Analisis Word2vec untuk Perhitungan Kesamaan Semantik antar Kata

Nabila Nanda Widyastuti[1], Arif Bijaksana, Ir., M.Tech., Ph.D[2], Indra Lukmana Sardi, S.T., M.T[3]

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

Abstrak

Implementasi perhitungan kesamaan semantik antar kata merupakan salah satu tugas yang dapat diselesaikan dalam bidang Natural Language Processing(NLP). Perhitungan kesamaan semantik antar kata dapat digunakan untuk membantu mesin dalam memahami bahasa manusia. Selain itu, perhitungan kesamaan semantik juga dapat digunakan sebagai dasar penelitian tahap selanjutnya pada bidang NLP. Penelitian ini dilatar belakangi oleh suatu masalah dimana pada saat ini pencarian sistem informasi banyak melibatkan teks atau dokumen, namun mesin belum dapat menyamakan persepsi manusia dengan baik sehingga mesin perlu dibantu untuk memahami teks atau dokumen tersebut. Sepasang kata dinyatakan mempunyai kesamaan semantik apabila memiliki kesamaan pada makna atau konsep. Pada penelitian ini, dilakukan implementasi perhitungan kesamaan semantik antar kata untuk bahasa Inggris. Korpus yang digunakan pada penelitian ini yaitu Brown Corpus, Berita Corpus, dan Harry Potter Corpus. Dokumen tersebut diubah kedalam bentuk vektor dengan Word2vec. Selanjutnya nilai kesamaan semantik yang dihasilkan dari vektor tersebut dibandingkan dengan dataset Gold Standard SimLex999 untuk mengukur nilai korelasinya. Hasil pengujian menunjukkan bahwa pengukuran Word2vec menghasilkan korelasi sebesar 0.192 dengan perhitungan korelasi Pearson.

Kata kunci : Kesamaan Semantik, Natural Language Processing, Word2vec

Abstract

The implementation of calculation semantic similarity between word is one of task that can be done by Natural Language Processing. The calculation of semantic similarity between word can used to help the machine to understanding of human language(natural language). Beside that, calculation of semantic similarity can be used as a basic of the next step in NLP's research. The main idea of this study is motivated by a problem where nowaday the seaarch of information sistem are involved by many text and document, so we need to help the machine to understand those texts or documents. A pair of word are similar if they have similarity to the level of meaning of concept. In this research, we are implement the calculation of semantic similarity between word in English. The corpus that used in this research are Brown Corpus, Berita Corpus, and Harry Potter Corpus. That documents are convert into vector space by using Word2vec. Next, the score of semantic similarity generated by vector are compared to SimLex999 Gold Standard dataset to measure their corelation. The result showed that Word2vec have corelation's score of 0.192 in Pearson corelation.

Keywords: Semantic Similarity, Natural Language Processing, Word2vec

1. Pendahuluan

Latar Belakang

Kesamaan semantik yaitu merupakan suatu pengukuran untuk mencari nilai yang menyatakan tingkat kesamaan atau kedekatan secara semantik, kalimat, atau teks. Sepasang kata dinyatakan semantik apabila memiliki kesamaan dari sisi makna atau konsep [11]. Perhitungan kesamaan semantik dilatar belakangi oleh suatu masalah dimana mesin belum dapat menyamakan persepsi manusia dengan baik, untuk itu perhitungan kesamaan semantik digunakan untuk membantu mesin memahami bahasa manusia. Selain untuk membantu mesin memahami bahasa manusia, perhitungan kesamaan semantik juga dapat digunakan dalam *Natural Language Processing* (NLP) untuk menangani masalah ambiguitas[13]. Selain ambiguitas, perhitungan kesamaan semantik juga dapat digunakan untuk penyelesaian masalah pada *information retrieval* seperti masalah pencarian, *query suggestion*, *automatic summarization*, dan pencarian gambar [5]. Kesamaan semantik juga merupakan isu pokok pada penelitian dalam berbagai bidang ilmu seperti Psikologi, Linguistik, Ilmu Cognitif, Biomedis, dan Kecerdasan Buatan [12].

Tugas yang dilakukan oleh kesamaan semantik yaitu untuk menghitung nilai kedekatan antara dua buah kata. Karena mesin hanya dapat membaca angka, maka kata-kata yang ada harus diterjemahkan dalam bentuk angka terlebih dahulu. Untuk mencari nilai kedekatan antara dua buah kata dapat dilakukan dengan barbagai cara, salah satunya yaitu dengan menghitung nilai vektor kata tersebut. Untuk mendapatkan nilai kesamaan semantik dengan vektor yaitu dengan cara menghitung perbedaan sudut antara dua buah vektor dengan rumus *cosine similarity*. Salah satu metode berbasis vektor untuk mencari nilai kesamaan semantik yaitu Word2vec, dimana Word2vec ini merupakan metode yang baru dikembangkan pada tahun 2013. Word2vec dengan pemodelan Skip-gram merupakan metode yang efisien untuk melatih representasi vektor terdistribusi berkualitas tinggi yang menangkap hubungan kata sintatik dan semantik yang tepat dalam jumlah besar[9].

Pada tugas akhir ini dilakukan perhitungan nilai kesamaan semantik antara dua buah kata dengan berbasis vektor. Metode yang digunakan yakni metode Word2vec, karena metode ini dapat memproses kedekatan vektor kata-kata dan dinilai memiliki nilai performansi yang baik [9]. Word2vec yaitu metode yang dapat menghitung nilai vektor hubungan antara sepasang kata atau lebih. Input dari Word2vec yaitu berupa korpus, sedangkan outputnya berupa vektor kata yang selanjutnya dapat menghasilkan nilai kesamaan semantik yang dihasilkan dari perhitungan cosine similarity. Pada tugas akhir ini pengukuran Word2vec akan diimplementasikan pada sejumlah pasangan kata dalam tiga dataset berbeda. Sistem yang dibangun diharapkan dapat menghasilkan performansi yang baik berdasarkan nilai korelasi yang dihitung menggunakan acuan data SimLex999 sebagai gold standard.

