Analisis Model Word2vec dalam Penyelesaian Soal Analogi pada Bahasa Indonesia

Abdul Raffi Malikul Mulki[1], **Moch. Arif Bijaksana**[2], **Arie Ardiayanti Suryani**[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]abdulraffi@students.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id,
[3]ardiyanti@telkomuniversity.ac.id,

Abstrak

Semantik adalah cabang ilmu lingustik dan salah satu komponen dalam suatu bahasa yang mempelajari arti atau makna suatu kata. Semantik kurang diperhatikan orang karena objek kajiannya berupa mak- na yang dianggap sangat sulit ditelusuri dan dianalisis strukturnya terutama untuk analogi suatu kata. Analogi kata merupakan cara untuk menunjukkan dua situasi yang didalamnya terdapat struktur rela- sional. Selain itu analogi kata memerlukan kemampuan kognitif yang lebih sedikit dan dapat digunakan diberbagai bidang. Maka dari itu Word2vec adalah solusi berupa model untuk merepresentasikan suatu kata menjadi vektor dengan besar dimensi yang ditentukan, sehingga dengan representasi word2vec dapat dilakukan operasi kesamaan dan keterkaitan antar kata. Word2vec telah banyak direkomendasikan dan digunakan pada penelitian pemerosesan bahasa alami, sehingga model ini menarik untuk dibahas dengan perbedaan konfigurasi pada model. Evaluasi yang dilakukan adalah membadingkan jawaban dari sistem dengan jawaban aktual dari persoalan analogi pada data tes. Hasil terbesar didari penelitian ini adalah 34% pada arsitektur *Skip-gram*, dimensi 100 dan *windows size* 10 serta 12. Hal ini dikarenakan jumlah korpus yang kecil serta distribusi kata pada koprus yang tidak merata.

Kata kunci : analogi, semantik, vector, word2vec

Abstract

Semantic is a branch of linguistics and one component in a language that learns the meaning of a word. The semantics are less noticed because the object of the study is in the form of meaning which is considered very difficult to trace and analyzed its structure especially for the analogy of word. Word analogy is a way to show two condition in which there is a relational structure. In addition, the analogy of words requires fewer cognitive abilities and can be used in various fields. Thus Word2vec is a solution in the form of a model to represent words into vectors with the dimensions specified. Word2vec has been widely recommended and used in natural language processing research, so this model is interesting to discuss with different configurations on the model. Evaluation is done by comparing the answers from the system with the actual answers to the problem of analogies on the datatest. The best results from this study is 34% on the *Skip-gram* architecture, dimension 100 and *windows size* 10 and *windows size* 12. This is due to the small number of corpus and the uneven distribution of words on the coprus.

Keywords: analogy, semantic, vector, word2vec

1. Pendahuluan

Dalam mencari nilai kesamaan dan keterkaitan perlu diketahui informasi berupa makna dari kata dalam bahasa. Dimana makna sangat bersifat arbitrer(berubah-ubah, tidak tetap), berbeda dengan morfem atau kata, sebagai sasaran dalam studi morfologi yang strukturnya tampak jelas dan dapat disegmen-segmenkan [1]. Oleh karena itu makna yang terkandung pada suatu bahasa, kode, atau jenis representasi lain yang dapat diketahui dan dipelajari oleh cabang ilmu linguistik [4].

Di era teknologi yang telah maju seperti sekarang ini, pemrosesan bahasa khususnya mencari nilai kesamaan dan keterkaitan kata telah banyak diminati. Namun, pemrosesan bahasa ini memerlukan komputasi pengolahan data yang cukup berat dikarenakan data teks yang tidak terstruktur serta mempunyai informasi yang kaya, dalam arti mempunyai fitur yang banyak dan berdimensi tinggi [13]. Bukan hanya itu, nilai untuk kesamaan dan ke- terkaitan suatu pasang kata atau analogi akan beragam karena dinilai dari sudut pandang yang berbeda dimana analogi kata merupakan cara untuk menunjukkan dua situasi yang didalamnya terdapat struktur relasional. Selain itu analogi kata memerlukan kemampuan kognitif yang lebih sedikit dan dapat digunakan pada bidang ekonomi, mengukur keterampilan untuk keberhasilan di perguruan tinggi, sekolah pasca sarjana dan bekerja [7]. Walaupun

secara manual, analogi sebuah kata dapat dicari dengan menggunakan kamus sebagai informasi [2] namun per- lu adanya sebuah penelitian untuk membuktikan bahwa analogi kata dapat dicari dengan sebuah sistem dengan metode tertentu.

