Analisis dan Implementasi Kesamaan Semantik Antar Teks Menggunakan Pendekatan *Alignment* dan Vektor Semantik pada Terjemahan Alquran

Analysis and Implementation Sematics Text Similiarity using Alignment and Semantic Vector in Translation of the Qur'an

Meiditia Mustika Rani[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]meiditiarani@gmail.com [2]arifbijaksana@gmail.com [3]saidalfaraby@gmail.com

Abstrak

Alquran adalah pedoman hidup umat Islam yang sudah memiliki banyak tafsiran agar mudah untuk dipahami. Untuk mempermudah pencarian kesamaan ayat maka dilakukanlah penelitian pada teks terjemahan Alquran. Banyak metode dalam bidang *text mining* dan NLP yang dapat digunakan untuk mengukur kesamaan semantik antar kalimat, beberapa diantara yaitu menggunakan *alignment* dan vektor semantik. *Alignment* adalah salah satu metode yang dapat menghitung kesamaan semantik dengan melakukan penyejajaran kata-kata. Fitur *alignment* yang digunakan dalam penelitian ini yaitu *identical words*, PPDB, *word sequences*, dan *named entities*. Sementara vektor semantik adalah metode yang digunakan untuk menghitung kemiripan sebuah kata dari distribusi kata-kata di sekitarnya. Agar memperoleh hasil yang lebih baik maka dilakukan kombinasi dari metode *alignment* dengan vektor semantik menggunakan regresi *ridge*. Selain itu, dilakukan juga pembuatan *gold standard* sebagai tolak ukur untuk mengetahui nilai korelasi yang menjadi kontribusi penulis untuk menambahkan data yang sudah ada pada penelitian tahun lalu. Kombinasi metode *alignment* dan vektor semantik untuk menghitung kesamaan semantik pada data terjemahan Alquran tahun 2016 menghasilkan nilai korelasi 0,90467, sementara untuk data terjemahan Alquran yang dikumpulkan pada tahun 2017 menghasilkan nilai korelasi 0,79756.

Kata kunci: *text mining*, nlp, *word alignment*, semantik, vektor semantik, quran, islam, regresi.

Abstract

Alquran is a way of life for Muslims and have many interpretations to make it easy to understand. To facilitate the search for similarity of the verse then conducted research on the text of the translation of the Quran. Many methods in the field of text mining and NLP can be used to measure semantic similarity between sentences, some of it is use alignment and semantic vectors. Alignment is one of method that can calculate semantic similarity by doing alignment of words. Features of alignment used in this research are identical words, PPDB, word sequence and named entity. While semantic vector is the methods used to calculate the similarity of a word from the distribution of adjacent words. For the use of the semantic vector it needs corpus text8 taken from Matt Mahoney. In order to obtain better results, a combination of alignment method and semantic vector using ridge regression was used. In addition, the manufacture of gold standard was made as a benchmark to determine the value of the correlation and become contribution of the authors to add data that already exist in research last year. The combination of alignment and semantic vector methods to calculate semantic similarity in Alquran translation data in 2016 resulted in a correlation value of 0,90467, while for the translation of Alquran data collected in 2017 resulted in a correlation value of 0,79756.

Keywords: *text mining, nlp, word alignment, semantic, semantics vector, quran, moslem, regression.*

