1. Pendahuluan

Latar Belakang

Keterkaitan dan kesamaan semantik berkaitan dengan bidang linguistik khususnya pada *Natural Language Processsing* (NLP) [2] yang belakangan ini menjadi topik yang menarik dan banyak diteliti. Kesamaan semantik dan keterkaitan kata memiliki peran penting dalam beberapa *task* dari *NLP* dan beberapa bidang terkait seperti *text classification, document clustering, text summarization*, dan lain sebagainya [8]. Kemiripan semantik adalah suatu proses yang digunakan untuk memperkirakan kekuatan hubungan semantik antara satuan bahasa, konsep atau mengetahui kesamaan kata. Mengevaluasi kesamaan dalam dokumen secara luas digunakan, salah satunya untuk aplikasi yang berkaitan dengan pencarian informasi, pengolahan bahasa alami, dll [7]. Evaluasi kata kesamaan (yaitu, *WordSim*) adalah salah satu metode alami tertua penilaian model distribusi semantik.

Al Qur'an merupakan kitab suci bagi umat Islam dan menjadi pedoman dan sumber hukum paling utama. Al Qur'an memiliki 30 Juz, 114 Surat dan 6236 ayat [8]. Buku ini saat ini sedang digunakan sebagai buku panduan kehidupan untuk 1.600.000.000 Muslim di dunia yang hidup saat ini [6]. Pada 6236 ayat ini terdapat banyak kosa kata yang sebenarnya memiliki kesamaan dan saling berkaitan [14], namun banyak yang terpisah jauh rentang katanya bahkan tidak terdapat pada satu surah atau satu juz, maka dari itu diperlukan pengkajian lebih dalam untuk dapat memahami makna yang terkait pada kata Al-Qur'an tersebut. Oleh karena itu juga, muncul gagasan tentang evaluasi dataset nilai kesamaan dan keterkaitan kata semantik pada pasangan kata Al-Quran seperti antara kata "Allah" dan "Tuhan"; "Surga" dan "neraka" atau bahkan "nabi" dan "rasul" dengan melalui deskripsi numerik yang diperoleh sesuai dengan perbandingan informasi pendukung yang berarti atau menggambarkan alam.

Dalam makalah ini, akan menyajikan karya dalam kesamaan kata semantik dengan memanfaatkan pembelajaran linguistik komputasi, terutama di bidang kemiripan semantik dan distribusi model semantik. Distribusi semantik (DS) adalah sebuah teori yang berkaitan dengan tradisi linguistik, yang menurut [10] seperti dalam proposal Zellig Harris, dari analisis distribusi sebagai batu linguistik fundamental sebagai disiplin ilmiah. Elemen linguistik dalam dua (ortogonal) jenis yang banyak belajar dengan hubungan distribusi dalam semantik distribusi penelitian (DS) hari ini: sintagmatik dan paradigmatik. Pada penelitian [15], menurut pendapat Milton, untuk mengukur kosakata yang dikuasai tidak dapat diperoleh ketika kata dituliskan hanya sebagai kata leksikal, namun dapat diukur ketika sudah tersebar dalam kalimat. Salah satu contoh adalah sinonim seperti "nabi" dan "rasul" dalam kalimat "Dia sebagai [nabi — rasul] Allah", dimana dua kata cenderung terjadi pada kalimat yang sama. Pada penelitian sebelumnya yang menggunakan dataset Al-Qur'an hanya dilakukan semantik untuk surah tertentu [1], keterkaitan kata dengan kosa kata bahasa arab [12] . Penelitian lain, membahas pembangunan dataset kesamaan dan keterkaitan semantik namun untuk penerapan pada bahasa Turki[11], dan penerapan evaluasi dataset untuk kata kerja [5] saja. Menurut [13] adanya penelitian lain yang datasetnya masih digunakan hingga saat ini seperti*Simlex-999, WordSim353, SimLex-666. Simlex-999* masih sedang digunakan hingga saat ini sebagai salah satu *gold standard* di model distribusi semantik penelitian pemodelan.