Oleh karena itu dalam penelitian ini dibangun sebuah model untuk mengetahui nilai analogi kata dalam bahasa. Model yang dibangun yaitu sebuah word2vec untuk merubah suatu kata lalu digambar dengan vector (sesuatu yang berarah dalam dimensi). Representasi dari word2vec ini dapat menemukan arti pada kata dengan menyamakan arah dan panjang vektor. Model ini telah banyak digunakan untuk mempresentasikan kata dalam penelitian terbaru di bidang lingustik komputasi[6].

1.1 Latar Belakang

Semantik adalah cabang ilmu lingustik yang mempelajari arti atau makna dalam suatu bahasa. Semantik mempelajari arti atau makna yang terkandung pada suatu Bahasa, kode, atau jenis representasi lain [4]. Penilaian untuk semantik pada umumnya disebut *gold standart* yaitu nilai yang diberikan ahli pada bidangnya.

Nilai untuk kesamaan dan keterkaitan suatu pasang kata akan beragam karena dinilai dari sudut pandang ber- beda. Nilai yang diberikan dapat dibantu dengan adanya kamus sebagai informasi dari kata yang akan dinilai [2]. Informasi yang didapat dari kamus akan berupa makna dan perlu dipelajari dengan salah cabang ilmu linguistik (ilmu bahasa) yaitu semantik.

Model word2vec merubah suatu kata lalu digambar dengan vector (sesuatu yang berarah dalam dimensi). Dengan representasi dari word2vec dapat menemukan arti pada kata dengan menyamakan arah dan panjang vektor. Metode ini banyak banyak digunakan untuk mempresentasikan kata dalam penelitian terbaru di bidang lingustik komputasi atau text mining[6]. Sehingga metode ini perlu di eksplorasi agar dapat mencapai nilai akurasi dan kecepatan yang maksimal.

1.2 Topik dan Batasannya

Topik yang dibahas adalah sistem yang dapat menyelesaian soal analogi yang terdapat pada tes psikotes dengan batasan sebagai berikut :

1. Analisis efektifitas Word2Vec
2. Korpus yang digunakan bahasa Indonesia

1.3 Tujuan

Adapun tujuan yang ingin dicapai pada penelitian ini adalah analisis konfigurasi *windows size* dan dimensi pada model Word2vec dalam menyelesaikan soal analogi

1.4 Organisasi Tulisan

Pada jurnal ini mencakup pendahuluan, studi terkait, sistem yang dibangun, evaluasi, dan kesimpulan. Pada bagian pendahuluan terdapat latar belakang, topik dan batasannya serta tujuan dari penelitian ini. Bagian studi terkait berisi tentang materi yang mendukung pada jurnal ini. Bagian selanjutnya berisi tentang sistem yang dibuat untuk mencapai tujuan jurnal ini. Adapun hasil dari sistem yang dibangun akan dibahas pada bagian evaluasi. Pada bagian evaluasi akan terdapat dua sub-bagian yaitu hasil pengujian dan analisis hasil pengujian. Untuk bagian kesimpulan bersisi tentang pekerjaan selanjutnya untuk meningkatankan capaian dari jurnal ini.