1 Pendahuluan

Semantic Textual Similarity (STS) adalah salah satu penelitian yang terkait dengan hubungan semantik antara dua konsep dalam penggunaan dan keterkaitannya [1]. Dalam bidang komputasi dan NLP terdapat kompetisi SemEval yang rutin diselenggarakan setiap tahunnya untuk mendorong peneliti agar menciptakan atau mengembangkan metode baru yang efektif dan efisien. Pada kompetisi tersebut terdapat berbagai Task yang dikelompokan berdasarkan bidangnya, salah satu task yang diperlombakan adalah STS. Pada tahun 2014 Sultan menggunakan metode *alignment* dan berhasil mendapatkan skor akurasi 0,8591[1]. Sementara pada tahun 2015 digunakan metode tambahan yaitu vektor semantik dan memperoleh skor akurasi tertinggi 0,8642[2]. *Alignment* merupakan metode STS dengan mensejajarkan kata dalam kalimat. Untuk mengatasi kekurangan pada *alignment* dalam mengidentifikasi kata non-frasa dibutuhkan vektor semantik. Vektor semantik merupakan metode yang digunakan untuk menghitung sebuah kata dari distribusi kata-kata di sekitarnya. Kriteria data pada penelitian SemEval sama seperti data Alquran. Alquran merupakan kitab suci yang berisikan teks pedoman hidup umat Islam terdiri dari 30 Jus, 144 surat, dan 6236 ayat. Untuk mempermudah mengartikan ayat dalam Alquran sudah banyak dibuat terjemahan dalam berbagai bahasa. Salah satunya adalah Alquran terjemahan bahasa Inggris. Pada tugas akhir ini akan digunakan data terjemahan Alquran bahasa Inggris sebagai input untuk diukur nilai kesamaannya pada sistem STS dengan menggunakan pendekatan *word alignment* dan vektor semantik. Untuk mengukur akurasi sistem STS yang akan dibangun pada penelitian ini dibutuhkan *gold standard* yang berisi penilaian kesamaan antar kalimat berdasarkan intuisi manusia. Hasil sistem kemudian dibandingkan dengan *gold standard* yang kemudian dihitung nilai korelasinya.

2 Dasar Teori

2.1 *Semantic Textual Similarity*

Semantic Textual Similiarity (STS) adalah salah satu cara untuk mengukur kesamaan semantik antar dua teks. Kesamaan diukur berdasarkan makna kata dan kalimat. Adapun penelitian STS dilakukan untuk mengukur kesamaan makna pada suatu teks dengan teks yang lainnya.

2.2 Data STS Alquran

Qursim merupakan penelitian yang melakukan pengumpulan data pasangan ayat Alquran untuk dijadikan data evaluasi berdasarkan keterkaitannya pada tafsir Ibnu Katsir. Data Qursim memiliki nilai *gold standard* untuk setiap pasangannya dengan rentang 0 hingga 2.

2.3 *Word Alignment*

Word Alignment merupakan salah satu metode yang digunakan untuk mengukur kesamaan dan keterkaitan antar teks dengan melakukan proses identifikasi frasa dan kata yang masih mempunyai hubungan arti dan makna pada dua buah kalimat yang berbeda. Fitur-fitur alignment yang digunakan adalah *align identical words, align* PPDB *align word sequences* dan *align named entities*.

2.3.1 *Align Identical Words*

Merupakan tahapan *alignment* dimana terdapat satu kata yang identik antar kedua kalimat. Kata dapat dikatakan identik dapat dilihat dari dua faktor yaitu identik berdasarkan *string* atau berdasarkan kontekstualnya.

2.3.2 *Align PPDB*

Tahap ini menggunakan *paraphrase database* dalam menentukan *alignment*, diadaptasi dari Sultan et al. [2]. Jika kata yang akan di-*align* terdapat dalam PPDB maka pasangan kata tersebut akan dilakukan *align*.

2.3.3 *Align Word Sequences*

Align word sequences digunakan untuk mengidentifikasi pasangan kata yang mempunyai urutan kata yang sama dengan minimal 2 kata yang identik. Jika terdapat kata identik yang mempunyai urutan yang sama antar dua kalimat maka akan dilakukan *align*.

[1] http://alt.qcri.org/semeval2014/index.php?id=evaluation-results
[2] http://alt.qcri.org/semeval2015/task2/index.php?id=results

2.3.4 *Align Named Entities*

Dilakukan *alignment* pada setiap kata antar kedua kalimat dan ditentukan *head word* untuk mengidentifikasi apakah kata berupa nama orang, tempat, kota, negara, objek, perusahaan, dll dengan bantuan *Part-Of-Speech tagging* (POS *tagging*) dan *Named Entity Recognition*.

2.4 Perhitungan Kesamaan Semantik *Alignment*

Dalam penelitian ini menggunakan *single proportion* yaitu perhitungan yang tidak memperhatikan proporsi kata *align* pada suatu kalimat terhadap jumlah konten pasangan kalimat dengan menggunakan persamaan 1.

