Implementasi dan Analisis Pengukuran *Cross Level Semantik Similarity* Dengan Metode *Alignment-Based Disambiguation* Dalam Pencarian Ayat AlQuran

Implementation and Measurement Analysis of Cross-level Semantics Similarity With Alignment-Based Disambiguation Method In Search of Al-Quran Verse

Mu'ti Cahyono Putro[1], Ir. Moch. Arif Bijaksana., Ph.D[2], Dr. Arief Fatchul Huda, S.Si., M.Kom[3]

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

[1] em.chepe@gmail.com, [2] arifbijaksana@gmail.com, [3] afhuda@gmail.com

Abstrak

Al Qur'an merupakan kitab suci umat islam yang diturunkan kepada Nabi Muhammad SAW melalui malaikat Jibril sebagai pedoman hidup. Al Qur'an memiliki 30 juz, 114 surat, 6.243 ayat. Tentunya dengan banyaknya ayat AlQur'an membuat sulit dalam proses pencarian suatu ayat. *Cross Level Semantic Similarity* merupakan pengukuran kesamaan antara dua buah variabel yang memiliki ukuran yang berbeda seperti pengukuran kata dengan kalimat. Metode *Alignment-Based Disambiguation* merupakan metode yang ditujukan untuk mengukur nilai kesamaan(*similarity*) suatu pasangan data yang memiliki ukuran yang berbeda. Oleh karena itu, pada tugas akhir ini akan akan digunakan pengukuran nilai kesamaan dengan menggunakan metode *Alignment-Based Disambiguation* dengan bantuan WordNet yang dapat diterapkan dalam pencarian ayat Al Qur'an.

Keywords: *Al Qur'an, Cross Level Semantic Similarity, Alignment Based Disambiguation, Similarity, WordNet*.

Abstract

The Qur'an is a sacred book of Muslims that was revealed to the Prophet Muhammad SAW through the angel Gabriel as a guide of life. The Qur'an has 30 juz, 114 letters, 6,243 verses. Certainly with the many verses of the Qur'an makes it difficult in the process of searching a verse. Cross Level Semantic Similarity is a measure of the similarity between two variables that have different sizes such as word measurement with sentences. Alignment-Based Disambiguation Method is a method intended to measure the value of similarity of data pairs that have different sizes. Therefore, in this final project will be used the measurement of similarity value by using Alignment-Based Disambiguation method with WordNet help which can be applied in searching Qur'anic verse.

Keywords: Al Qur'an, Cross Level Semantic Similarity, Alignment Based Disambiguation, Similarity, WordNet.

1. Pendahuluan

Al Qur'an merupakan kitab suci umat islam yang diturunkan kepada Nabi Muhammad SAW melalui malaikat jibril sebagai pedoman hidup [8]. Al Qur'an memiliki 30 juz yang terbagi atas 114 surat dan tersusuan atas 6.243 ayat dengan bahasa arab sebagai bahasa penyusunnya. Meskipun saat ini sudah banyak Al Qur'an terjemahan dalam berbagai Bahasa seperti Al Qur'an terjemahan dalam Bahasa inggris oleh Abdullah Yusuf Ali [2]. Meskipun begitu sulit untuk sistem dalam hal mencari kesamaan dan keterkaitan ayat Al-Quran, karena tidak memiliki kemampuan intuisi seperti manusia. Dengan intuisi yang dimiliki, manusia dapat dengan tepat menentukan kesamaan dan keterkaitan dari informasi yang disediakan.

