text pieces paired back as before. Semantic similarity value afterward measured using implementations of a method of Word Alignment and Semantic Vector using Word Embedding with the word2vec toolkit. To combine the results of semantic similarity between these two features used Ridge Regression Model. Having obtained the results of the regression, correlation calculations were done by comparing the semantic similarity value system

with the gold standard that already exists using the Pearson Correlation. An overview of the system to be built can be seen in figure 1.

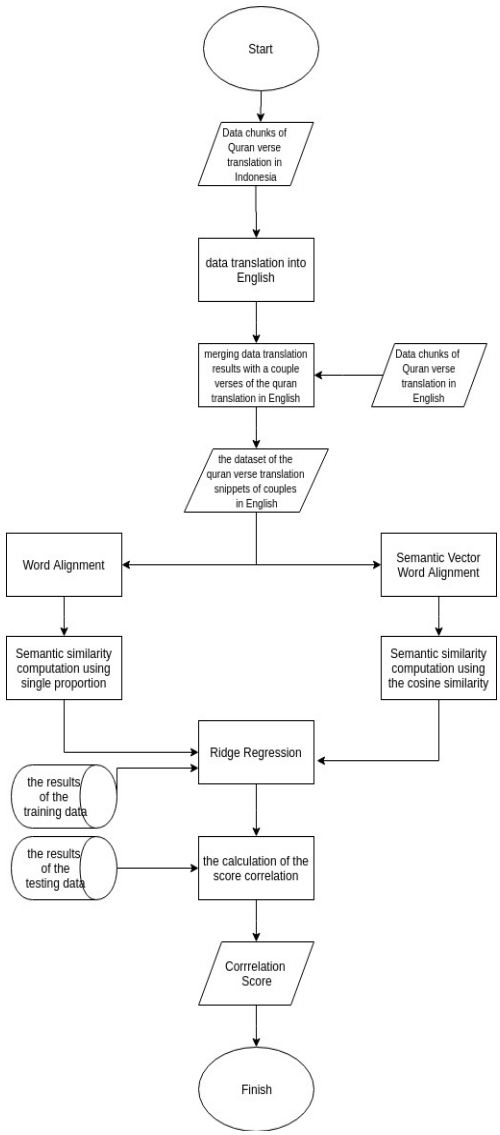

Fig. 1. System Design

3.2 Data Collection

Data Pair of The Quran Verses. Data to data obtained from the verse couples research conducted by Mustika Meiditia Rani and Dwi Jayanti Wulandari in 2017. In

the study, 550 pairs of verse translations are taken from the Qursim research results based on the interpretation of the Quran corpus linkages Ibn katheer Abdul-Baquee Sharaf and Eric m. s. Atwell, and as much as 250 couples verses taken from the Thematic Index Ministry of religious affairs of the Republic of Indonesia and the Centre for the study of Hadith Al-Mughni. So that the total of all the data on the research of the year 2017 totalling 800 data.

Then the data selected from the overall data 260 tafseer Ibn katheer and 40 thematic index data from the study the year 2017 and 100 selected data from tafseer Ibn katheer research year 2016 for the dataset. The previous translation is done in advance of one verse in one pair of verses into the language of Indonesia based on tafseer Ibn katheer Abdul-Baquee Sharaf and Eric m. s. Atwell. The following example data couple verses on table 1 and table 2.

Table 1. Tafseer Ibn Katheer Data

Verse Index	Verse 1	Verse 2
2-4, 4-136	Dan mereka yang beriman kepada Kitab alQuran yang telah diturunkan kepadamu dan kitab-kitab yang telah diturunkan sebelumnya serta mereka yakin akan adanya kehidupan akhirat. (And who believe in what has been revealed to you O Muhammad and what was revealed before you and of the Hereafter they are certain in faith.)	O you who have believed believe in Allah and His Messenger and the Book that He sent down upon His Messenger and the Scripture which He sent down before.