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

dimana $n_c^a(S^{(x)})$ untuk jumlah token pada kalimat x yang dilakukan *alignment* dengan token pada kalimat pasangannya. Sementara $n_c(S^{(x)})$ merupakan jumlah keseluruhan kata pada kalimat x.

2.5 Vektor Semantik

Vektor semantik adalah metode yang digunakan untuk menghitung sebuah kata dari distribusi kata-kata di sekitarnya [3]. Kata-kata ini umumnya diwakili sebagai vektor atau deret angka yang terkait dengan beberapa cara untuk diperhitungkan.

2.6 Word2vec

Terdapat cara lain untuk merepresentasikan vektor kata yaitu dengan word2vec yang telah dibangun oleh Google. Vektor dari word2vec ini bisa mewakili makna dari sebuah kata dan dapat diukur dengan beberapa vektor sebagai perbandingan. Word2vec juga dapat dilakukan learning. Terdapat dua arsitektur yang digunakan untuk *learning* yaitu dengan model *Continuous Bag-of-Words* (CBOW), dan model *Skip-gram* [4].

2.7 Regresi *Ridge*

Regresi *ridge* adalah salah satu regresi yang mengatasi masalah korelasi tinggi antar beberapa variabel bebas. Multikolinieritas dalam regresi linier berganda yang mengakibatkan matriks $X^T X$-nya hampir singular yang pada gilirannya menghasilkan nilai estimasi parameter yang tidak stabil [5].

2.8 Evaluasi Performansi Sistem

Evaluasi performansi sistem dilakukan dengan mengukur nilai korelasi antara nilai kesamaan semantik yang dihasilkan sistem dengan *gold standard* menggunakan metode *pearson correlation*.

2.9 *Paraphrase Database* (PPBD)

Paraprhase database (PPDB) adalah basis data yang berisi kumpulan parafrase yaitu suatu kata yang memiliki makna yang sama namun dituliskan dalam bahasa yang berbeda.

3 Perancangan Sistem

3.1 Gambaran Umum Sistem

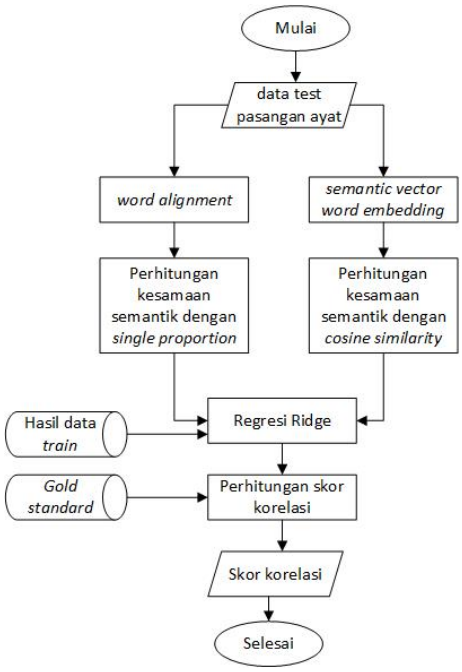
Sistem yang akan dibangun dalam tugas akhir ini bertujuan untuk mencari tingkat kesamaan semantik antar dua ayat dalam Alquran. Nilai kesamaan sematik berdasarkan hasil impelemntasi antara *word alignment* dan vektor semantik dengan *word embedding* menggunakan word2vec *toolkit* yang kemudian dilakukan evaluasi untuk perhitungan semantik antar kedua fitur dan *gold standard* menggunakan *ridge regression*.

Penjelasan terkait dengan tahapan dan proses utama sistem pada Gambar 1 adalah sebagai berikut:

a. membaca data input berupa pasangan potongan ayat Alquran terjemahan bahasa Inggris.

b. Sistem terbagi dalam dua tahap yaitu tahap *word alignment* dan vektor semantik *word embedding*.

c. Pada tahap *alignment* sistem melakukan tahapan *preprocessing* data, yaitu dengan melakukan tokenisasi, *stopwords removal*, dan *lemmatization* terhadap data input pasangan potongan ayat Alquran.



