Analisis dan Implementasi Perhitungan *Semantics Similarity* Pada Ayat Al-Quran Dengan Pendekatan *Word Alignment* Berdasarkan *Support Vector Regression*

Analysis and Implementation Semantics Similarity Measurement Of Al-Quran Verses with Word Alignment Approach Based on Support Vector Regression

Agung Wardhana Z. Nasution[1], **Moch. Arif Bijaksana, Ph.D.**[2], and **Said Al Faraby, S.T., M.Sc.**[3]

[1,2,3] Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom.

[1]dhanasution@student.telkomuniversity.ac.id [2]arifbijaksana@telkomuniversity.ac.id

[3]saidalfaraby@telkomuniversity.ac.id

Abstrak

Al-Quran adalah kitab suci yang menjadi pedoman hidup bagi umat islam. Pada Al-Quran terdapat pengulangan ayat yang sama pada ayat lain. Salah satu cara untuk memahami Al-Quran adalah dengan mencari kesamaan dan keterkaitan antar ayat. Oleh karena itu, diperlukan penelitian yang dapat menilai kesamaan antar ayat dengan ayat lainnya. Salah satu penelitian dalam penyejajaran kata-kata yang memiliki kesamaan adalah *word alignment*. *Word alignment* memperhatikan kesamaan konteks dalam penyejajaran berdasarkan *identical word sequence, named entities, word dependency* dan *surrounding words*. Pada penelitian ini dilakukan penambahan *database* parafrase yang berhubungan dengan Al-Quran. Selain itu dalam penyejajaran ayat dapat dilakukan dengan merepresentasikan kedalam bentuk vektor dengan menggunakan *word2vec*. Untuk pengukuran nilai kemiripan berdasarkan vektor dapat menggunakan perhitungan *cosine similarity* [1]. Evaluasi yang dilakukan menggunakan *Support Vector Regression* (SVR) untuk mengukur nilai prediksi data pasangan ayat Al-Quran terjemahan bahasa Inggris berdasarkan *alignment* dan *word2vec*. Penggunaan metode *word alignment*, *word2vec* berdasarkan SVR pada penelitian ini menghasilkan nilai *pearson correlation* 0,81221.

Kata kunci: Al-Quran , kesamaan semantik, *word alignment*, *word2vec*, svr, *pearson correlation*.

Abstract

Qur'an is a holy book that became the guide of life for Muslims. In the Qur'an there is a repetition of the same verse in another verse. One way to understand the Qur'an is to seek similarity and interrelationship between verses. Therefore, it is necessary research that can assess the similarity between verses with other verses. One of the research in aligning words that have similarities is word alignment. Word alignment aligns words based on contextual similarities of identical word sequences, named entities, word dependencies and surrounding words. The research also added a paraphrase database related to the Qur'an. In addition, alignment can be done by representing a sentence into a vector form using word2vec. For the measurement of semantic vectors can use cosine similarity calculation [1]. Evaluation was performed using Support Vector Regression (SVR) to measure the prediction value of value of Quranic verses pair English translation based on alignment and word2vec. The use of word alignment method, word2vec based on SVR in this research resulted pearson correlation 0,81221.

Keywords: Qur'an, semantic similarity, word alignment, word2vec, svr, pearson correlation.

1 Pendahuluan

Al-Quran adalah kitab suci yang menjadi sumber dan pedoman hidup bagi umat islam. Al-Quran yang terdiri dari 30 Juz, 114 Surat dan 6236 ayat umumnya merupakan kalimat pendek. Pada Al-Quran juga terdapat makna yang sama atau terkait namun tersebar dalam beberapa surat atau juz. Salah satu cara untuk memahami Al-Quran adalah dengan mencari kesamaan dan keterkaitannya agar diperoleh kandungan informasi yang lengkap. Sulit bagi sistem dalam mencari kesamaan ayat Al-Quran, karena tidak memiliki kemampuan intuisi seperti manusia. Oleh karena itu, diperlukan sistem yang dapat menilai kesamaan antar ayat atau kalimat yang disebut *Semantic Textual Similarity*.

Semantic Textual Similarity disingkat STS merupakan konsep yang mengukur kesamaan makna semantik berdasarkan potongan teks yang berpasangan dengan algoritma dan model komputasi yang meniru kinerja manusia. STS banyak digunakan dalam penerapan ilmu *Natural Language Processing* (NLP) dan *text mining* [2]. Sebagai wadah untuk mendukung para peneliti bidang komputasi *linguistic* dalam mengembangkan pendekatan baru STS dilaksanakan *Semantic Evaluation* disingkat SemEval. Salah satu metode yang sering digunakan dalam penyelesaian *task-task* tersebut adalah *word alignment based similarity*.

