Monolingual word alignment for parallel corpus of Al-Qur'an translation

Kurniawan Adhiethama[1], Moch. Arif Bijaksana[2] and Ibnu Asror[3]

[1,2,3] Telkom University, Bandung, Indonesia
adhiethama@students.telkomuniversity.ac.id,
arifbijaksana@telkomuniversity.ac.id,
iasror@telkomuniversity.ac.id

Abstract. To measure the semantic correlation between words there are many meth- ods that can be used, one of which is word alignment. Word alignment is a method that aligns words that have a letter correlation or meaningful correlation between two sentences. This study focuses on using word alignment in the translation of the Al-Quran verse. This method was developed to align the sentence pair data but can be used to measure the semantic correlation between verses. By using the al- gorithm back to basic word alignment developed by Sultan et al.[7] the researcher re-develops to research the alignment between verses in the Al-Quran, to find out the effect if used in the translation of the Al-Quran as the dataset. The Al-Quran dataset used will be converted into the MSR-RTE[2] dataset format by researchers, with the aim of providing new research results in the context of the Al-Quran word alignment. In Back to Basic Word Alignment there is a pipeline alignment that con- tains the use of a tour map sequence, the tour used in this research, align identical word, align PPDB, align word sequences, align named entities, align content words (dependency), align content words using surrounding words (text neighbor), align stop words, align PPDB Extended[7]. These features will be combined to determine the correlation value between two Al-Quran verses (F1 score). The best correlation value between verses that can be produced in this study was 51.02 % compared to the baseline research by Sultan et al. that is 91.7%. The correlation value between verses in this study can be concluded as a sufficient value, and can still be improved by adding features, knowledge base, or using a combination of different translators of the Al-Quran.

1 Introduction

To be able to understand the Qur'an, one of them is by understanding the translation of the verses of the Qur'an, it would be better to compare the two versions of the translation of the Qur'an as a reference of learning that has a correlation and rele-

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

vance of meaning to each the word in a verse, so that understanding the verses can get complete information about its meaning. This method is also applied in the interpretation of the Qur'anic verse, which is to interpret a verse with another verse, however, to measure the correlation of meanings between two text documents is still very difficult to do by the computer/system. Because the computer/system does not have intuition abilities like humans who are able to find correlations from two text documents. Therefore, a system is needed that can provide information on the correlation between verses in the Qur'an with accuracy similar to the results of human intuition. So, it can help Muslims in learning and understanding the Qur'an. In this case the system that can be used is text mining, which is a process of mining data in the form of text where the data source is usually obtained from a text document and the aim is to obtain useful information from a collection of documents so that connectivity can be done between documents [6] . In this research is carried out mining text from two translations of Al-Qur'an verses which will produce translation performance values in the Al-Qur'an verse.

In this research, one approach in mining text to solve a case using the word alignment approach to get a word translation from the Qur'anic verse. Word alignment [6] identifies the words or phrases contained in two pairs of sentences to determine the correlation of meanings between the two sentences. The approach back to basic word alignment [7] was chosen because the method uses the most basic algorithm and is easiest to implement in word alignment.

The problem that will be raised in this research is to produce correlation values between 2 translations in the Al-Quran verse. This problem is raised because the translation of the Qur'anic verse has a variety of translations, with each translation varying in the choice of words, phrases, or sentences to translate the verses of the Qur'an, but have the same core. In the system that will be built, use the corpus of the Al-Quran translation in English.

2 Related Study

In the research of Sultan et al. Monolingual alignment is simple and easy to imitate but still shows good performance in performing word by word alignment performance in sentences, then made the Back to Basics algorithm where the algorithm for performing performance has no supervision and uses few external resources. Based on the research hypothesis of Sultan et.al. It is said words with similarity exemplify prospective pairs for alignment if placed in similar contexts. [7].

With the summary of MSR the alignment corpus [2] was established from Recognizing Textual Entailment (RTE), the 2nd challenge data [1], evaluating directly and Monolingual word alignment for parallel corpus of Al-Qur'an translation 3 aligning evaluations is feasible. The 1st aligner tested and evaluated on the corpus is a phrasal

aligner called MANLI [5]. MANLI only combines only a few features in the characterization of contextual similarities, namely the comparative positions of two phrases aligned (or not) in two sentences and boolean features that represent the goodness of the previous token from two similar phrases. In this research we made translations of Al-Quran corpus following the MSR corpus writing format.