Gold Standard Data. Gold standard data contains the data values for the semantic similarity of pairs of pieces of verse translations of the Quran obtained manually by human judgement and would later serve as a comparison to the results of semantic similarity score that is generated by the system . Gold standard has a range of values from 0 to 5. The higher the value of the gold standard, meaning that the two pieces of the verse are very similar.

Table 2. Thematic Index Data

Verse Index	Verse 1	Verse 2
9-32, 37-8	Syaitan syaitan itu tidak dapat mendengar dengarkan pembicaraan Para Malaikat dan mereka dilempari dari segala penjuru. (They want to extinguish the light of Allah with their mouths but Allah refuses except to perfect His light although the disbelievers dislike it.)	So they may not listen to the exalted assembly of angels and are pelted from every side.

3.3 Data Translation of The Quran Verses

Before it can be further processed by the system, the data chunk translation of verses of the Quran is translated in the past using the google translator in its web site translate.google.co.id. This is done by referring to the cross-lingual subtask on the SemEval 2016 competition. [7] a similar study using the Google Translate tool has also been done by Vulic in 2013 [2].

3.4 Prepocessing Data

Data pairs the piece of the translation of the verses of the Quran in both Indonesia and English one which has been translated into English Language is not yet ready to be used in this study. There for it is necessary to do preprocessing steps to make data ready to be used. Preprocessing stages use in this research are Tokenization, Stopwords Removal, and Lemmatization. After this step is complete, go to the next step, namely Alignment Feature will be explained in the next section

3.5 Alignment Features

The features used in the Alignment as indicated in Figure 2 below:

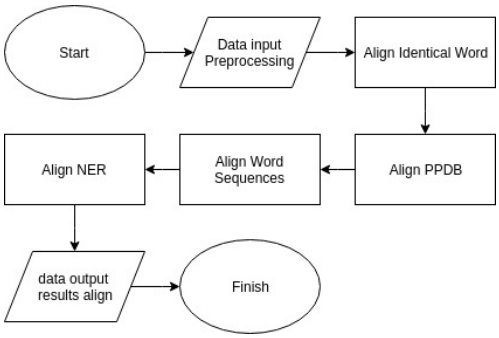

Fig. 2. Alignment Features Phase

From figure 2, examples of its application are as follows.
Given there is a pair of verses with preprocessing results as follows:

- Verse 1: [’mention’, ’said’, u’angel’, ’Prostrate’, ’Adam’, ’prostrated’, ’except’, ’Iblees’]
- Verse 2: [’remember’, ’said’, u’angel’, ’prostrated’, ’Adam’, ’prostrated’, ’except’, ’Iblis’]

Then the results obtained on every feature of the alignment are:

- Align Identical Word: [[said,said], [angels,angels], [Adam,Adam], [prostrated,prostrated], [prostrated,prostrated], [except,except]]
- Align PPDB: []

- Align Word Sequence: [Adam,Adam], [prostrated,prostrated], [except,except], [said,said], [angels,angels]
- Align Named Entities: [Adam,Adam]

The end result of this phase is a Word from the first and second sentences are aligned with the index word in the sentence. Word aligned are as follows:

[[said,said], [angels,angels], [Adam,Adam], [prostrated,prostrated], [prostrated,prostrated], [except,except]]

These features run in order, if there are features that give off the same alignment results are then to be taken is the alignment of the results of the previous feature in order to avoid redundancies. Semantic similarity of results obtained using the equation as follows:

$$sts(S^{(1)}, S^{(2)}) = \frac{n_c^a(S^{(1)}) + n_c^a(S^{(2)})}{n_c(S^{(1)}) + n_c(S^{(2)})} \quad (1)$$

$$Semantic\ Score = \frac{6+6}{8+8} = 0{,}67 \times 5 = 3{,}3$$

3.6 Semantic Vector

Semantic similarity computation of Semantic Vector on this research is by using the equation of Cosine Similarity. Cosine Similarity is a calculation to measure similarity or similarity used in this research. In its use, any sentence that would have measured the sameness represented in vector space models [11].