Word alignment based similarity adalah metode yang dikembangkan oleh Sultan et al untuk penyejajaran kata-kata dalam kalimat yang sama dan diukur kesamaannya berdasarkan identifikasi urutan dan posisinya. Dalam penelitian SemEval 2015 dan 2016, fitur *alignment* dikombinasikan dengan fitur lain dengan tujuan untuk menghasilkan model yang lebih baik. Salah satu fitur yang dikombinasikan adalah model perhitungan kata dalam dokumen yang disebut TF-IDF. *Word alignment* pada dasarnya mengandalkan *paraphrase* dalam penyejajaran kata. Kelemahan dalam penyejajaran kata kata yang bukan *paraphrase* ini dapat diatasi dengan merepresentasikan kata kedalam bentuk vektor [1] yang dikenal dengan istilah vektor semantik.

Penelitian ini akan melakukan analisa dan implementasi perhitungan nilai semantik kemiripan kata pada ayat Al-Quran terjemahan bahasa Inggris menggunakan pendekatan *word alignment* dan vektor semantik *word2vec*. sebagai evaluasi dari penelitian ini digunakan model *Support Vector Regression* (SVR). Model SVR mencoba memprediksi tingkat kesamaan semantik antar kata atau *phrase* dengan asumsi bahwa hal tersebut dapat diwakili oleh probabilitas *annotator* (manusia) secara acak yang telah meng-*annotate* (membubuhi) keterangan pasangan *paraphrase* [3]. Hasil anotasi tersebut dijadikan *gold standard*, nilai acuan penelitian dengan skala 0 sampai 5.

Pendekatan *word alignment*, vektor semantik dan model regresi SVR ini dipilih karena merupakan salah satu metode yang diterapkan oleh Thomas Brychcin et al dari tim UWB yang meraih peringkat kedua pada kompetisi SemEval 2016 [1].

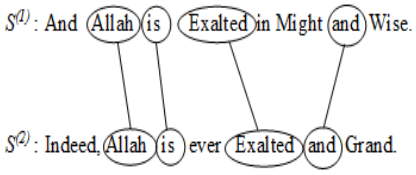
2 Dasar Teori

2.1 *Semantic Textual Similarity*

Semantik merupakan ilmu kebahasaan yang mempelajari tentang makna, tanda baca, dan struktur dalam kalimat yang menentukan pemahaman pembaca terhadap kalimat. *Semantic Textual Similarity* merupakan suatu konsep yang dapat mengukur kesamaan makna dalam konteks teks pendek. Teks yang dibandingkan dapat berupa kata, kalimat pendek, dan sebuah dokumen [2]. Penelitian tentang STS terus berkembang seperti pada kompetisi tahunan di bidang komputasi linguistik, SemEval (*Semantic Evaluation*) pada *task* STS.

2.2 *Word Alignment Based Similarity*

Word alignment merupakan metode *unsupervised learning* dalam bidang pemrosesan bahasa alami (NLP) yang berfungsi sebagai identifikasi keterhubungan makna antarkata pada kalimat yang berbeda, baik secara makna maupun arti. Antarkata sepasang kalimat dihubungkan dengan garis untuk kata-kata yang memiliki relasi. Dengan metode ini, sistem akan menentukan pola kesejajaran *alignment* yang cocok dari kalimat yang diinputkan. Gambar 1 dibawah ini merupakan contoh penggunaan *word alignment* dalam menyejajarkan kata.



Gambar 1: Contoh *Word Alignment*

[1] http://alt.qcri.org/semeval2016/task1/index.php?id=results

2.3 *Contextual Similarity*

Untuk melihat kesamaan konteksnya digunakan *alignment* berdasarkan *identical word sequence, named entities, word dependency* dan *surrounding words*.

2.3.1 *Identical Word Sequence*

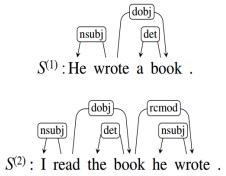
Alignment terhadap urutan kata yang sama pada kedua buah kalimat yang mengandung minimal dua kata berurutan yang sama persis atau identik.

2.3.2 *Named Entities*

Alignment terhadap kata atau kalimat yang merupakan untuk menentukan *head word* dari kata tersebut apakah kata tersebut merupakan nama tempat, kota, nama objek dan sebagainya.

2.3.3 *Word Dependency*

Alignment berdasarkan *dependency* dari masing-masing kata atau kalimat. Hal yang terlebih dahulu dilakukan adalah proses POS *Tagging* dan ekstraksi *contextual evidence*. Dicontohkan seperti pada Gambar 2 dibawah ini.



Gambar 2: Contoh *Dependency Equivalence* [4]

2.3.4 *Surrounding Words*

Alignment berdasarkan sifat ketetanggaan kata yang mendefenisikan konteks dari sebuah kata dalam kalimat sebagai *fixed window*.