Gambar 1: Gambaran Umum Sistem

d. Tahapan yang dilakukan untuk membangun sistem *word alignment* ialah dengan menentukan fitur-fitur *alignment*-nya terlebih dahulu, fitur utama yang relatif mendasar ialah *identical words* dan PPDB, namun dalam penelitian ini dilakukan penambahan fitur yaitu seperti *textual neighborhood, dependency, name entity*.

e. Pada tahap vektor semantik *word embedding* sistem membangun sebuah metode dengan menggunakan *tool* word2vec. Metode tersebut kemudian digunakan untuk menguji data input pasangan potongan ayat Alquran.

f. Sistem melakukan perhitungan nilai kesamaan semantik hasil dari tahapan *word alignment* dan menghasilkan *output* berupa angka nilai kesamaan semantik dari data pasangan potongan ayat Alquran dan disimpan dalam bentuk *.txt*.

g. Sistem juga melakukan perhitungan nilai kesamaan semantik hasil dari vektor semantik *word embedding* serta dilakukan perhitungan nilai kesamaan semantik menggunakan *cosine similiarity*. Output yang dihasilkan berupa angka nilai kesamaan semantik dari data pasangan potongan ayat Alquran dalam bentuk *.txt*.

h. Evaluasi yang dilakukan dengan cara sistem membaca hasil nilai kesamaan semantik tersebut dan melakukan perhitungan dengan menggunakan *ridge regression* dimana nilai X merupakan hasil kesamaan semantik dari tahap *word alignment* dan vektor semantik *word embbeding*, sementara nilai Y merupakan *gold standard* yang sudah dipersiapkan. Pada tahapan ini dibutuhkan data latih untuk membangun metode regresinya. Data latih yang digunakan yaitu data pasangan terjemahan ayat Alquran tahun 2016 yang dikumpulkan oleh Dwi Jayanti Wulandari. Hasil dari regresi tersebut kemudian dilakukan perhitungan korelasi menggunakan *pearson correlation* dengan *gold standard* sehingga menghasilkan nilai korelasi yang menandakan tingkat akurasi sistem yang dibangun dengan data yang digunakan.

3.2 Pengumpulan dan Pengolahan Data

Data pasangan ayat dikumpulkan secara manual penulis sebanyak 400 pasangan ayat terjemahan diambil dari Qursim hasil penelitian keterkaitan *corpus* Alquran berdasarkan tafsir Ibn katsir oleh Abdul-Baquee M. Sharaf dan Eric S. Atwell yang dimuat dalam situs web *textminingthequran.com* sebanyak 350 pasangan dan ayat Indeks Tematik Kementrian Agama Republik Indonesia dan Pusat Kajian Hadist Al-Mughni sebanyak 50 pasangan. Data tersebut kemudian ditambahkan dari data yang sudah dikumpulkan tahun lalu oleh Dwi Jayanti Wulandari sebanyak 400 pasangan ayat. Setelah data pasangan ayat dibangun selanjutnya dilakukan pengolahan penilaian kesaman pasangan ayat oleh domain pakar yang dilakukan secara manual dan memberikan angka penilaian 0 hingga 5 untuk dijadikan sebagai data *gold standard* yang diperoleh melalui data kuisioner.

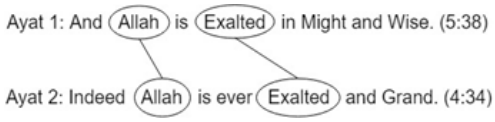
3.3 Proses *Word Alignment*

3.3.1 *Preprocessing*

Preprocessing yang digunakan pada penelitian ini adalah pembersihan data, *tokenization, stopword removal,* dan *lemmatization*.

3.3.2 *Align Identical Words*

Data yang telah melewati *preprocessing* akan dibandingkan dengan perbandingan *string* atau hurufnya, setiap kata yang ada di kalimat kedua dibandingkan dengan setiap kata di kalimat pertama seperti Gambar 2.