Dengan pendekatan Semantic Text Similarity (STS) dapat diketahui tingkat kesamaan suatu ayat dengan ayat lainnya, penilaian dilakukan dengan cara memberi nilai tertentu. STS telah banyak digunakan dalam penerapan disiplin ilmu *Natural Language Processing* (NLP) dan *Text Mining* seperti *information retrieval, machine translation, question answering, text summarization* dan lain sebagainya [6]. Hingga saat ini penelitian terkait *Semantic Text Similarity* masih terus dikembangkan, salah satu topik penelitian *Semantic Text Similarity* adalah *cross level STS* yaitu pengukuran antara pasangan data yang memiliki ukuran berbeda seperti kata dengan kalimat, kata dengan phrase, paragraph dengan kalimat, *word* dengan *sense* dan lain sebagainya sebagai contoh kata *lord* (Tuhan) dengan surat An-Nahl ayat 4: *He hath created man from a drop of fluid, yet behold! he is an open opponent.* (Dia telah menciptakan manusia dari mani, tiba-tiba ia menjadi pembantah yang nyata) dan surat Al Fatihah ayat 1: *In the name of Allah, the Entirely Merciful, the Especially Merciful* (Dengan menyebut nama Allah yang Maha Pengasih lagi Maha Penyayang) memiliki kesamaan tinggi dalam makna yaitu menerangkan

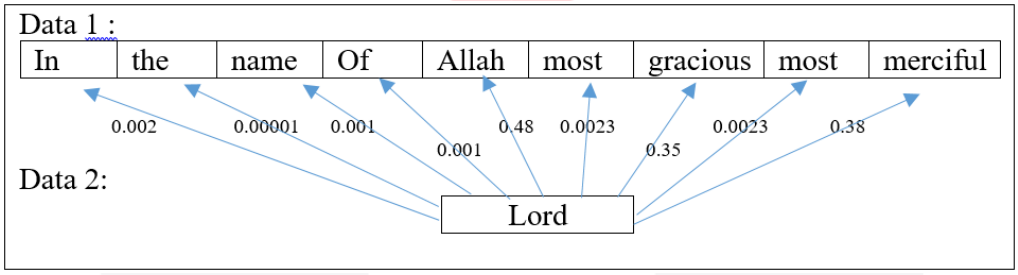
tentang ketuhanan. Salah satu metode untuk menghitung tingkat kesamaan *Semantic Similarity* adalah dengan metode *Alignment-Based Disambiguation* [10] dengan menggunakan corpus Wornet dan Semantic Signature yang berbentuk PageRank vectors yang dibangun oleh Pilehvar dkk.

Pada tugas akhir ini akan dibangun sistem yang mampu menghitung nilai similarity dari dua buah data yang memiliki ukuran berbeda dan akan dibangun sistem pencarian ayat Al Quran dengan menerapkan metode Alignment-Based Disambiguation. Metode tersebut dipilih karena penelitian terkait metode tersebut menunjukan hasil yang baik dalam mencari nilai similarity pada pasangan data yang memiliki ukuran berbeda.

2. Dasar Teori dan Perancangan

2.1. Alignment-Based Disambiguation

Metode *Alignment-Based Disambiguation* akan membentuk pasangan synsets dari setiap pasangan token kemudian akan mencari nilai *similarity* dari setiap pasang *synsets*, pasangan dengan nilai tertinggilah yang akan pilih. Pada penelitian ini akan menggunakan metode *Jaccard Similarity* untuk menghitung nilai similarity setiap pasangan synset. Sehingga memungkinkan untuk melakukan perbandingan pada pasangan teks pendek atau pasangan kata yang memiliki informasi kontekstual yang sedikit [10]. Dengan menggunakan metode Alignment-Based Disambiguation urutan kata pada kalimat tidak diperhatikan. Pasangan yang dipilih adalah pasangan token yang memiliki nilai similarity tertinggi. Contoh seperti pada surat Al Fatihah ayat 1 terjemahan Bahasa inggris yang berbunyi “in the name of Allah most gracious most merciful” dengan kata “lord” seperti pada gambar 1.



Gambar 1. Ilustrasi metode *Alignment Based Disambiguation*

2.2. Jaccard Similarity

Perhitungan Jaccard atau Jaccard Similarity merupakan metode yang diperkenalkan oleh Paul Jaccard, pada tahun 1901. Jaccard Similarity yaitu pengukuran menggunakan ranking untuk mengidentifikasi query. Metode Jaccard Similarity dapat menghasilkan performansi yang baik dalam similarty kata [14]. Dalam menghitung nilai similarity antara pasangan, metode ini menggunakan persamaan 2.2:

$$s = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} = \frac{|V_1 \cap V_2|}{|V_1| + |V_2| - |V_1 \cap V_2|} \quad (2.2)$$