2.4 *Paraphrase Database*

PPDB adalah *database* leksikal yang berisi parafrase dan parafrase sintaksis. Parafrase adalah dua kata atau lebih yang memiliki makna sama (bisa menggantikan). Misalkan kata *beautiful* memiliki parafrase *good looking*. Pada penelitian ini dilakukan penambahan PPDB *Extended*. PPDB yang ditambahkan berisi tentang kemiripan kata berkaitan dengan Al-Quran terjemahan bahasa Inggris.

2.5 *Term Frequency-Invers Document Frequency (TF-IDF)*

Metode *Term Frequency-Invers Document Frequency* merupakan metode pembobotan suatu kata *(term)* terhadap dokumen. Dalam penggunaannya, metode ini menggabungkan konsep *Term Frequency* (TF) dan *Document Frequency* (DF). *Term frequency* merupakan metode pembobotan yang menghitung jumlah kemunculan suatu kata *(term)* dalam suatu dokumen. *Invers Document Frequency* adalah metode pembobotan yang menghitung jumlah kemunculan suatu kata *(term)* pada setiap dokumen. Bobot dari suatu dokumen dapat dihasilkan melalui perkalian hasil perhitungan TF dan IDF yang dirumuskan dalam persamaan dibawah ini.

$$w(t,s) = tf(t,s) \times idf(t,D) \tag{1}$$

2.6 Vektor Semantik

Vektor semantik atau *word embedding* adalah salah satu cara untuk merepresentasikan struktur kalimat yang akan di *align* dengan memanipulasi kalimat kedalam bentuk vektor. Terdapat cara lain untuk merepresentasikan vektor semantik yaitu dengan *word2vec*. *Word2vec* adalah salah satu *toolkit* yang digunakan untuk ekstraksi vektor menjadi 2,8 miliar token dari *corpus* dan merepresentasikan makna dari suatu kata. Pada *word2vec* kata-kata yang secara semantik mirip terkadang letaknya berdekatan. Oleh karena itu, *embedding* yang bagus dalam melakukan prediksi *neighboring* juga bagus dalam merepresentasikan *similarity* [5].

2.7 *Support Vector Regression*

SVR merupakan penerapan dari s*upport vector machine* (SVM) untuk kasus regresi. SVM merupakan kumpulan teknik klasifikasi dan regresi yang merupakan pengembangan algoritma nonlinear (Bermolen, 2008). SVR dapat digunakan untuk regresi bilangan rill maupun kontinu. Metode ini dapat mengatasi masalah *overfitting*, sehingga menghasilkan performa yang lebih baik (Smola dan Scolkoph, 2004).

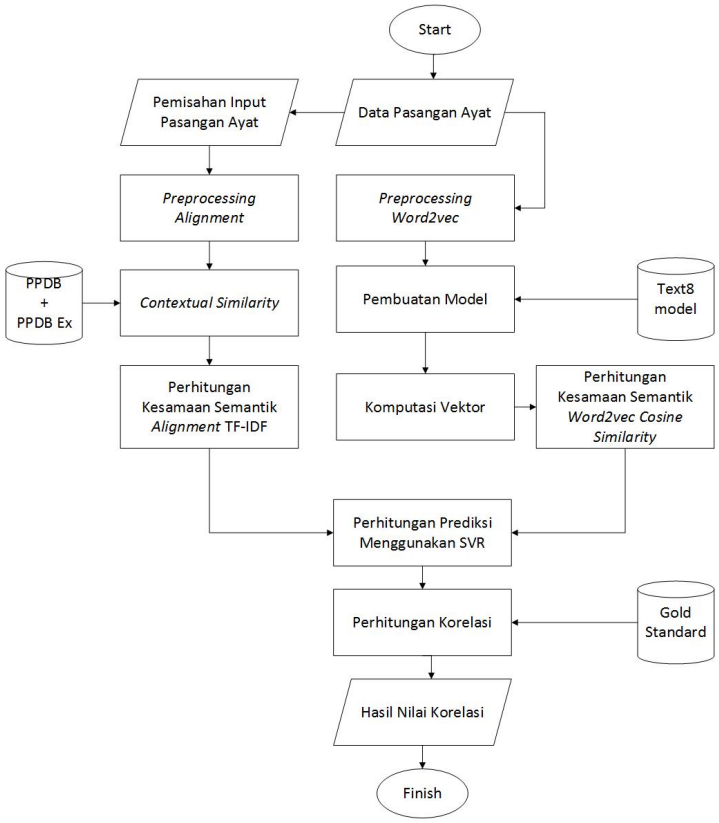
2.8 STS Al-Quran

Qursim merupakan salah satu penelitian tentang teks Al-Quran. Penelitian ini berkaitan dengan pembuatan *database* kebahasaan yang dikenal dengan *corpus* untuk evaluasi kesamaan, keterkaitan teks pendek pada kitab suci ini. Data yang digunakan merupakan eksplorasi dari tafsir Ibnu Katsir.