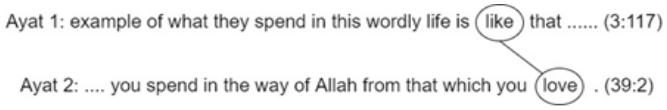


Gambar 2: Contoh *Align Identical*

Output dari setiap *aligner* adalah pasangan kata yang di-*align* dan indeksnya, untuk *aligner identical word* dengan contoh pasangan kalimat seperti Gambar 2 output yang dihasilkan adalah `[Allah, Allah], [Exalted, Exalted]`.

3.3.3 *Align* PPDB

Setiap kata pada kalimat inputan akan diperiksa apakah pasangan kata tersebut terdapat didalam PPDB atau tidak. Jika ada maka akan dilakukan *align* seperti pada Gambar 3.

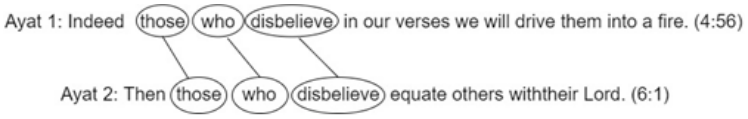


Gambar 3: Contoh *Align PPDB*

sehingga *output* yang dihasilkan dari fitur ini adalah `[like, love]`. Namun jika tidak ada dalam PPDB maka tidak dilakukan *align*.

3.3.4 *Align Word Sequences*

Akan dilakukan *align* dengan ketentuan minimal dua kata yang berurutan dan memiliki kesamaan penulisan seperti pada Gambar 4.



Gambar 4: Contoh *Align Word Sequence*

Output dari fitur *align* ini dengan contoh kasus diatas adalah `[those, those], [who, who], [disbelieve, disbelieve]`.

3.3.5 *Align Named Entities*

Fitur *align named entities* melakukan *align* terhadap *named entities*. Fitur ini digunakan dengan bantuan *stanford name entity recognizer* agar sistem dapat mendeteksi entitas lokasi, organisasi dan nama orang.

3.3.6 Perhitungan Kesamaan Semantik *Word Alignment*

Perhitungan kesamaan semantik *word alignment* menggunakan persamaan 1, yaitu dengan membagi jumlah total kata yang di-*align* dengan total kata hasil *preprocessing*. Hasil bagi tersebut kemudian dikalikan dengan 5 agar menyesuaikan dengan data *gold standard*.

3.4 Proses Vektor Semantik

3.4.1 *Preprocessing*

Preprocessing yang digunakan sebelum proses komputasi vektor pada penelitian ini hanya digunakan *tokenization*.

3.4.2 *Load* Model *Learning*

Dibutuhkan model *learning* untuk mempermudah proses komputasi vektor. Dalam penelitian ini digunakan model *text8* yaitu ringkasan *corpus* Wikipedia yang dibangun oleh Matt Mahoney.

3.4.3 Komputasi Vektor

Komputasi vektor dilakukan dengan memasukan potongan ayat yang sudah dilakukan tahap *preprocessing* agar menghasilkan nilai dalam bentuk vektor dari metode *text8.bin* seperti pada Tabel 1.

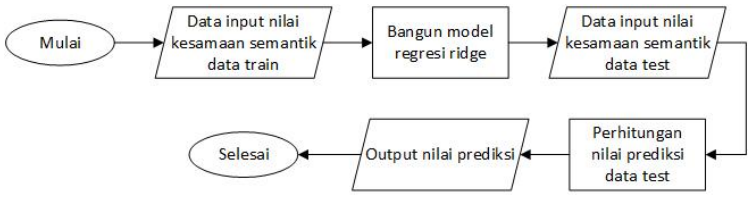
Tabel 1: Contoh Komputasi Vektor

Input Data		
Ayat 1 Token	:	['and', 'let', 'not', 'those', 'who', 'disbelieve', 'think', 'they', 'will', 'escape'].
Ayat 2 Token	:	['then', 'those', 'who', 'disbelieve', 'equate', 'others', 'with', 'their', 'lord'].
Hasil Komputasi Vektor		
Vektor Ayat 1	:	[0,33492102 0,06178305 -0,11226578 -0,1497349 -0,43190207 0,19694783 0,47498547 0,1854155 0,5367676 -0,2415434 0,00473845 0,03588425 ... 0,58831825 -0,31863664 -0,33278767 0,25012344 -0,03292524].
Vektor Ayat 2	:	[0,49925773 0,20194609 -0,15479319 -0,05784126 -0,93383033 0,16123282 0,03097209 0,30738768 0,59329191 -0,29155186 0,03253944 0,10368647 ... 0,16017136 -0,21065286 -0,32201015 0,53435173 0,13147605].