2.3. Weighted Overlaps

Merupakan sebuah pengukuran nonparametric similarity dengan membandingkan kesamaan ranking/bobot untuk elemen vektor yang sama dari setiap pasanganan vektor. Weighted Overlap mewakili jumlah dari bobot setiap vektor [5]. Untuk menghitung weighted Overlap yang terlebih dahulu harus mencari elemen elemen yang sama antar kedua vektor. Kemudian dihitung nilai similarity menggunakan persamaan 2.3 dimana n adalah list elemen yang sama pada kedua vektor, sedangkan r adalah bobot elemen:

$$x_1 = \sum_{i=1}^{|n|} (r1[n_i] + r2[n_i])^{-1} \quad (2.3)$$

Hasil dari persamaan 2.3 adalah *weighted overlap* yang belum dinormalisasi. Untuk menghasilkan nilai kesamaan dalam 0 atau 1(normalisasi), menggunakan persamaan 2.4:

$$x_1 = \sum_{i=1}^{|n|} (2i)^{-1} \quad (2.4)$$

Kemudian untuk menghitung nilai *similarity* diperoleh dengan membagi hasil dari formula 2.3 dengan hasil yang diperoleh dari formula 2.4.

2.4. Pengukuran Evaluasi

2.4.1. Pearson Correlation

Pearson product-moment correlation atau yang biasa disebut pearson correlation memiliki penggunaan yang lebih luas dalam mengukur keterkaitan antara duavariabel[12] dengan nilai korelasi(r) berapa pada *range* -1 hingga 1[1]. Formula yang digunakan untuk mengukur keterkaitan dua variabel menggunakan pearson correlation yaitu pada persamaan 2.6 dimana x adalah nilai *gold standard* dan y adalah nilai yang dihasilkan sistem.

$$r = r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \times \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad (2.6)$$

2.4.2. Spearman Correlation

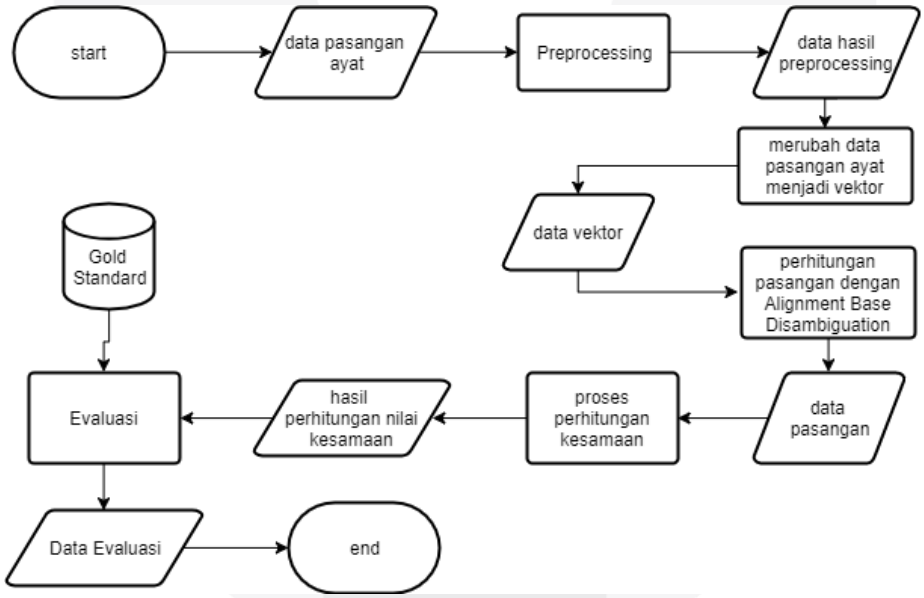
Merupakan pengukuran korelasi yang berdasarkan rank keterhubungan antara dua variabel atau lebih. Koefisien spearman correlation didefinisikan sebagai koefisien pearson correlation antara ranked variables [13]. Formula yang digunakan dalam mengukur correlation dengan spearman correlation yaitu pada persamaan 2.7:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2.7)$$

2.5. Perancangan

2.5.1. Gambaran Umum Sistem

Sistem yang akan dibangun pada penelitian ini adalah sistem yang mempu melakukan perhitungan pasangan dataset dalam Bahasa Inggris dengan menerapkan pendekatan semantic text similarity menggunakan metode *alignment-based disambiguation*. Perhitungan *semantic similarity* yang digunakan pada penelitian ini yaitu: *Weighted Overlap* dan *Jaccard Similarity*. Secara umum sistem digambarkan pada Gambar 2.