3 Perancangan Sistem

3.1 Gambaran Umum Sistem

Sistem yang dibangun bertujuan untuk menghasilkan nilai kesamaan semantik antar dua pasangan ayat Al-Quran. Nilai kesamaan semantik diperoleh berdasarkan pengimplementasian metode *word alignment*, vektor semantik dan menggunakan perhitungan regresi *support vector regression*. Evaluasi yang dilakukan adalah dengan menggunakan perhitungan korelasi sebagai tolak ukur kemiripan sistem yang dibangun dengan nilai *gold standard*. Gambaran umum sistem yang dibangun dapat dilihat pada Gambar 3.



Gambar 3: Gambaran Umum Sistem

Dari Gambar 3 diatas maka tahapan gambaran umum sistem adalah sebagai berikut:

1. Sistem membaca data input pasangan ayat Al-Quran terjemahan bahasa Inggris kemudian memisahkan ayat satu dan ayat dua menjadi kumpulan file data input yang terpisah dengan format .txt.
2. Sistem melakukan tahapan *preprocessing alignment*, yaitu dengan melakukan tahapan tokenisasi, lemmatization terhadap setiap file input ayat satu dan ayat dua.
3. Sistem melakukan *contextual Similarity* yaitu melakukan proses identifikasi *alignment* berdasarkan *identical word sequences, named entities, dependencies* dan *surrounding content words* terhadap file inputan. Proses identifikasi ini menggunakan bantuan database PPDB dan PPDB *Extended*.

4. Sistem melakukan perhitungan nilai kesamaan semantik *alignment* dengan menggunakan metode TF-IDF terhadap data input pasangan ayat Al-Quran terjemahan. Nilai kesamaan semantik disimpan dalam bentuk *list* dan dioutputkan kedalam file format *.txt*.

5. Sistem melakukan tahapan *preprocessing word2vec*, yaitu dengan melakukan tahapan tokenisasi dan *stemming* terhadap setiap file input ayat satu dan ayat dua.

6. Sistem melakukan tahapan pembuatan model dengan bantuan *corpus text8*. Model ini digunakan untuk mengetahui nilai vektor pada setiap input ayat satu dan ayat dua.

7. Sistem melakukan komputasi vektor yaitu mengubah kata menjadi vektor berdasarkan model yang telah dibangun.

8. Sistem melakukan perhitungan nilai kesamaan semantik *word2vec* dengan menggunakan *cosine similarity* dari perhitungan komputasi vektor. Nilai kesamaan semantik disimpan dalam bentuk *list* dan dioutputkan kedalam file format *.txt*.

9. Sistem melakukan tahapan evaluasi dengan menggunakan perhitungan regresi *support vector regression*. Dibutuhkan data *train* dan data *test* dalam perhitungannya. data *train* yang digunakan merupakan kumpulan data pasangan ayat terjemahan bahasa Inggris yang dilakukan oleh penelitian sebelumnya dengan jumlah 400 pasangan ayat terjemahan. Sedangkan, data *test* yang digunakan merupakan kumpulan pasangan ayat terjemahan yang dilakukan oleh peneliti dengan jumlah 400 pasangan ayat terjemahan.

10. Sistem melakukan perhitungan prediksi data *test*. Parameter yang digunakan dalam perhitungan regresi ini adalah *SVR(C=1,0, cache size=200, coef0=0,0, degree=3, epsilon=0,1, gamma='auto', kernel='rbf', max iter=1, shrinking=True, tol=0,001, verbose=False)*. Nilai prediksi disimpan dalam bentuk *list* dan dioutputkan kedalam file format *.txt*.

11. Sistem melakukan perhitungan korelasi dari nilai prediksi *support vector regression* dibandingkan dengan nilai *gold standard*.

3.2 Pengumpulan Data

Pengumpulan data pasangan potongan ayat Al-Quran dan data *gold standard* yang digunakan sebagai *knowledge base* pada penelitian ini. Data yang digunakan adalah data pada penelitian tahun lalu oleh Dwi Jayanti Wulandari. Data tersebut berjumlah 400 data ayat terjemahan Bahasa Inggris. Terbagi menjadi dua jenis berdasarkan pengumpulan datanya yaitu data Qursim hasil penelitian keterkaitan *corpus* Al-Quran berdasarkan tafsir Ibn katsir oleh Abdul-Baquee M. Sharaf dan Eric S. Atwell sebanyak 200 data, dan data Indeks Tematik Kementrian Agama Republik Indonesia dan Pusat Kajian Hadist Al-Mughni sebanyak 200 data. penulis juga menambahkan data pasangan ayat Al-Quran sebanyak 350 pada data Qursim dan 50 data ayat Indeks Tematik. Total data yang dikumpulkan menjadi 800 pasangan ayat. Contoh data yang digunakan seperti pada Tabel 1.