Untuk melengkapi penelitian sebelumnya, penelitian ini membuat daftar pasang kata dalam Al-Qur'an dengan bentuk kelas kata benda dan kata kerja yang menyajikan kesamaan kata dan keterkaitan kata (yaitu, Asosiasi). Evaluasi dataset ini bertujuan untuk menyediakan bidang pemodelan semantik untuk bahasa Indonesia yang diambil dari kosa kata Al-Qur'an dengan sumber evaluasi intrinsik berdasarkan definisi kata yang didapat dari Kamus Besar Bahasa Indonesia (KBBI). Melihat berkembangnya ilmu semantik yang ada untuk mempelajari Al-Qur'an, penelitian ini diciptakan pada saat yang sama untuk membuat dataset, sehingga penelitian di masa depan dengan fokus diskusi lain dapat menggunakan dataset ini untuk membantu kemajuan penelitian. Seperti penelitian [11], [5], tingkat kesamaan kata ini akan diukur dalam rentang kontinu [0,10] untuk setiap sepasang kata yang telah dipilih sebelumnya. Skor 10 menunjukkan kesamaan dan keterkaitan maksimum, sementara 0 tidak menunjukkan kesamaan dan keterkaitan. Nilai kesamaan dan Asosiasi diperoleh dengan melibatkan *gold standard* yang hasilnya dijelaskan ke dalam fungsi vektor dan fungsi tipe relasi. Evaluasi yang dibangun diharapkan dapat menghasilkan performansi yang baik berdasarkan nilai korelasi yang dihitung. Nilai korelasi yang dimaksud adalah yang didapat dari perhitungan *Spearman Rank*. Tantangan penelitian ini adalah untuk meniru intuisi manusia dalam mengukur kesamaan dan keterkaitan kata dalam semantik.

Topik dan Batasannya

Berdasarkan dari latar belakang yang telah dijelaskan, topik-topik yang diangkat dalam tugas akhir ini sebagai berikut:

1. Kelas Kata Terjemahan Al-Quran

   Topik yang diangkat berupa kosa kata terjemahan Bahasa Indonesia dari Al-Qur'an. Namun tidak semua kata Al-Qur'an menjadi topik dalam penelitian. Kelas kata yang diangkat menjadi topik yaitu kelas kata terbuka yang hanya berupa kata benda dan kata kerja.

2. Nilai *Similarity and Relatedness*

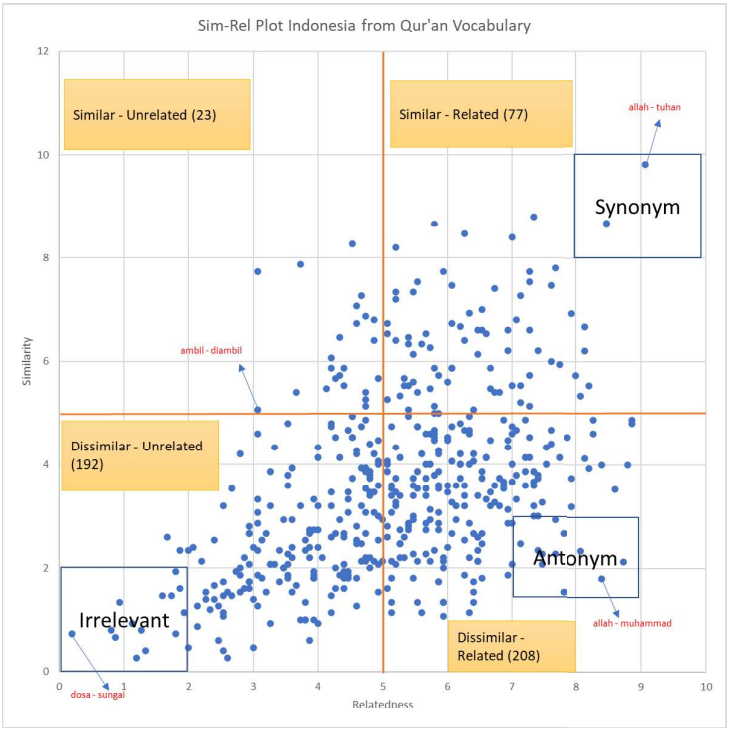
   Hasil dari tugas akhir ini berupa nilai kesamaan dan keterkaitan kata yang menunjukan tingkat kedekatan antar pasang kata. Kemudian dari nilai-nilai kesamaan tersebut akan dikelompokkan berdasarkan tingkat keterkaitannya dengan batas fungsi vektor tertentu.

3. Input dan Output

   Input dari sistem ini berupa pasang kata yang ingin diperediksi nilai kesamaan dan keterkaitan kata dengan pasang kata lainnya dalam bentuk bahasa indonesia yang diambil dari makna Al-Qur'an.Dengan S merepresentasikan nilai Similaritas atau Kesamaan dan R adalah Relatedness atau Keterkitan Output dari sistem ini berupa nilai kesamaan dan keterkaitan dari masing-masing pasang kata. Kemudian, nilai tersebut akan menentukan letak dari masing-masing pasang katanya pada titik koordinat untuk *SU (Similar-Unrelated) ; SR (Similar-Related) ; DU (Dissimilar-Unrelated) dan DR (Dissimilar-Related)* dan juga dapat menentukan tipe relasi (mis. Sinonim, Antonim dan *Irrelevant*). Berikut contoh untuk input dan output dengan Kata1 sebagai Kata Benda dan Kata2 sebagai Kata Kerja :