2. Studi Terkait

Pada jurnal [7] menjelaskan bahwa analogi merupakan hal penting karena memiliki hubungan yang kuat antara kemampuan kognitif umum dan akuisisi pengetahuan dan keterampilan. Dalam penelitian pada paper ini disarank- an sebuah Miller Analogies Test (MAT) untuk mengetahui hubungan antara dua istilah yang diberikan dan mencari cara untuk mengetahui hubungan yang sama antara istilah tersebut. Jawaban yang benar harus dipilih dengan me- nyimpulkan hubungan antara dua istilah yang kemudian akan dipetakan ke pasangan yang dibentuk oleh istilah yang diberikan.

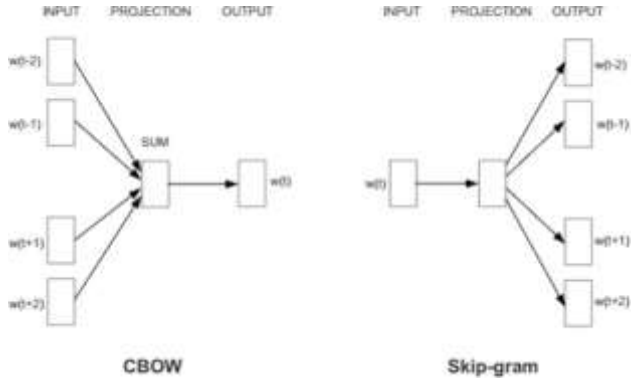
Semantik adalah ilmu bahasa yang mempelajari tentang makna [4]. Semantik dapat pula dapat dikatakan sebagai keterkaitan kata, sebagai contoh Jepang dan Indonesia memiliki makna yang sama dan keterkaitan yang sama pula yaitu negara. Sehingga soal analogi dapat dimasukan dalam sematik, contoh soal analogi ”Burung

: Terbang = Ikan : ?". Pada soal tersebut jawabannya tentu saja "Berenang" karena kata "Berenang" memiliki makna perilaku dengan kata "Terbang" dan keterkaitan dengan "Ikan" sebagai pelaku.

Jurnal[2] membahas tentang implementasi dan analisis kesamaan semantik pada Bahasa Indonesia dengan metode berbasis vector. Model yang dibangun menggunakan rumus kesamaan kosinus dan tf-idf sebagai pembobot. Aturan yang dibangun yaitu nilai korelasi pearson dengan nilai terbaik didapatkan dengan menambahkan pengaruh definisi sinonim dari kedua kata yang dibandingkan dan parameter terbaik mempengaruhi nilai kesamaan semantic.

Pada jurnal[6] membahsa mengenai self-training naïve bayes berbasis Word2Vec untuk kategori barita Bahasa Indonesia. Sistem yang dibangun yaitu menggunakan data berlabel dan tidak berlabel untuk membuat model klasifikasi. Hasil yang didapatkan dalam penelitian menggunakan model ini yaitu didapatkan nilai rata-rata dengan dimensi baik walaupun perubahan ukuran dari Word2Vec tidak berpengaruh kepada hasil klasifikasi secara signifikan. Ini disebabkan karena penentuan nilai batas ambang confidence juga menentukan kinerja dari algoritma self-training naïve bayes yang diusulkan.

Penelitian yang dilakukan oleh Mikolov dalam menghasilakan model *Word2Vec* yang berasal dari NNLM (Neural Network Language Modeling) untuk merubah kata menjadi vector[9]. Dalam melakukan vektorisasi terdapat nilai yang dibutuhkan yaitu jumlah *windows size* dan dimensi. Jumlah *windows size* menentukan jumlah kata sebelum dan sesudah dari kata yang akan di vektorisasi, sedangkan dimensi adalah jumlah titik pada hasil vektorisasi. *Word2Vec* dapat membantu memecahkan soal analogi yang sering muncul pada tes psikotes. Word2vec memili- ki dua arsitektur yaitu Skip-gram dan CBOW (Continuous Bag-of-Word). Pada gambar 1 terlihat perbedaannya, CBOW memprediksi kata dari konteks pada kata, sedang Skip-gram memprediski konteks dari kata. Serta pada tabel 1 merupakan hasil yang didapat oleh Mikolov dengan data latih sebanyak enam triliun kata.