Tabel 1: Contoh Data Pasangan Terjemahan Al-Quran Bahasa Inggris

Data Qursim Tafsir Ibn Katsir		
Ayat 1	:	and also those who spend of their wealth to be seen by the people and believe not in Allah nor in the last day. (4:38)
Ayat 2	:	never will you attain the good reward until you spend in the way of Allah from that which you love. (3:92)
Data Indeks Tematik		
Ayat 1	:	The punishment will not be lightened for them nor will they be reprieved. (2:162)
Ayat 2	:	The punishment will not be lightened for them nor will they be reprieved. (3:88)

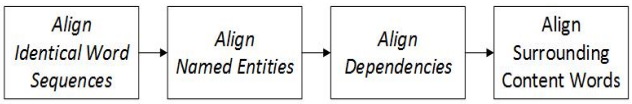
3.3 Rancangan Sistem

3.3.1 *Preprocessing Alignment*

Dalam pengimplementasiannya, data input yang digunakan adalah data *input* potongan ayat Al-Quran terjemahan bahasa Inggris yang sudah terpisah antara kalimat pertama Ayat 1 dan kalimat kedua Ayat 2. Selanjutnya dilakukan tahapan *preprocessing* yaitu tokenisasi, *parsing, lemmatization, Part Of Speech Tagging* (POS *Tagging*).

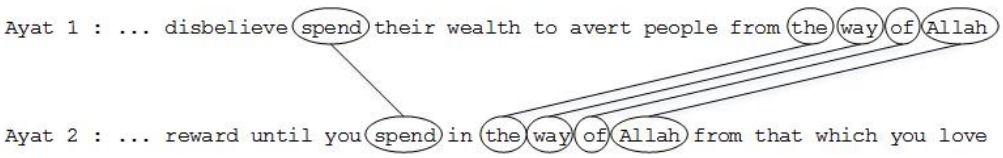
3.3.2 *Alignment Based Similarity*

Alignment Based Similarity dibangun untuk mengidentifikasi nilai kesamaan semantik dan mengetahui performansi sistem terhadap data pasangan potongan ayat Al-Quran terjemahan Bahasa Inggris. Fitur *alignment* yang dilakukan pada tugas akhir ini merujuk kepada metode Sultan et al. (2014) [4]. Urutan proses *alignment* dapat dilihat pada Gambar 4 dibawah ini.



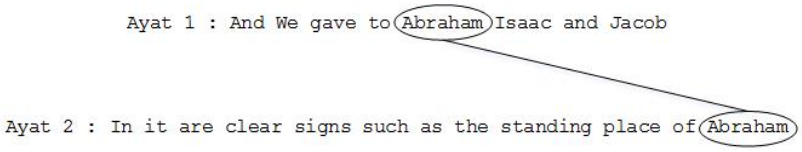
Gambar 4: Tahapan *Alignment Based Similarity*

a. *Align Identical Word Sequences*
Identical word sequences merupakan bentuk sederhana dari *contextual evidence* ini. Mengukur kemiripan antara pasangan kalimat secara sederhana dapat dilihat pada kata-kata yang identik dan urutannya dari kedua kalimat. *String* kata yang mirip akan di *align* seperti ilustrasi potongan ayat (Q.S. Al-Anfal 8:36) dan (Q.S. Al-Imran 3:92) pada Gambar 5 dibawah ini.



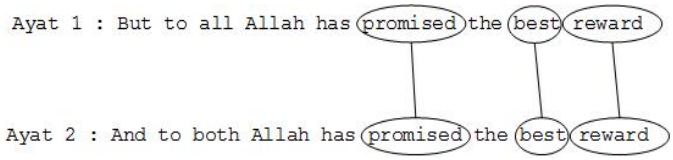
Gambar 5: Ilustrasi *Align Identical Word Sequences*

b. *Align Named Entities*
Align Named Entities dilakukan untuk mensejajarkan kata-kata berdasarkan entitas-entitas nama yang dimiliki pada suatu kalimat. Entitas-entitas yang dikenali berupa nama *organization, location,* dan *person*. Contoh *align named entities* terdapat pada potongan ayat (Q.S. AlAnkabut 29:27) dan (Q.S. Al-Imran 3:97) sesuai Gambar 6 dibawah ini.