3.4.4 Perhitungan Kesamaan Vektor Semantik

Hasil komputasi vektor tersebut kemudian akan dilakukan proses perhitungan kesamaan semantik dengan menggunakan *cosine similarity*. Hasil dari perhitungan *cosine similarity* dikalikan dengan 5 agar setara dengan *gold standard*.

3.5 Regresi Antara Hasil *Word Alignment* dengan Vektor Semantik



Gambar 5: Alur Regresi

Seperti pada Gambar 5 hal pertama yang dilakukan adalah mencari nilai kesamaan semantik yang dihasilkan oleh *alignment* dan vektor semantik pada data latih dan data uji. Kemudian dilakukan pembangunan metode regresi menggunakan *ridge* dengan menggunakan modul *python sklearn* dengan parameter *alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None, normalize=False, random_state=None, solver='auto', tol=0.001* parameter tersebut merujuk kepada penelitian yang dilakukan oleh Sultan et al [2].

4 Pembahasan

4.1 Implementasi

Pada penelitian ini sistem dibangun sesuai dengan perancangan sistem. Kemudian sistem diimplementasikan untuk mengukur kesamaan semantik pada data pasangan ayat yang telah dibangun. Data pasangan ayat tersebut yaitu, data pasangan berdasarkan indeks tematik, data Qursim dan gabungan data indeks tematik dan Qursim yang dikumpulkan pada tahun 2016 dan 2017. Data tersebut dipisah agar dapat dijadikan sebagai data latih dan data uji.

4.2 Hasil Pengujian

Berdasarkan hasil pengujian dengan menggunakan metode *word alignment*, vektor semantik, serta kombinasi keduanya menggunakan data pasangan terjemahan Alquran bahasa Inggris didapatkan hasil yang disertakan dalam Tabel 2.

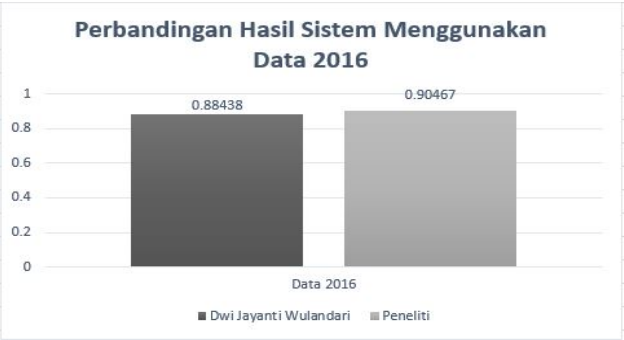
Tabel 2: Hasil Data Alquran 2016 dan 2017 Pada Masing-masing Metode

Data	Metode	Korelasi	*Standard Error*
2016	*Word Alignment*	**0,89978**	0,0011
	Vektor Semantik	0,855	0,0013
2017	*Word Alignment*	**0,77776**	0,0017
	Vektor Semantik	0,66625	0,0020

Tabel 3: Hasil Data Alquran 2016 dan 2017 Kombinasi Metode

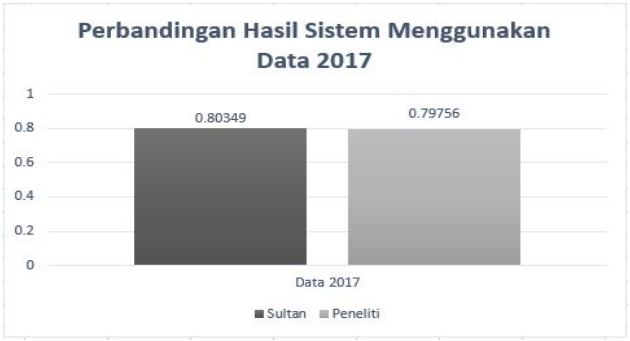
Data	Korelasi	*Standard Error*
2016	**0,90467**	0,0011
2017	**0,79756**	0,0017