Gambar 1. Arsitektur CBOW dan Skip-gram

Tabel 1. Hasil Uji Coba Word2Vec yang dilakukan oleh Mikolov [9]

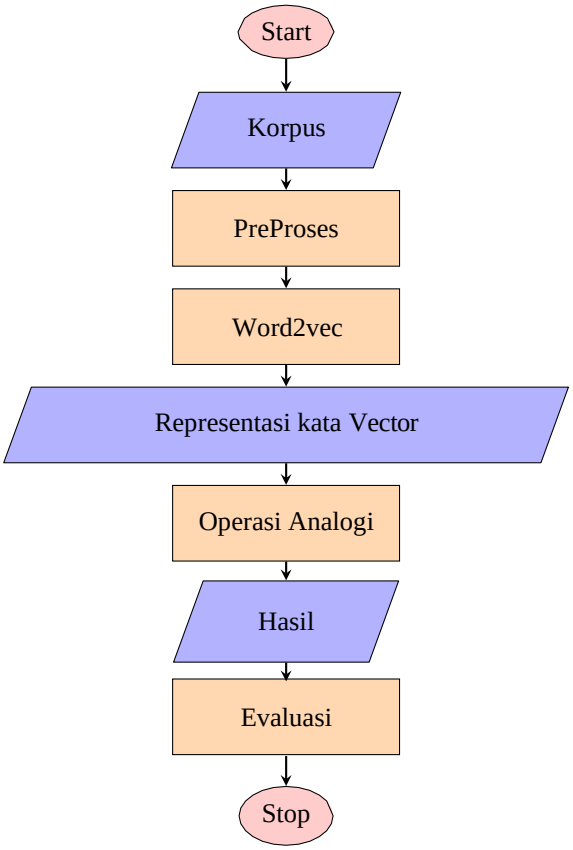
Model	Vector Dimensionality	Accuracy [%]			Training time [days x CPU cores]
		Semantic	Syntactic	Total	
NNLM	100	34.2	64.5	50.8	14 x 180
CBOW	1000	57.3	68.9	63.7	2 x 140
Skip-gram	1000	66.1	65.1	65.6	2,5 x 125

Korpus adalah kumpulan dokumen teks dari berbagai sumber, penulis dan tipe jenis tulisnya. Bentuk dokumen dapat berbentuk buku, artikel hingga ujaran. Dulu bentuk korpus adalah hard-copy yang artinya ada fisiknya, akan tetapi sekarang dapat berbentuk soft-copy yang dapat diproses secara otomatis atau semi otomatis [10]. Korpus yang paling banyak digunakan adalah Wikipedia karena Wikipedia tersedia dalam berbagai bahasa dan cakupannya luas. Data Test Set adalah dokumen yang sudah tersturuktur digunakan sebagai bahasan evaluasi dari keluaran sistem.

3. Sistem yang Dibangun

Pada gambar 2 menggambarkan sistem yang akan dibangun. Diawali dengan korpus yang memalui prap- roses untuk *transform case*. Setelah itu masuk ke proses vektorisasi yang digambarkan oleh model Word2Vec. Selanjutnya hasil vektor yang berupa model dapat digunakan pada operasi analogi sesuai soal yang ditentukan.

Masukan dalam model ini berupa pasang kata yang memiliki keterkaitan (premis), satu kata sebagai lawan dari keterkaitannya, serta pilihan untuk melengkapi keterkaitan lawan kata seperti dalam tes psikotes tentang sinonim kata.



Gambar 2. Diagram Arus Sistem Penyelesaian Soal Analogi

Operasi aljabar yang dimaksud sebagai contoh vektor(”king”) - vektor(”man”) + vektor(”women”) hasilnya mendekati vektor(”queen”)[9], sehingga sistem memberikan pilihan jawaban yang mendekati dengan keterkaitan pada soal. Penyelesaian soal analagi dapat diselesaikan dengan operasi aljabar tersebut menggukan model representasi dari Word2Vec.