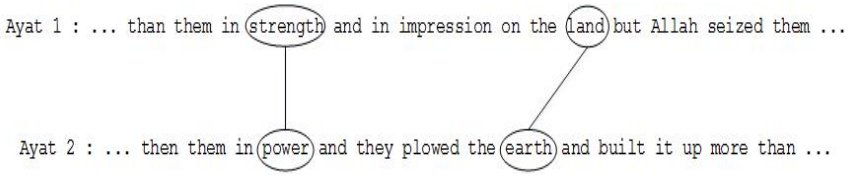
Gambar 6: Ilustrasi *Align Named Entities*

c. *Align Dependencies*
Dependency merupakan salah satu *contextual evidence* yang penting dalam pensejajaran kalimat maupun kata. *Dependency type* yang bisa di *align* harus memenuhi syarat antara kedua kata harus merupakan anggota dari struktur *equivalent dependency* yang bersifat *lexical* terdiri dari *verbs, nouns, adjectives, adverbs* [6]. Contoh penerapan *align dependencies* dapat dilihat pada potongan pasangan ayat (Q.S. Al-Hadid 57:10) dan (Q.S. An-Nisa 4:95) dibawah ini sesuai Gambar 7.



Gambar 7: Ilustrasi *Align Dependencies*

d. *Align Surrounding Content Words*
Proses *align* menggunakan *evidence* ini dilakukan dengan memeriksa kata-kata di sekitarnya atau *neighborhood*. Penelitian ini menerapkan pengambilan 3 kata sebelah kiri dan 3 kata sebelah kanan dari kata yang akan di *align* sesuai Sultan et al. [6]. Contoh *align surrounding content words* terdapat pada potongan ayat (Q.S. Ghafir 40:21) dan (Q.S. Ar-Rum 30:9) sesuai Gambar 8 dibawah ini.



Gambar 8: Ilustrasi *Align Surrounding Content Words*

3.3.3 Perhitungan Kesamaan Semantik *Align*

Setalah dilakukan proses *alignment* dari pasangan potongan ayat Al-Quran, dilakukan perhitungan kemiripan dengan menggunakan metode TF-IDF. Pertama, menghitung banyaknya TF dan DF dari suatu kalimat pada setiap dokumen input pertama dan kedua. TF dihitung berdasarkan frekuensi kemunculan kata dalam suatu kalimat. Sedangkan DF dihitung dari banyaknya suatu kata pada suatu dokumen. Jumlah DF pada kata akan bertambah bila ditemukan pada kalimat berikutnya.

3.3.4 *Preprocessing Word2vec*

Tahapan *preprocessing word2vec* dilakukan terhadap data input pasangan ayat Al-Quran terjemahan Bahasa Inggris. Data *input* tersebut lalu dilakukan proses tokenisasi dan *stemming*.

3.3.5 Proses Vektor Semantik

Tahapan yang dilakukan dalam pembangunan model adalah *load* dan *unzip corpus text8* yang dikumpulankan dan dikembangkan oleh Matt Mahoney. Vektor yang dihasilkan berupa angka kedekatan kata tersebut dengan kata yang lain yang terdapat dalam model *text8.bin*. Data tersebut kemudian dilakukan tokenisasi setiap ayatnya. Setiap kata pada hasil token kemudian diubah menjadi vektor dengan memeriksa nilai *array* matriks setiap kata pada model *text8.bin*. Matriks *array* yang dihasilkan setiap token tersebut kemudian dijadikan dalam satu vektor pada setiap ayat. Sehingga terdapat 2 vektor yang dihasilkan. Untuk menghitung kesamaan semantik pada kedua vektor digunakanlah *cosine similarity*.

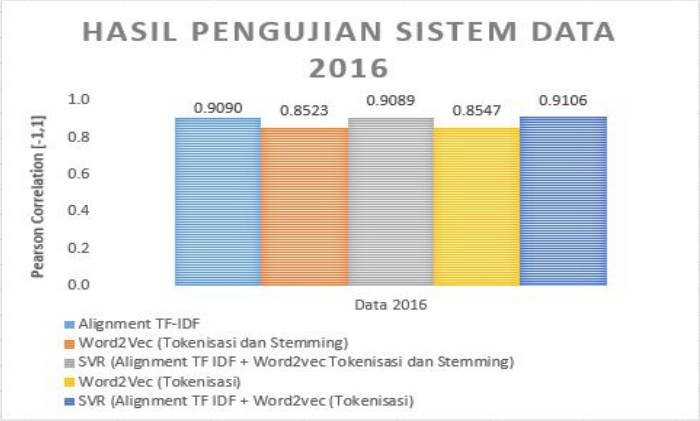
3.4 Evaluasi

Setelah hasil dari fitur *alignment* dan vektor semantik dikumpulkan, maka dilakukan evaluasi kombinasi kedua fitur dengan menggunakan regresi. Pada penelitian ini regresi yang digunakan adalah SVR *library scikit-learn* dengan parameter *SVR(C=1,0, cache_size=200, coef0=0,0, degree=3, epsilon=0,1, gamma='auto', kernel='rbf', max_iter=1, shrinking=True, tol=0,001, verbose=False)* merujuk kepada [7]. Data yang digunakan dalam evaluasi menggunakan SVR yaitu hasil kesamaan *alignment* dan *word2vec* pada data *train* sebagai acuan prediksi untuk data *test*. Hasil prediksi data *test* tersebut dilakukan perhitungan korelasi dengan membandingkan terhadap nilai *gold standard* menggunakan *pearson correlation*.