3 Experiment Scenario

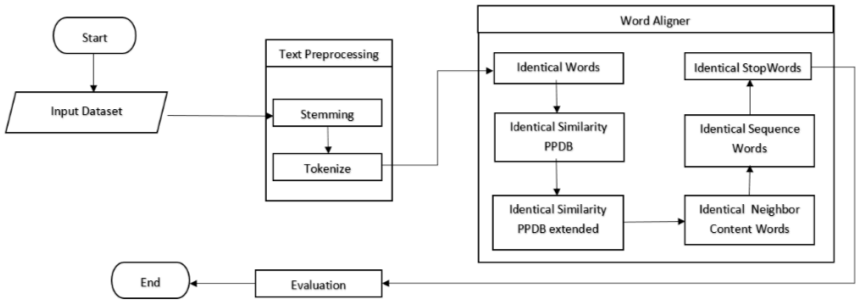
3.1 System Overview

The system built in this study is a system of measuring the correlation value of the translation of Al-Quran verses by taking a dataset from the translation of the English Al-Quran. This system was built to determine the relationship between words from each verse based on the feature feature of Back to Basic Word Alignment and calculate F1 score of each verse pair.

In figure 1 below, the dataset that has been converted into the MSR format by the author is inputted into the system. next, first processed with preprocessing using tokenize and stemming. after that go to the word alignment process using back to basic algorithm by using the feature feature alignment to find align of each word between translations, as long as this process uses knowledge database base paraphrase English and the extended version are included. the results of align between words will be calculated correctly and predictions whose value is converted to f1 score, which is the reference value of the relationship between words. Here is a description of flowchart from what describes this research in general:

Fig. 1. Flowchart System

3.2 MSR-RTE Data



MSR data is data from Microsoft Research made in 2006 with the aim that data can be used in various studies from information retrieval, semantic similarity to summarization [2]. MSR data is data that has information between text and alignment. This

dataset has alignment characteristics from one word to one word in one set of sentences and relatively does not consider the semantic correlation. The Example of data:

```
#sentence pair 1
Salvadoran reporter Mauricio Pineda , a sound techni-
cian for the local canal Doce television station , was
shot and killed today in Morazan department in the
eastern part of the country . NULL (/ /) Mauricio (3 p1
p2 / /) Pineda (4 p2 p1 / /) was (17 / /) killed (20 /
/) in (22 / /) Morazan (23 / /) . (32 / /)
```

Sentence pair is a sentence pair marker, the line right after sentence pair is the first sentence and the word *NULL* is a marker for the second sentence along with the gold standard, the number in parentheses is *gold standard*, if only the number then it is gold standard which is *sure* or must align and if there is a letter p then that is *gold standard* which is *possible* or may *align*. How to read *gold standard*, for example, Mauricio will *align* with Mauricio's word in the third index in the first sentence and p1, p2 meaning Mauricio *may align* with the Salvadoran and reporter [2].

3.3 Dataset

The data used in this study, using the Al-Qur'an verse English translation. Al-Qur'an verse data that is used has two hundred eighty six pairs of verses of the Al-Qur'an in the same English translation, with different translators.

The process of selecting the data set of the pair of Al-Qur'an verses in the same English translation, namely for the translation of the first Al-Qur'an verse is used the full translation of the Qur'anic verse, and for the translation of the second Qur'anic verse, the verse is used Al-Qur'an translations are changed according to the format of writing the MSR dataset. both versions are translated by Muhammad Sarwar and Wahiduddin Khan with 286 verses each. Example pair translation of Al-Qur'an verses :

```
verse 1 : He will say, "Oh, would that I had provided
beforehand for my life!" verse 2 : He will say, "Would
that I had done some good deeds for this life".

Convert into :

He will say , " Oh , would that I had pro-
vided beforehand for my life ! " NULL ( / /
) He(1 / / ) will ( 2 / / )say( 3 / / ), ( 4
7 / / )"( 5 18 / / ) Would( 8 / / ) that( 9
/ / ) I( 10 / / ) had( 11 / / ) done(12 / /
```

```
)some( / / ) good( / / )deeds( / / )for(14 /
/ ) this ( / / )life(16 / / )"(5 18 / / ).(
/ / )
```

3.4 Preprocessing

In the process of text pre-processing which is done the first time is tokenization to get the word token that is in the translation of the first and second verses. Then the Stemming process is done in the two translation verses to get the basic words of each word in the two translations of the verse.