3.1 Word2Vec Konfigurasi

Banyak parameter dan variabel yang dapat menentukan hasil akurasi dan efektifitas pada penggunaan sistem Word2Vec, sebagiai contoh jumlah dimensi, windows size serta lainnya. Dan pada penelitian ini difokus pada perbedaan arsitektur, dimensi, serta besar *window size*. Arsitektur Word2vec akan diuji kedua-duanya. Dimensi yang akan di uji pada dimensi 100, 400, dan 600, sedangkan untuk *windown size* pada 5, 8, dan 10.

3.2 Pre-processing

Pada penelitian ini dilakukan *text preprocessing* untuk data yang akan di vektorisasi guna meningkankan keterkaitan makna. Pada data tersebut akan menggunakan fitur *transform case* dengan fitur ini semua kata akan diubah menjadi huruf kecil semua[5] untuk menghidndari duplikasi kata dengan perpedaan kapital huruf.

Ada dua fitur lainnya pada *text preprocessing* yaitu fitur *stopword removal* dan fitur *stemming dan lemmatization*. Fitur *stopword removal* berfungsi untuk menghilangkan kata yang dikira kurang bermakna seperti yang, dia, saya, dsb. Pada penelitian ini fitur *stopword removal* diterapkan setelah mendapatkan hasil akurasi maksimal untuk melihat perubahan pada nilai akurasi. Dengan menerapkan fitur ini akan menggagalkan sistem menjawab satu persoalan karena pada data tes terdapat kata yang daftar di-*stopword* bahasa Indonesia[3]. *Stemming dan lemmatization* berfungsi untuk memotong kata berimbuhan seperti ”pemberani” menjadi ”berani”, sehingga didapatkan kata dasar yang akan meningkatkan makna pada penggunaan kata tersebut. Fitur ini juga tidak digunakan karena ada data tes yang memiliki kata berimbuhan, seperti keberanian, kemenangan, dsb.

3.3 Korpus

Wikipedia.org (id.wikipedia.org untuk bahasa Indonesia) adalah situs informasi umum yang ada dapat diakses, dan disunting oleh semua orang serta sudah mencapai empat ratus ribu lebih (400.000+) artikel berbahasa indo-neisa, apalagi Wikipedia sendiri memberikan akses bebas untuk mendapatkan data kumpulan artikel atau yang disebut korpus. Penulis menggunakan korpus Wikipedia bahasa Indonesia [12] serta artikel berita dari kompas dan tempo [8], dengan komposisi 693,5 MB dari 367.591 artikel dan 74,3 MB dari 29.177 artikel

3.4 Data Test

Data test berupa soal-soal yang biasanya terdapat pada soal psikotes analogi verbal. Pada tabel 2 merupakan contoh soal yakan dilakukan pengujian terhadapnya. Jumlah soal terdapat 61 (enam puluh satu) yang telah dipilih berdasarkan keberadaaan kata pada soal dengan kata pada korpus. Data test yang akan digunakan dalam penelitian adalah soal psikotes verbal[11].

Tabel 2. Data Test

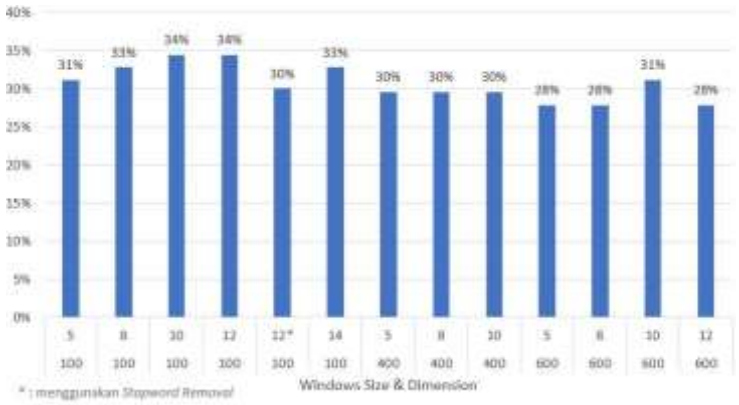
Premis	Kesimpulan	Pilihan
mobil : bensin	pelari : . . .	makanan; sepatu; kaos; lintasan
busur : panah	senapan : . . .	peluru; senjata; berbahaya; tembakan
tambang : emas	laut : . . .	badai; kapal; nelayan; karang
ramalan : astrologi	bangsa : . . .	etnologi; psikologi; demografi; antropologi
perusahaan : karyawan	sekolah : . . .	pengawas; pelajar; ujian; kelas
.	.	.
.	.	.