Semantic Vector is the approach used to address limitations in word alignment which can only identify the word semantics for lexical paraphrasing. Research on semantic vector using the word process embedding is done using the toolkit word2vec which was later constructed sentence vector. As for the steps undertaken in this stage is like in figure 3 as follows:

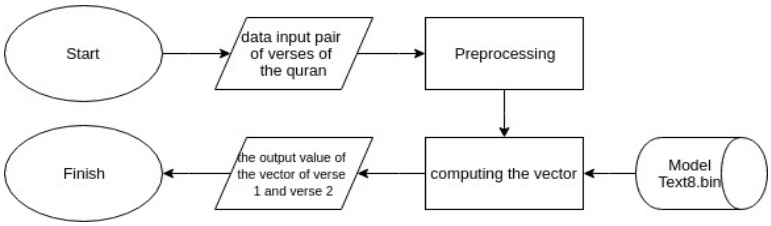

Fig. 3. Semantic Vector word2vec Phase

The stages being performed against the data preprocessing is doing a couple pieces of verse translations of the Koran language United Kingdom, then calculated the average vector of each sentence by using the help of the word2vec toolkit. With the example as follows:

Given an input pair with preprocessing results as follows:

- Verse 1 : [’\’, ’o’, ’you’, ’who’, ’have’, ’believed’, ’believe’, ’in’, ’allah’, ’and’, ’his’, ’messenger’, ’and’, ’the’, ’book’, ’that’, ’he’, ’sent’, ’down’, ’upon’, ’his’, ’messenger’, ’and’, ’the’, ’scripture’, ’which’, ’he’, ’sent’, ’down’, ’before’, ’.’,’\’, ’..’]
- Verse 2 : [’\’, ’and’, ’those’, ’who’, ’believe’, ’in’, ’the’, ’scripture’, ’which’, ’has’, ’been’, ’sent’, ’down’, ’to’, ’you’, ’and’, ’the’, ’books’, ’before’, ’it’, ’,’, ’and’, ’they’, ’are’, ’sure’, ’of’, ’the’, ’hereafter’, ’.’, """]

Then the computational results of the vector are as follows:

- Verse 1 : [2.20467597 0.61992224 -1.16566291 ….. -2.55170158 3.37612481 -0.96519544]
- Verse 2 : [1.76773137e+00 5.01511968e-01 -1.40453816e+00 …… -1.78670245e+00 2.41034764e+00 -4.68070671e-01]

For the computation of semantic similarity vectors both sentences use cosine similarity as in equation as follows:

$$Cosine\ Similarity = \frac{V1 \cdot V2}{\|V1\|\|V2\|} = \frac{238{,}626731129}{272{,}94530412}$$
$$= 0{,}874265750414 \times 5$$
$$= 4{,}37132875207$$

Where the value of the V1 · V2 is the vector dot product of multiplication sentence 1 (V1) with sentence 2 vector (V2).

3.7 Regression Results between Word Alignment with Semantic Vector

In this research, to combine the results of semantic similarity between the Word Alignment and Semantic Vector used method of Ridge Regression. The plot of this method are described in figure 4 as follows:

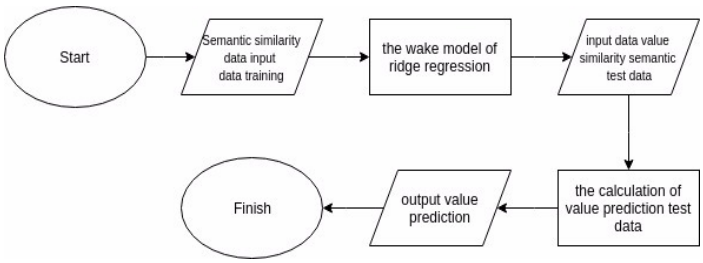

Fig. 4. Regression Phase

First, look for the value of the resulting semantic similarity in Word Alignment process and Semantic Vector on data training and test data. After that is done using the Regression model Ridge development ridge with the python module sklearn with the parameter alpha = 1.0, copyX = True; fitintercept = True; maxiter = None; normalize = False; randomstate = None; Solver = 0 auto 0; ToL = 0; 001.