Gambar 2. Gambaran Umum Sistem

2.5.2. Preprocessing

Proses *preprocessing* melakukan proses pengolahan data input mencakup tokenisasi yaitu proses pemecahan kalimat menjadi kata(token), *POS Tagging* yaitu proses pencarian jenis kata dari suatu *word*, *lemmatization* yaitu proses untuk mendapatkan bentuk asli dari suatu kata, *stopwords removal* yaitu proses mengeliminasi kata yang diangkap tidak penting.

2.5.3. Convert to Vector

Dataset dirubah kedalam bentuk vektor dengan Wordnet dan *Semantic Signatures*. Dimulai dengan mencari setiap *sysonim set(synset)* untuk setiap kata yang menyusun pada data input menggunakan WordNet. Kemudian akan di cari nilai vektor untuk setiap *synset* menggunakan *Semantic Signatures*.

[1]http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

2.5.4. Proses Perhitungan

Proses perhitungan terjadi sebanyak dua kali untuk satu pasangan data input yang pertama perhitungan terjadi pada metode *Alignment-Based Disambiguation* dengan metode *jaccard similarity* dengan parameter pasangan vektor yang akan memilih pasangan vektor yang memiliki nilai tertinggi untuk setiap pasangan token. Proses perhitungan kedua terhadap data yang dihasilkan pada prose perhitungan pertama menggunakan *weighted overlap.*

3. Pembahasan

3.1. Analisis Pengaruh *Alignment-Based Disambiguation*

Tabel 1. Hasil Pengujian dataset *SemEval 2014 task 3 prh2word.txt*

Sistem	*Spearman Correlation*	*Pearson Correlation*
Dengan ABD	0.2948	0.235
Tanpa ABD	0.2744	0.187

Nilai korelasi yang dihasilkan dengan menggunakan metode *Alignment-Based Disambiguation* lebih tinggi dibandingkan dengan sistem yang tidak. Hal tersebut dikarenakan, pada proses *Alignment-Based Disambiguation* akan memilih *synsets* pada token yang memiliki nilai kemiripan tertinggi dengan *synsets* pada token pasangannya. Sedangkan sistem tanpa menggunakan metode *Alignment-Based Disambiguation* akan menghasilkan *array* vektor yang lebih panjang dari pada sistem yang menggunakan metode *Alignment-Based Disambiguation* hal tersebut tentunya akan mempengaruhi nilai indeks atau pembobotan dari setiap element vektor dan juga akan mempengaruhi nilai *similarity* yang dihasilkan.

3.2. Analisis Pengaruh Proses Proprocessing *Lemmatization*

Tabel 2. Hasil Pengujian Pengaruh *Preprocessing*

Sistem	*Spearman Correlation*	*Pearson Correlation*
Tanpa Proses Lemmatization	0.278	0.216

Sistem ini dibangun menggunakan WordNet sebagai corpus utama sehingga proses identifikasi suatu kata sangat bergantung pada WordNet. Kata yang tidak dikenali WordNet maka tidak memiliki vektor sehingga tidak dapat diperoleh informasi yang terdapat pada kata tersebut, seperti *directions#n, bummed#v, knocked#v, expecting#v, fest#a, male-dominated#a, room-temperature#a, down-low#n, overthe-shoulder#a, hanky-panky#n, multi-national#a*. Pada sistem ini proses preprocessing menggunakan library Stanford core NLP namun tidak semua hasil dari proses preprocessing tidak dapat di kenali oleh WordNet seperti "*using#v*" dimana token "*using*"dikenali sebagai kata kerja(*verb*) dengan library Stanford Core NLP namun pada WordNet token "*using*" dikenali sebagai kata benda (*Noun*) sehingga butuh suatu proses untuk melakukan perbaikan pada hasil *preprocessing* agat dapat dikenali oleh WordNet.