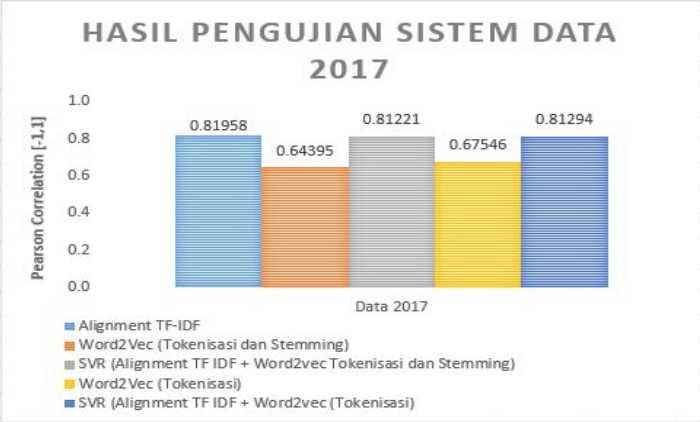
4 Pembahasan

4.1 Hasil Pengujian

Dari sistem *alignment* dan *word2vec* yang dibangun, diperoleh hasil pengujian pada Gambar 9 dan Gambar 10. Sistem yang menggunakan data 2017 sebagai data *test* dengan inputan 350 ayat terjemahan Bahasa Inggris Ibnu Katsir dan 50 ayat indeks tematik memperoleh nilai korelasi *alignment* 0,81958, nilai *word2vec* 0,64395 dan nilai korelasi SVR yang diperoleh 0,81221. Nilai SVR ini, memiliki nilai lebih kecil 0,00737 dibandingkan *alignment* TF-IDF dikarenakan kompleksitas pada masing-masing data dan pengaruh fitur *word2vec*. hal tersebut tidak terjadi jika data *test* menggunakan data 2016 dengan fitur tanpa *stemming* pada *word2vec* menghasilkan nilai SVR 0,9106 lebih besar 0,0016 dari hasil *alignment*.



Gambar 9: Hasil Pengujian Sistem Menggunakan data Alquran 2016



Gambar 10: Hasil Pengujian Sistem Menggunakan data Alquran 2017

Pada Tabel 2 dapat dilihat bahwa hasil *alignment* yang diperoleh untuk sistem bernilai lebih tinggi 0,0094 dari sistem *alignment* Sultan et al. dan lebih tinggi 0,02463 dari sistem yang dikembangkan Dwi Jayanti Wulandari. Sama halnya dengan dataset 2017 yang dikumpulkan peneliti.

Tabel 2: Perbandingan Hasil Data Menggunakan *Alignment*

Sistem	*Pearson Correlation* [-1,1]	
	Data 2016	Data 2017
This Work (TF-IDF)	0,90902	0,81958
Sultan et al.	0,89962	0,80349
Dwi Jayanti Wulandari	0,88439	-

Tabel 3: Perbandingan Hasil Data Menggunakan Semua Fitur Penelitian ini

Sistem	Data 2017
This Work (All System)	0,81221
Sultan et al.	0,80349

Pada Tabel 3 nilai sistem yang diperoleh menggunakan semua fitur pada penelitian ini jika dibandingkan dengan sistem yang dibangun oleh Sultan et, al. menghasilkan nilai korelasi yang lebih tinggi.

4.2 Analisis *Alignment* dan Pola Pasangan Terjemahan Ayat Al-Quran

Pada penelitian ini dilakukan penambahan PPDB *Extended* menyesuaikan dengan data pada penelitian ini yaitu data Al-Quran terjemahan bahasa Inggris. Hasil *alignment* dengan menggunakan penambahan PPDB *Extended* mengalami peningkatan 0,05291 dari *alignment* jika hanya menggunakan PPDB saja. Hal ini dikarenakan ada beberapa kata yang tidak di *align* jika menggunakan PPDB saja. Jika dilakukan pengecekan terhadap *database* PPDB, kata `Lord` tidak disandingkan dengan kata `Allah`. Permasalahan ini dapat ditangani oleh PPDB *Extended*.

Dilihat dari fitur *identical word sequence* tidak terjadi kesalahan. Namun, fitur *named entity* tidak sepenuhnya dapat mengidentifikasi *tag* kata kedalam bentuk *tag* sebenarnya. Perhitungan TF-IDF dalam fitur *alignment* memiliki beberapa keuntungan yaitu, mudah dihitung kemiripannya karena dokumen di ekstraksi kedalam bentuk yang deskriptif. Ukuran dokumen pada perhitungan TF-IDF akan sangat berpengaruh dalam mengukur nilai kesamaan.