3.5 Word Alignment

Word alignment used in this study is one method that is classified into the category of unsupervised methods. For MSR [2] datasets the system will issue word pairs that are similar in writing or in meaning, the results of the system will then be compared with the results of human annotations. The features used in this study are: word align align, align PPDB, align word sequences, align content words using surrounding words (text neighbor), align stop words, align PPDB Extended [7]. Align Identical Words features or modules in alignment that can be used are identical sequences of words. There are two indicators in determining the word correlation, the first is identical in string and the second is contextually [7]. Identical words in strings or letters between two sentences will be categorized as align. Align PPDB is aligner that relies on paraphrase database in determining alignment, adapted from paper Sultan et al. [8] submitted to SemEval, said the one that will be aligned will be checked in PPDB, if the word that will be texted (align) is in PPDB then the word pair will be categorized as the word align. Align Word Sequences feature is used to identify word pairs that have the same word order with a minimum of 2 word [7] correlations. Align Content Words that is an alignment that is devoted to words that have meaning. In doing alignment content words the alignment content words method is used by looking at the words in the neighbor or text neighbor [7]. Text Neighbor In identifying words that align in sentences, you can also use textual neighborhood, that is, checking the neighboring word pairs from the words that will be align and neighboring words will be grouped with 3 words to the right and 3 left words from the word which will be align which then 3 words right and 3 words from the word in the sentence will be cross product with 3 words on the right and 3 words left. In this study if the value for the word pair to be align is more than 0.9 [7] then the word pair is categorized as the word pair align. Align StopWords works similar to aligner text neighbor, the difference is stop words which will be aligned not content words. Each stop words in the sentence pair will be seen in 3 words on the right and left, if after calculating the value of the words around the word stop words is equal to 0.9 [7] or more then the stop words pair will be align. Align PPDB Extended will do alignment by checking back into the database created by the author by studying the dataset, the way it works is the same as aligner PPDB but in aligner This is not just one word that will be checked but up to three words.

3.6 F-Measure

F-measure is used to measure system accuracy monolingual alignment to data alignment [3]. The equation is as follows:

$$F1 = 2\ \frac{(Precision\ .\ Recall)}{Precision + Recall} \tag{1}$$

3.7 Paraphrase Database

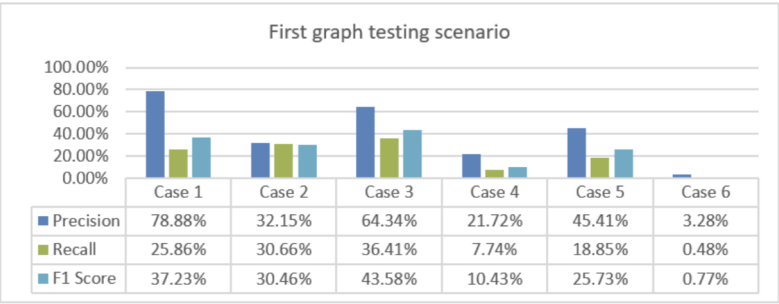
Paraphrase database is database which contains 220 million paraphrase including 72 million phrase and 8 million lexical paraphrase and 140 million paraphrase patterns[9]. PPDB was created by Juri Ganit kevitch and tim [4]. By using the text paraphrase database, the research work will be helped in measuring the accuracy of the proximity between the phrase that can be used to sort and calculate the semantic value of a pair of sentences. PPDB used in the system uses PPDB 2.0 XXL in English.

4 Analysis and Test Result

System testing is done using 286 verses that have been altered according to the format of writing MSR datasets that contain sentence pairs and the results of alignment based on annotator (gold standard). The system testing scenario is as following:

4.1 First Scenario Testing Analysis

Tests the Quran verse dataset with each feature in the system, with the final result in the form of performance value obtained from the alignment system, then compared to gold standard. The aim is to determine the effect of any text alignment features that have the greatest influence and which are not.