4. Evaluasi

Bagian ini berisi dua sub-bagian, yaitu Hasil Pengujian dan Analisis Hasil Pengujian. Pengujian dan analisis yang dilakukan selaras dengan tujuan TA sebagaimana dinyatakan dalam Pendahuluan.

4.1 Hasil Pengujian

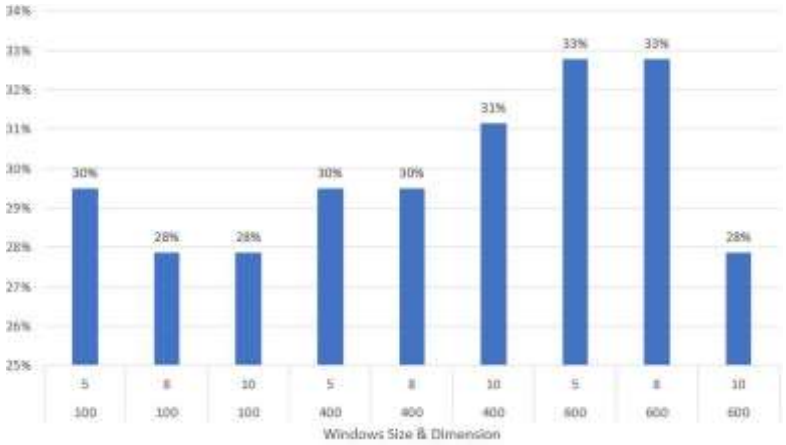
Pada gambar 3 dan gambar 4 yang disusun berdasarkan dimensi, dikarenakan perubahan yang cukup signi-fikan adalah perubahan tiap *windows size*. Hasil yang didapat pada arsitektur *Skip-gram* lebih besar dari 1% dari pada arsitektur CBOW sebesar 33% pada konfigurasi *windows size* 8 dan dimensi 600 sedangkan *skip-gram* pada *windows size* 12 dan dimensi 100.



Gambar 3. Hasil Pengujian dengan Arsitektur Skip-gram pada variasi *Windows Size* dan Dimensi

Pada pengujian arsitektur *skip-gram* penulis mencoba konfigurasi *windows size* lebih besar dari rencana pengu-jian dikarekan pada dimensi 100 setiap penambahan *windows size* mendapatkan nilai yang meningkat. Akan tetapi hal tersebut tidak benar karena pada *windows size* 14 mengalami penurunan sebesar 1%. Pada pengujian arsitektur CBOW penulis tidak mencoba penambahan konfigurasi karena nilainya akurasinya lebih kecil. Selanjutnya pada

konfigurasi dengan nilai terbesar diterapkan praproses fitur *stopword removal* didapatkan nilai akurasi menurun pada nilai 30% dengan menggagalkan satu soal seperti yang dijelaskan pada poin 3.2 tentang *pre-procrssing*.



Gambar 4. Hasil Pengujian dengan Arsitektur CBOW pada variasi *Windows Size* dan Dimensi

4.2 Analisis Hasil Pengujian

Untuk menganalisa konfigurasi dengan nilai terbaik diperlukan membandingkan setiap nilai akurasi pada konfigurasi yang dilakukan. Membandingkan grafik pada gambar 3 dan gambar 4 nilai akurasi yang terbesar terdapat pada arsitektur *skip-gram* dengan konfigurasi *windows size* 12 dan 10 pada dimensi 100. Jawaban pada dua konfigurasi tersebut hanya berbeda pada empat soal. Pada empat soal tersebut masing-masing dua soal di jawab benar. Sembilan belas soal dijawab benar dan sama, tiga puluh empat dijawab salah dan sama, dan empat soal dijawab salah dan bebeda oleh kedua konfigurasi tersebut. Nilai akurasi yang kecil ada kemungkinan sedikitnya konteks pada korpus.