Sistem yang dibangun, dapat dengan mudah mengidentifikasi kalimat pasangan kalimat yang banyak memiliki kesamaan kata atau makna dengan baik. Kebanyakan pasangan kalimat yang tidak dapat dinilai kesamaannya adalah kalimat-kalimat yang secara bentuk tidak sama dan tidak ada dalam *database* PPDB. Data yang sulit untuk di identifikasi dalam penelitian ini adalah pasangan kalimat yang memiliki nilai kesamaan semantik tinggi tetapi pada pasangan kalimatnya hanya memiliki sedikit pasangan kata yang sama secara penulisan atau makna.

4.3 Analisis *Preprocessing* Vektor Semantik *Word2vec* dan Penggunaannya

Pada penelitian ini dilakukan analisis penggunaan *preprocessing* yang digunakan. Tahapan yang digunakan adalah tokenisasi dan *stemming*. Nilai korelasi yang dihasilkan menggunakan hanya *preprocessing* tokenisasi lebih tinggi 0,0308 dari *preprocessing*menggunakan tokenisasi dan *stemming*. Hal ini disebabkan terdapat perbedaan *array* vektor dari kalimat yang sudah di *stemming*, sehingga mempengaruhi nilai kesamaan vektor. Pasangan kalimat yang memiliki nilai kesamaan vektor tinggi adalah pasangan kalimat dengan kata-kata yang terdapat dalam model vektor. nilai korelasi dataset 2017 yaitu 0,64395 mengalami penurunan karena kompleksitas pasangan ayat yang tidak dapat diidentifikasi sistem. Sistem dengan model yang dibangun tidak sepenuhnya dapat mengidentifikasi kata-kata serapan yang berhubungan dengan Al-Quran ke dalam bentuk vektor, seperti kata `alMasjid, alHaram, tawaf, fitnah`. Ketidakmampuan identifikasi tersebut dikarenakan dimensi model *corpus* yang digunakan berukuran kecil.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan dari penelitian yang telah dilakukan adalah:

1. Nilai korelasi dari sistem *word alignment* dan *word2vec* berdasarkan regresi *support vector regression* (SVR) lebih besar daripada menggunakan sistem *alignment* yang dibangun oleh Sultan et, al. namun lebih rendah jika dibandingkan dengan nilai *alignment* menggunakan perhitungan TF-IDF.

2. Nilai korelasi dari sistem *word alignment* dengan PPDB *Extended* lebih tinggi daripada tanpa menggunakan PPDB Extended.

3. Nilai kesamaan vektor lebih tinggi jika hanya menggunakan *preprocessing* tokenisasi dibandingan dengan menggunakan *preprocessing* tokenisasi dan *stemming*.

4. Nilai kesamaan *word alignment* menggunakan TF-IDF pada penelitian ini menghasilkan nilai korelasi yang lebih baik dan mendekati penilaian manusia, dibandingkan penelitian sebelumnya.

5.2 Saran

Untuk penelitian yang lebih baik dari sistem ini dalam mengidentifikasi kesamaan kalimatada beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya, yaitu:

1. Mencoba membangun dan menggunakan metode perhitungan kesamaan *align* dan vektor lain untuk meninggatkan nilai korelasi.
2. Mencoba menambahkan parafrase yang berhubungan dengan Al-Quran.
3. Mencoba membangun model dimensi vektor dengan ukuran *corpus* yang lebih besar.
4. Membandingkan perhitungan regresi dengan regresi lain, seperti *Linier Regression, Gaussian Regression* dan model lainnya.

Daftar Pustaka

[1] Wu Hao, Huang Heyan, and Lu Wenpeng. Sentence Similiarity Based on Alignment and Vector with Weight of Information Content. *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1225–1259, 2015.

[2] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. A pilot on semantic textual similarity. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, 2012.

[3] R.M. Karampatsis. SemEval-2015 CDTDS: Predicting Paraphrases in Twitter via Support Vector Regression. *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics*, pages 75–79, 2015.

[4] 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) Association for Computational Linguistics*, pages 241–246, 2014.

[5] Daniel Jurafsky and James. H. Martin. Speech and Language Processing. *Stanford University*, pages 802–811, 2015.

[6] M.A. Sultan, S. Bethard, and T. Sumner. Back to basics for monolingual alignment. *Exploiting word similarity and contextual evidence, Association for Computational Linguistics*, pages 219–230, 2014.

[7] Fu.B.A Cheng and H.L.S Xianpei. ”Sentence Similarity Based on Support Vector Regression using Multiple Features. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2016). Association for Computational Linguistics*, pages 642–646.