	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Precision	78.88%	32.15%	64.34%	21.72%	45.41%	3.28%
Recall	25.86%	30.66%	36.41%	7.74%	18.85%	0.48%
F1 Score	37.23%	30.46%	43.58%	10.43%	25.73%	0.77%

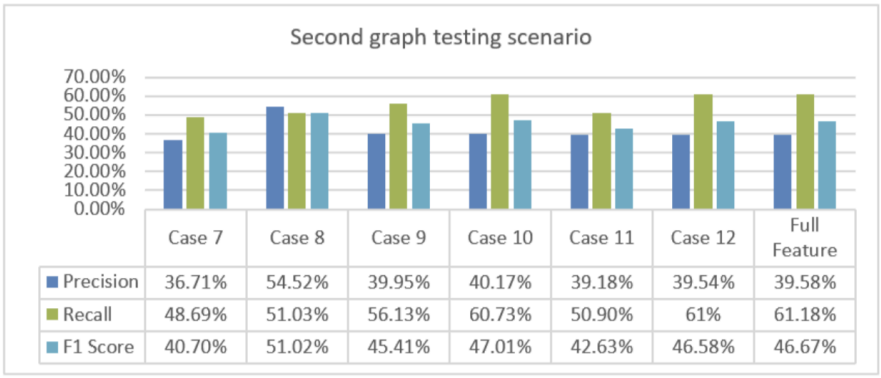
Fig. 2. First Scenario Testing

Based on the first test (figure 2), by testing the align (case 1 Identical Word, case 2 Sequence, case 3 PPDB, case 4 Neighbor, case 5 Stopwords, case 6 PPDB Extended) with the highest F1 score of 43.58% in case 3.

Case 3 can obtain the best f1 score because there are many words from the dataset that successfully align with words in the used 2.0 XXL Paraphrase database. The opposite thing happens in case 6 with the smallest f1 score because the number of words from the PPDB content is extended only slightly so that only a few words can align with the words in the dataset. Case 1 only gets f1 score of 37.23%, this means that words that have string and contextual similarities between words in the first version of the translation and the second version of the translation are only a few that succeed in aligning. Case 2 only gets f1 score of 30.46%, this means that word pairs that have the same word order with a minimum of 2 word-to-word correlations in the first version of the translation and the second version of the translation are only a few that successfully align. Case 4 only gets 10.43% f1 score, meaning that there are very few pairs of neighboring words from the first version of the translation that will be aligned with the neighboring words pair of words in the second version of the translation that is successfully crossed product between these words. Case 5 only gets a score of 25.7% f1, this means that only a few stopword non-content words pairs succeed in aligning between words in the first version of the translation and the second version.

4.2 Second Scenario Testing

Tests Qur'anic verse datasets using a combination of 5 features in the system, with the final result in the form of performance values obtained from the alignment system, then compared to gold standard. The aim is to determine the effect of each feature alignment.



	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12	Full Feature
Precision	36.71%	54.52%	39.95%	40.17%	39.18%	39.54%	39.58%
Recall	48.69%	51.03%	56.13%	60.73%	50.90%	61%	61.18%
F1 Score	40.70%	51.02%	45.41%	47.01%	42.63%	46.58%	46.67%

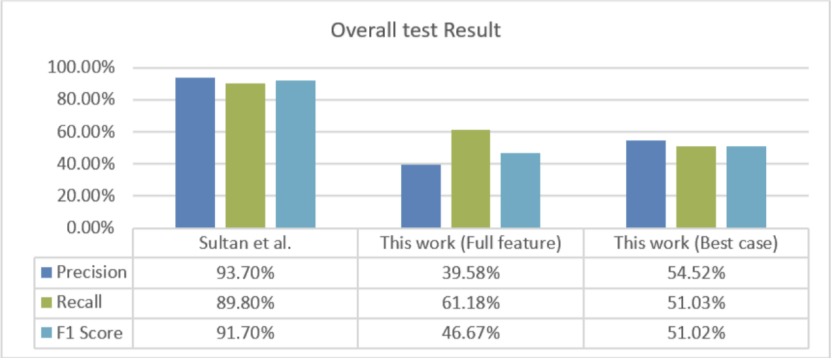
Fig. 3. Second Scenario Testing

Based on the second test (figure 3), by testing the combination of 5 features in the system (case 7 Sequence + PPDB + Neighbor + Stopwords + PPDB Ext, case 8 Identtical Word + PPDB + Neighbor + Stopwords + PPDB Ext, case 9 Identtical Word + Sequence + Neighbor + Stopwords + PPDB Ext, case 10 Identtical Word + Sequence + PPDB + Stopwords + PPDB Ext, case 11 Identtical Word + Sequence + PPDB + Neighbor + PPDB Ext, case 12 Identtical Word + Sequence + PPDB + Neighbor + Stopwords) with the highest F1 score of 51.02% in case 8.