These parameters refer to the results of research by 2016 SemEval MD Sultan [8]. If the Regression model Ridge is already built, and then calculated the value of semantic similarity predictions on the test data. For example, the value of semantic similarity results of Word Alignment is 2.8, semantic similarity and Semantic result Vector is 4.4, then the results of the regression is 3.6. The regression results are stored in a .txt file and then calculated its correlation value using Pearson Correlation.

3.8 Evaluation

Evaluation of results in this research was done by comparing the semantic similarity between the results system with semantic similarity values based on human judgement in the form of the Gold Standard. This process is done by using the equation of Pearson Corellation. Pearson Correlation equation is as follows:

$$r = \frac{n\sum(X_iY_i) - \sum(X_i)\sum(Y_i)}{\sqrt{n\sum(X_i^2) - (\sum(X_i))^2}\ x\ \sqrt{n\sum(Y_i^2) - (\sum(Y_i))^2}} \quad (2)$$

In this study, the variable X in the equation represents the semantic similarity value is generated by the system, while the variable Y is the value of the gold standard.

4 Result and Discussion

Based on the testing that has been done in this study using the method of Word Alignment, Semantic Vector, as well as the merger between both of these methods use the data pair Quran verse translation of the language of Indonesia with the United Kingdom, obtained results in table 3 as follows:

Table 3. The Results of Testing System

Methods	Correlation Result
Word Alignment	0,63231
Semantic Vector	0,34841
Word Alignment and Semantic Vector	0,62007

From table 3 can be aware that the correlation value obtained using the method of Word Alignment for measuring semantic similarity values on data pairs the piece of the translation of the verses of the Qur'an translation of Indonesia with the United Kingdom is of 0.63231. Obtaining the correlation value is affected by the Alignment of features that are used with the following details:

Table 4. The Results of The Influence of The Combination of The Alignment Feature

Features	Correlation Result
Identical Word	0,58367
Identical Word + PPDB	0,63434
Identical Word + PPDB + Word Sequence	0,63434
Identical Word + PPDB + Word Sequence + NER	0,63231

From table 4 it can be seen that the use of the Word produces Identical features Align value correlation is quite high, i.e. 0.58367. This feature is simple enough in determining the value of semantic similarity between two verses of the Quran, i.e. by specifying words that are similar or identical in these two verses. However, this feature is most dominant in determining the alignment results. This is because the data pairs the piece of the translation of the verses of the Koran were used in this study has a lot of similarities between the words without regard to the overall meaning of the sentence. The combination of Align Identical Word features and Align PPDB produce higher correlation value but with the difference that the values are not too far away with Align Identical Word, i.e. 0.63434. This is because the results of the Indonesia language translations use the vocabulary that many are not contained on the PPDB, or there are loan words that are not changed when translated.

The combination of Align Identical Word features, Align PPDB features and Align Word Sequence features yields the same correlation with a combination of Align Identical Word features and Align PPDB features instead. This is due to the identical

words sequence had already been aligned by the Identical Word feature so that the results of the Word Align Sequence feature does not affect the results of correlation.

The addition of Align Named Entities features, little affects the results of correlation, but if done rounding to two decimal places, this feature also does not affect the results of correlation and still remain on the results of correlation 0.63 as in the merger of three previous features. This is because at Al-Quran verse pieces used in this study rarely found names of people, places, or the like that are able to be identified by POS Tagging and Name Entity Recognition.