Nilai korelasi yang dihasilkan sistem menggunakan data pasangan Alquran 2016 pada kedua metode lebih tinggi dibandingkan dengan nilai korelasi yang dihasilkan sistem menggunakan data pasangan Alquran 2017, ini dikarenakan perbedaan perbandingan jumlah variasi data yang berbeda antara kedua data tersebut, data 2016 terdiri dari 200 pasangan Ibnu katsir yang berisi pasangan kalimat yang cukup kompleks sementara sisa 200 pasangan Indeks Tematik yang berisi data pasangan kalimat sangat sederhana. Berbeda dengan data 2017 yang memiliki data Ibnu Katsir lebih banyak yaitu 350 pasangan sementara data Indek Tematiknya hanya 50 pasangan. Namun jika dilakukan kombinasi antar kedua hasil kesamaan semantik antara *word alignment* dengan vektor semantik menggunakan regresi kedua data tersebut mengalami kenaikan akurasi. Pada data 2016 mengalami kenaikan akurasi menjadi 0.90472 sementara pada data 2017 menjadi 0.73594. *Standard error* pada masing-masing hasil tidak mempengaruhi hasil analisis diatas dikarenakan *error* yang dihasilkan kecil.



Gambar 6: Perbandingan Hasil Korelasi Sistem Menggunakan data Alquran 2016

Seperti pada Gambar 6 hasil korelasi yang dilakukan peneliti dengan menggunakan data 2016 lebih tinggi 0,02029 jika hasil sistem yang dibangun pada penelitian ini dibandingkan dengan hasil sistem yang dibangun oleh Dwi Jayanti Wulandari yaitu dengan menggunakan *alignment* dengan *semantic role labelling*. Tetapi hal berbeda dengan hasil korelasi dengan menggunakan data 2017 seperti pada Gambar 7 hasil penelitian ini lebih kecil 0,00593 jika dibandingkan hasil korelasi yang dilakukan menggunakan sistem yang dibangun oleh Md Sultan Arafat.



Gambar 7: Perbandingan Hasil Korelasi Sistem Menggunakan data Alquran 2017

4.3 Analisis Penggunaan Vektor Semantik

Menghasilkan nilai korelasi untuk data 2016 sebesar 0,855 namun untuk data 2017 hanya mendapat nilai korelasi 0,66625, hal ini dikarenakan data pada 2017 lebih beragam tingkat kompleksitas kalimatnya serta banyak pasangan kalimat yang secara susunan kata berbeda tetapi memiliki nilai yang relatif tinggi kesamaan maknanya atau sebaliknya pasangan kalimat tersebut memiliki banyak kesamaan kata namun memiliki makna yang berbeda. Hal ini dikarenakan metode yang digunakan pada word2vec yaitu *text8* memiliki kata-kata yang tidak terlalu banyak sehingga ada beberapa kata yang tidak memiliki nilai array vektornya. Hal tersebut menyebabkan nilai kesamaan semantik yang dihasilkan metode ini tidak mendekati *gold standard*. Beberapa contoh kata yang tidak memiliki *array* vektor adalah *disbelieve, inclining, polythesists, crucify, transgressed, lifelessness, recompensed*, kata-kata yang jarang digunakan dalam pembuatan kalimat. Selain kata tersebut, kata-kata yang sering digunakan dalam Alquran yang bentuk kata Arab latin dengan bahasa Inggrisnya tidak memiliki perubahan juga tidak memiliki *array* vektor, sebagai contoh kata *fitnah, iblees, almasjid, alharam, tawaf, jinn, kabah*, dan kata-kata lain yang tidak memiliki perubahan bentuk.