4. Kesimpulan

Kesimpulan yang diperoleh dari penelitian ini adalah:

1. Nilai kesamaan yang dihasilkan dengan menggunakan metode *Alignment-Based Disambiguation* menghasilkan nilai korelasi yang lebih tinggi dari sistem yang tidak menggunakan metode *Alignment-Based Disambigution*, yaitu : sistem dengan metode *Alignment-Based Disambigution* mendapat nilai 0.2948 untuk nilai *Spearman Correlation* dan 0.235 untuk *Pearson Correlation* dan untuk sistem tanpa metode *Alignment-Based Disambiguation* mendapat nilai 0.2744 untuk *Spearman Correlation* dan 0.187 untuk *Pearson Correlation*.
2. Nilai korelasi yang dihasilkan dengan menggunakan metode *Alignment-Based Disambiguation* dengan corpus WordNet dan *Semantic Signatures* menghasilkan nilai korelasi yang lebih tinggi yakni sebesar 0.235 dibandingkan dengan nilai korelasi yang dihasilkan dengan sistem Omiotis yang dibagun pada penelitian yang sebelumnya yakni sebesar 0.111 untuk *Pearson Correlation* menggunakan dataset *SemEval 2014 task 3:prh2word.txt*.
3. Sistem yang menggunakan WordNet sebagai corpus saat bergantung pada proses preprocessing. Karena tanpa proses preprocessing ada kemungkinan suatu token tidak dapet dikenali oleh WordNet. Nilai kesamaan sistem tanpa proses *preprocessing(lemmatization)* mendapat nilai sebasar 0.278 untuk *Spearman*

Correlation dan 0.216 untuk *Pearson Correlation*. WordNet juga belum bisa mengenali sebuah nama dari suatu entitas seperti "*Aqsa*", dll. Dan untuk konteks keislaman WordNet belum mendukung secara penuh.

Daftar Pustaka

[1] DR. Moch. Arif Bijaksana, Akip Maulanaand, M. Syahrul Mubarok. Perancangan Semantic Similarity based on Word Thesaurus Menggunakan Pengukuran Omiotis Untuk Pencarian Aplikasi pada I-GRACIAS. 2016.

[2] Abdullah Yusuf Ali. The Meaning of the Holy Qur'an: Complete Translation with Selected Notes. 1934.

[3] M. T. Pilehvar D. Jurgens and R. Navigli. Cross-Level Semantic Similarity. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014.

[4] D.A. de Vaus. Survey in Social Research. *5th Edition (New South Wales: Allen and Unwin, 2002)*, 2002.

[5] Philipp Cimiano Elena Simperl and dkk Axel Polleres. The Semantic Web: Research and Applications: 9th Extended Semantic Web.

[6] Daniel Cer Eneko Agirre, Mona Diab 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 theshared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluatio*, page 385–393, 2012.

[7] Taher H. Haveliwala. Proceedings of the 11th international conference on world wide web.

[8] G. Lambert. The Leaders Are Coming!. *WestBow Press*, 2007.

[9] H. U. K M. I. Saeed M. Shoaib, M. N. Yasin and M. S. H. Khiyal. Relational WordNet Model for Semantic Searchin Holy Quran. *IEEE-5th International Conference on Emerging Technologies ICET*, 2009.

[10] D. Jurgens M. T. Pilehvar and R. Navigli. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, page 1341–1351, 2013.

[11] G. A. Miller. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM Vol. 38, No. 11*, 1995.

[12] Chung M.K. Correlation Coefficient. *The Encyclopedia of Measurement and Statistics*, 2007.

[13] J. L. Myers and A. D. Wel. Research Design and Statistical Analysis. 2003.

[14] Ekkachai Naenudorn Suphakit Niwattanakul*, Jatsada Singthongchai and Supachanun Wanapu. Using of Jaccard Coefficient for Keywords Similarity. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I*, 2013.

[15] Syark Al-Mukhtashar Sa'duddin Taftazani. ch. 1.