In table 3 it can also be noted that the use of the Semantic features of Vector generating value correlation of 0.34841, much smaller than the results feature of Word Alignment. This is because the research data on the verses of the Qur'an through the process of translation into the language of the United Kingdom first to use Google Translate before the calculated value of semantic similarity with the verses of the Qur'an which is the language of the United Kingdom her partner. The results of this translation is often not appropriate word choice and arrangement of words to use so that there is a pair of verses that were originally has the meaning and order of similar sentences become much different semantic similarity values and generate a small system , does not match the Gold Standard pre-set. In addition the text8 method used on the word2vec toolkit also have words that are not so much that there are some words that have no value to the array whose vectors. This also causes the value of semantic similarity several pairs of verses of the Al-Quran have value equality semantics are not approaching gold standard. Some examples of words that have no value to the array whose vectors is the United Kingdom which is the origin of the Arabic language, for example the word almasjid, alharam, iblees, kabah, jinn, tawaf, fitnah and other words that do not change shape when translated from the language of Indonesia into the language of United Kingdom, supposing the word Rabb.

5 Conclusion

Cross-lingual semantics similarity research are possible to be implemented. From this research, it was obtained the result of correlation using combination of Word Alignment and Semantic Vector method on the paired translation of the verses of Al-Quran in Indonesian and English in this research is 0.62007. While the result of correlation using Word Alignment method obtained the result of 0.63231, and last using Semantic Vector method obtained the result of 0.34841. Those values are quite small from the expected values. The results comfirmed that using Word Alignment only more accurate than the other two. The combination of Word Alignment and Semantic Vector gets smaller results because it influenced by the semantics result. The result of semantic vector is very small because the research data on the Indonesian verses of the Al-Quran through the process of translation into English Language first using Google Translate. Many results of the translation do not match the actual translation

from the Al-Quran, and that makes the similarity value different with the golden standard pre-set.

References

1. T. Mikolov, K. Chen, G. Corrado and J. Dean.: Efficient Estimation of Word Representations in Vector Space, CoRR, vol. abs/1301.3781 (2013).
2. I. Vulic and M.-F. Moens.: Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. ACL, pp. 106-116 (2013).
3. I. Vulic and M.-F. Moens.: Probabilistic Models of Cross-Lingual Semantic Similarity in Context Based on Latent Cross-Lingual Concepts Induced from Comparable Data. ACL, pp. 349-362 (2014).
4. E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio ˜ , M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria and J. Wiebe.: SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. ACL, pp. 252-263 (2015).
5. M. . A. Sultan, S. Bethard and T. Sumner.: Sentence Similarity from Word Alignment and Semantic Vector Composition. ACL, pp. 148-153 (2015).
6. D. Ataman, J. G. C. d. Souza, M. Turchi and M. Negri.: FBK HLT-MT at SemEval-2016 Task 1: Cross-lingual Semantic Similarity Measurement Using Quality Estimation Features and Compositional Bilingual Word Embeddings. ACL, pp. 570-57 (2016).
7. E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau and J. Wiebe.: SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. ACL, pp. 497-511 (2016).
8. M. A. Sultan, S. Bethard and T. Sumner.: DLS @ CU at SemEval-2016 Task 1: Supervised Models of Sentence Similarity. ACL, pp. 650-655 (2016).
9. D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio and L. Specia.: SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and. CoRR, vol. abs/1708.00055 (2017).
10. S. Xie and Y. Liu.: Using Corpus and Knowledge-Based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization. IEEE, pp. 4985-4988 (2008). https://doi.org/10.1109/ICASSP.2008.4518777
11. C. Saunders, A. Gammerman and V. Vovk.: Ridge Regression Learning Algorithm in Dual Variables. ICML (1998).
12. T. A. Amal.: Rekontruksi sejarah al-Quran. Pustaka Alvabet, Tangerang (2005).
13. M. Khumaini.: in Tafsir Asan, pp. 304.