Konteks pada korpus sangat berpengahur besar, hal ini penulis temutakan pada soal nomor 58(lima puluh delapan) dilampiran satu. Soal menanyakan keterkaitan ”bangsa” dan ”etnologi” dengan premis ”ramalana : astrologi”. Konteks kalimat yang berhubungan dengan soal tersebut adalah ”Etnologi adalah ilmu yang mempelajari asas kebudayaan manusia di dalam kehidupan masyarakat suku bangsa di seluruh dunia”. Walaupun jumlah kata ”etnologi” hanya sebanyak 150(seratus lima puluh), soal tersebut dapat dijawab oleh sistem dengan semua konfigurasi terlampir.

Berbeda dengan soal nomor satu yang mempertanyakan keterkaitan ”pelari” dan ”makanan” dengan premis ”mobil : bensin”. Jumlah kata ”pelari” sebanyak 425(empat ratus dua puluh lima) tetapi tidak berdekatan dengan kata ”makanana” maka sistem pun tidak dapat menjawab soal tersebut.

5. Kesimpulan

Berdasarkan penelitian ini, konfigurasi *windows size* dan dimensi, serta konteks kalimat pada korpus sangat berpengaruh besar pada nilai akurasi. Nilai akurasi yang hanya mencapai 34% pun, itu disebabkan oleh korpus yang digunakan pada penelitian ini kurang beragam *genre* (aliran; fiksi, majalah, dll), sehingga korpus yang digunakan sangat kurang akan konteks. Seperti yang telah dijabarkan bahwa kontesk sangat berpengaruh. Maka daripada itu untuk penelitian selanjutnya diharapkan memperbanyak *genre* korpus sehingga korpus yang digunakan kaya akan konteks.

Daftar Pustaka

[1]C. Abdul. Korpus dalam kajian penerjemahan. In *Pengantar semantik bahasa Indonesia*. Rineka Cipta, 1990.

[2]R. F. Ade Romadony, Said Al Faraby. Implementation and analysis of semantic similarity on bahasa indonesia by vector-based method. *e-Proceeding of Engineering*, 4:4641, 2017.

[3]K. R. Andy Librian. Sastrawi python. Accessed: 2019-06-21.

[4]K. Aris. Pengertian semantik dan contohnya lengkap. Accessed: 2019-03-21.

[5]J. H. Aris Tri. Preprocessing text untuk meminimalisir kata yang tidak berarti dalam proses text mining. In *Jurnal Infromatika UPGRIS*. Universitas PGRI Semarang, 2015.

[6]Dkk and J. Santoso. Self-training naive bayes berbasis word2vec untuk kategorisasi berita bahasa indonesia. *JNTETI*, 7:158, 2018.

[7]E. Don Meagher. Korpus dalam kajian penerjemahan. In *Understanding Analogies*, 2006.

[8]K. Kurniawan. Indonesian nlp resources. Accessed: 2019-02-22.

[9]T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10]T. Setiawan. Korpus dalam kajian penerjemahan. In *Dipresentasikan pada Seminar Nasional Perspektif Baru Penelitian Linguistik Terapan: Linguistik Korpus dalam Pengajaran Bahasa, Yogyakarta: UNY*, 2017.

[11]C. Team. Tpa (psikotes) - model dan pelajaran verbal analogy korelasi makna. Accessed: 2019-05-06.

[12]T. Wikipedia. Wikimedia download. Accessed: 2019-02-22.

[13]P. Yulius Denny, M. Tedi Lesmana, and S. Meylisa. Pembentukan vector space model bahasa indonesia menggunakan metode word to vector. *Jurnal Buana Informatika*, 10.1:29–40, 2006.