Case 8 became the best combination of features in the second scenario compared to other features with f1 score of 51.02%, in this combination of features do not use the identical word feature that makes the system do not need to find the same word pair between words in the first version translation and the second version so that the f1 score can be 4.35% higher than full feature. Case 7 became the lowest feature combination in the second scenario compared to other features with f1 score of 40.70%, in combination this feature does not use word squares align feature which makes the system do not need to look for word pairs that have the same word order with a minimum of 2 word correlations between words in the first version of the translation and the second version of the translation so that the f1 score can be 5.97% lower than the full feature Case 9 can reach f1 score of 45.41% with a combination of features without align PPDB feature which makes the system does not need to search for pairs of words in a suitable dataset based on the words in the Paraphrase database used so that the f1 score can be 1.16% lower than full feature Case 10 can reach f1 score of 47.01% with a combination of features without the align neighbor content word feature that makes the system do not need to look for pairs of neighboring words from the word in the first version of the translation that will be aligned with the pair of neighbor words from the word on the second version of the translation that succeeded in cross-producting between these words so that the f1 score can be 0.34% higher than the full feature. Case 11 can reach f1 score of 46.58% with a combination of features without the align stopwords feature that makes the system do not need to search for stopword non-content pairs so that the f1 score can be 0.09% lower than full feature. Case 12 can reach f1 score of 45.41% with a combination of features without extended PPDB align feature which makes the system does not need to search for pairs of words in a suitable dataset based on the words in the extended database Paraphrase used so that the f1 score can be 0.76% lower of full features.

4.3 Overall Test Result

The system built in this study produces a not so high f1 score of 46.67% when using the whole feature, whereas if using feature combination align Identical Word, align ppdb, align neighbor, align stopword, and align ppdb the extended f1 score increases to 51,02% for this dataset as the best result. This combination without using align sequences which means the system does not identify a word pair that has the same word order with at least two words in common.



	Sultan et al.	This work (Full feature)	This work (Best case)
Precision	93.70%	39.58%	54.52%
Recall	89.80%	61.18%	51.03%
F1 Score	91.70%	46.67%	51.02%

Fig. 4. Overall Test Result

However, this result is still far from the baseline, which is 91.7%, this is because there are significant differences that affect the measurement results, in the baseline the features used are word identicality, align contextual evidence, align depedeencies, align text neighbors , and align stopwords, and align PPDB while in this study use the align identical word feature, align sequence, align text neighbor, align stopword, align PPDB and in addition specifically align extended PPDB. In the baseline using the

MSR-RTE dataset that was built based on the results of an accurate research, while in this study a dataset constructed of 2 translated versions of the Al-Quran verses formed individually by the authors followed the MSR dataset writing format. Another difference is the addition of extended PPDB as a phrase dictionary specifically for the Al-Quran dataset.

5 Conclusion

Based on the results of the experiment we found that the feature combination scenario in case 8 was the optimum scenario with F1 score of 51.02%. These results mean that the translation version of Muhammad Sarwar and the Wahidduddin Khan version of the translation have a similarity value of 51.02% between the words for the sample dataset used in this study. This value can change when using another translation pair as a sample of the dataset. The results of this study can still be improved by adding a more diverse feature alignment.

References

1. R. Bar Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor.: The second pascal recognising textual entailment challenge (2006).
2. C. Brockett,: Aligning the rte 2006 corpus. Microsoft Research (2007).

3. J. Euzenat.: Semantic precision and recall for ontology alignment evaluation. In: IJCAI, vol. 7, pp. 348353 (2007).
4. J. Ganitkevitch, B. Van Durme, and C. Callison-Burch.: Ppdb: The paraphrase database. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 758–764 (2013).
5. B. MacCartney, M. Galley, and C. D. Manning.: A phrase-based alignment model for natural language inference. In: Proceedings of the conference on empirical methods in natural language processing, pp 802–811. Association for Computational Linguistics (2008).
6. M. Song.: Handbook of research on text and web mining technologies. IGI global (2008).
7. M.A.Sultan, S.Bethard, and T.Sumner.: Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. Transactions of the Association for Computational Linguistics, vol. 2, pp. 219–230 (2014).
8. M. A. Sultan, S. Bethard, and T. Sumner.: Sentence similarity from word alignment. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp. 241–246 (2014).
9. M. A. Sultan, S. Bethard, and T. Sumner.: Sentence similarity from word alignment and semantic vector composition. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) pp. 148–153 (2015).