4.4 Analisis Perhitungan Kesamaan Semantik

Berdasarkan hasil penelitian pasangan kalimat yang memiliki nilai semantik yang relatif tinggi adalah pasangan kalimat yang banyak memiliki kesamaan kata atau kesamaan makna katanya. Nilai yang dihasilkan *alignment* dan vektor semantik tersebut sama seperti *gold standard* sehingga menghasilkan nilai semantik yang mendekati *gold standard*. Hal tersebut dikarenakan kata-kata pada kedua kalimat memiliki banyak kata yang sama secara penulisan dan makna yang terdapat dalam PPDB, serta semua kata dalam kalimat memiliki *array* vektor. Sistem masih dapat menghasilkan nilai kesamaan semantik yang cukup baik pada kalimat yang sedikit rumit dengan perbedaan panjang kalimat. Namun ada beberapa pasangan kalimat yang memiliki *gold standard* rendah tetapi sistem menghasilkan nilai yang tinggi. Untuk kalimat yang terlalu panjang dan pasangan yang memiliki panjang sangat berbeda sistem belum bisa menangani baik *alignment* ataupun vektor semantik. Juga terdapat kasus bentuk kalimat yang memiliki *gold standard* tinggi, vektor semantik menghasilkan nilai yang tinggi namun hasil dari *alignment* sangat rendah dan jauh dari *gold standard*. Hal ini disebabkan karena kata-kata dalam kalimat tersebut memiliki sedikit kesamaan kata secara penulisan. Dan kebalikan kasus tersebut yaitu pada pasangan kalimat yang memiliki *gold standard* yang rendah, hasil sistem menggunakan *alignment* rendah, namun hasil sistem menggunakan vektor semantik tinggi.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan pada Tugas Akhir ini, dapat disimpulkan bahwa:

1. Penggunaan kombinasi metode antara *word alignment* dengan vektor semantik dengan mencari nilai kesamaan semantik menggunakan regresi *ridge* menghasilkan nilai korelasi realtif tinggi dengan menggunakan data Alquran 2016 ataupun 2017.

2. Fitur yang relatif berpengaruh dalam melakukan *alignment* adalah fitur *identical words*, fitur PPDB memiliki pengaruh namun tidak terlalu besar, fitur *words sequence* dan fitur *align named entities* tidak memiliki pengaruh.

3. Penggunaan metode vektor semantik pada penilaian kesamaan semantik lebih rendah dibandingkan dengan hasil penilaian kesamaan semantik menggunakan *alignment* hal ini dikarenakan data *corpus* yang digunakan untuk membangun model vektor dimensinya terlalu kecil, dan terdapat kata-kata yang tidak memiliki perubahan bahasa dari bahasa Arab latin ke bahasa Inggris.

4. Kombinasi *alignment* dan vektor semantik menggunakan regresi pada penelitian ini dapat digunakan untuk mengukur kesamaan semantik dan menghasilkan nilai yang lebih baik dibandingkan dengan hanya masing-masing fitur.

5.2 Saran

Adapun Saran yang diperlukan untuk pengembangan pada Tugas Akhir ini adalah sebagai berikut:

1. Membuat kumpulan parafrase Alquran terjemahan bahasa Inggris agar dapat menangani parafrase yang berhubungan dengan Alquran.
2. Mencoba membuat name entities khusus untuk data Alquran agar *tag* yang dihasilkan dapat sesuai dengan nama-nama yang ada dalam Alquran.
3. Mencoba membangun model vektor dengan data corpus yang berdimensi lebih besar.
4. Mencoba menggunakan regresi lain karena terdapat kemungkinan menghasilkan korelasi yang lebih baik.

Daftar Pustaka

[1] B. Danushka, M. Yutaka, and I. Mitsuru. Measuring Semantic Similarity between Words Using Web Search Engines,. 2007.

[2] dan S. Tamara Md. A. Sultan, B. Steven. Supervised Models of Sentence Similarity,. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2016)*, 2016.

[3] dan G. Kruszewski M. Baroni, G. Dinu. Dont Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors,. *ACL 2014*, pages 238–247, 2014.

[4] Q. V. Le dan T. Mikolov. Distributed representations of sentences and documents,. *In Proceedings of the 31st International Conference on Machine Learning*, 2014.

[5] N. Draper dan H. Smith. Analisis Regresi Terapan Edisi 2 (Terjemahan Sumantri),. *Jakarta: PT. Gramedia Pustaka Utama*, 1981.