Tabel 1. Contoh Input Pasang Kata

Pasang Kata IND		**Nama1**		**Nama2**		**Nama3**		**Nama4**		**Nama5**	
Kata 1	**Kata 2**	**S**	**R**	**S**	**R**	**S**	**R**	**S**	**R**	**S**	**R**
adam	manusia	4	4	7	6	7	8	5	5	7	7
petunjuk	alquran	7	4	9	7	9	9	8	8	8	9
allah	tuhan	10	2	10	0	10	0	10	0	10	0
buku	alquran	9	5	4	3	8	4	8	9	5	4
hukuman	siksa	0	2	7	7	8	4	9	8	7	8
bersujud	membaca	3	6	0	0	0	2	0	2	0	1
membaca	melihat	4	8	6	8	7	8	4	7	5	4
membacakan	mengatakan	2	0	6	8	7	8	4	7	1	4
istirahat	bersujud	0	5	2	0	3	2	2	2	0	0
mencari	mengetahui	7	0	7	9	7	6	0	6	0	1



Gambar 1. Hasil Output *Scatter Plot*

Adapun batasan dari permasalahan yang ada pada tugas akhir ini sebagai berikut:

1. Analisis kata-kata yang dilakukan menggunakan basis vektor dari *Similarity and Relatedness*. .
2. Hanya mencakup dari terjemahan Al-Qur'an untuk diambil sebuah kelas kata benda dan kata kerja dari Al-Qur'an
3. Skor penilaian responden berkisar antara 0 - 10 yang digunakan sebagai nilai *gold standard*
4. Memilih responden sebanyak 15 orang Mahasiswa dari universitas dan jurusan yang berbeda

Tujuan

Berikut adalah tujuan yang ingin dicapai pada penulisan proposal/TA :

1. Membuat dataset Al-Qur'an untuk Bahasa Indonesia pada Kata Benda dan Kata Kerja
2. Menganalisa nilai evaluasi dari dataset keterkaitan dan kesamaan model semantik antar kosa kata dalam Al-Qur'an;
3. Mengetahui pengelompokkan pasang kata sesuai dengan inputan bahasa Indonesia dari Al-Qur'an
4. Mengetahui akurasi yang didapat dari pendekatan dalam menganalisis keterkaitan dan kesamaan semantik antar kata dalam bahasa Indonesia pada makna Al-Qur'an.

2. Studi Terkait

2.1 *Similarity-Relatedness* Ruang Vektor

Seperti jurnal referensi, makalah ini akan mencoba menggunakan metode dari jurnal tersebut untuk mengumpulkan pasangan kata dalam mencari nilai kesamaan, dan hubungannya adalah Ruang *Sim-Rel Vector*. Tulisan ini akan mencoba membuktikan apakah kosakata bahasa Indonesia yang berasal dari Al-Qur'an menggunakan rumus fungsi. Sumbu x mewakili hubungan dengan skor r, dan sumbu y mewakili kesamaan dengan skor s dari setiap pasangan kata dalam setiap dataset. Pembagian grup ditandai dengan, SU (serupa-tidak terkait); SR (sejenis); DU (berbeda-tidak terkait); DR (berbeda-terkait) adalah label kategorikal dari kemungkinan semantik sub-ruang atau *ss*, kemudian fungsi $ss = f(r,s)$ adalah,

$$ss = f(r,s) \begin{cases} SU; & \text{if } s \geq 5 \; dan \; r < 5 \\ SR; & \text{if } s \geq 5 \; dan \; r \geq 5 \\ DU; & \text{if } s < 5 \; dan \; r < 5 \\ DR; & \text{if } s < 5 \; dan \; r \geq 5 \end{cases}$$

Gambar 2. Fungsi *Similarity-Relatedness* Ruang Vektor.

Kemudian metode ini juga menyediakan rumus fungsi dengan t = 2 yang menunjukkan ambang variabel yang mewakili jenis hubungan titik batas ruang dimana sinonim, antonim, *irrelevant* adalah label kategorik yang mungkin jenis hubungan semantik *rt*, *rt = f2 (r,s)*

$$f_2(r,s) = \begin{cases} synonym, & if \; 10-t \leq s \; and \; 10-t \leq r \\ antonym, & if \; 10-t \leq r \; and \; s \leq t \\ irrelevant, & if \; t \geq r \; and \; t \geq s \end{cases}$$

Gambar 3. Fungsi Tipe Relasi