Topik dan Batasannya

Berdasarkan latar belakang masalah yang telah diuraikan diatas, maka perumusan masalahnya adalah sebagai berikut :

1. Bagaimana analisis penggunaan Word2vec untuk perhitungan kesamaan semantik antarkata?
2. Bagaimana nilai korelasi yang dihasilkan dari penggunaan metode Word2vec untuk perhitungan kesamaan semantik antar kata jika dibandingkan dengan gold standard?
3. Apa sajakah yang dapat mempengaruhi hasil korelasi yang dihasilkan dari penggunaan Word2vec?

Adapun batasan masalah pada tugas akhir ini adalah,

1. Model Word2vec yang digunakan hanya model Skip-gram.
2. Sistem yang dibuat hanya diaplikasikan untuk menghitung nilai kesamaan semantik untuk sepasang kata.
3. Dataset yang digunakan adalah SimLex999 yang memuat 999 pasang kata.
4. Keluaran dari sistem ini adalah skor kesamaan semantik antar sepasang kata.
5. Skor kesamaan semantik yang dihasilkan oleh setiap pasangan memiliki skala 0-1.

Tujuan

Berdasarkan perumusan masalah di atas, maka tujuan dari tugas akhir ini adalah sebagai berikut:

1. Menganalisis penggunaan Word2vec untuk perhitungan kesamaan semantik antar sepasang kata.
2. Melakukan perhitungan korelasi antara nilai kesamaan yang dihasilkan sistem dengan nilai pada gold standard.
3. Menganalisis faktor-faktor yang dapat mempengaruhi hasil korelasi dari Word2vec dengan *gold standard*.

Organisasi Tulisan

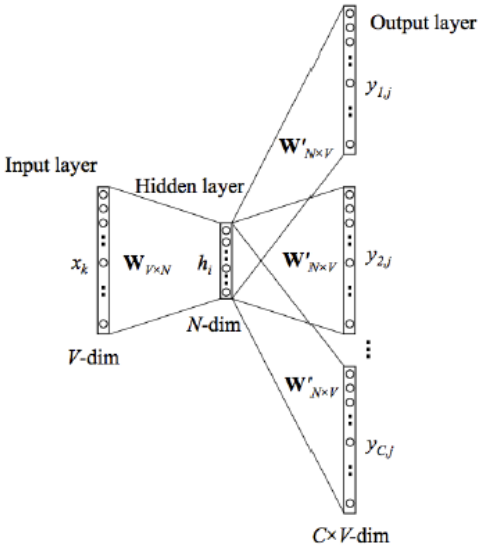
Organisasi penulisan laporan ini adalah sebagai berikut: Bagian 2 menunjukkan studi-studi terkait dengan tugas akhir ini. Bagian 3 yaitu menjelaskan sistem yang akan digunakan untuk perhitungan kesamaan semantik antar kata menggunakan Word2vec. Bagian 4 akan membahas tentang hasil serta analisis pengujian. Dan kesimpulan akan dibahas pada bagian 5.

2. Kajian Pustaka

2.1 Word2vec

Word2vec merupakan suatu alat yang baru dikembangkan oleh Thomas Mikolov. Word2vec dapat mengolah kata-kata dari *dataset* yang sangat besar dalam waktu yang relatif singkat dengan nilai akurasi yang lebih baik

dibandingkan dengan alat yang pernah ada sebelumnya [10]. Cara kerja alat ini yaitu dengan mengambil korpus teks sebagai input, lalu menghasilkan representasi vektor setiap kata yang ada pada korpus teks tersebut sebagai output [10]. File vektor yang dihasilkan dapat digunakan untuk penelitian pada pemrosesan bahasa alami dan aplikasi pembelajaran mesin. Vektor kata tersebut juga dapat digunakan untuk mengukur jarak kedekatan antar vektor kata yang lain. Word2vec memiliki dua arsitektur pemodelan yang dapat digunakan untuk merepresentasikan vektor kata, arsitektur tersebut yaitu *continous bag-of-word* (CBOW) dan Skip-gram. Untuk penelitian kali ini yang penulis gunakan yaitu arsitektur Skip-gram. Arsitektur ini dapat dilihat pada Gambar 1.



Gambar 1. Arsitektur Skip-gram [8]

Arsitektur Skip-gram bekerja dengan tiga layer, layer yang pertama yaitu layer input, layer yang kedua yaitu hidden layer, dan layer yang ketiga yaitu output layer. Di antara input layer dengan hidden layer terdapat matriks weight (W) yang didapatkan dari nilai random, layer ini berfungsi untuk mengaktifkan hidden layer. *Output* layer dihasilkan dari perhitungan *hidden* layer dengan matriks weight'(W'). Setelah itu menggunakan Softmax funtion pada setiap neuron yang dihasilkan oleh layer output untuk mengubah total jumalah *output* menjadi bernilai antara 0 - 1, dan jumlah dari semua nilai output akan bertambah sehingga nilainya 1[15]. Persamaan 1 Softmax function adalah sebagai berikut :

$$Softmaxfunction = \frac{e^x}{\sum e^x} \tag{1}$$

Dimana x adalah nilai ouput yang dihasilkan oleh pemodelan Skip-gram.

2.2 *Gold standard*

Gold standard merupakan suatu nilai hasil dari sekumpulan pendapat manusia yang dijadikan sebagai acuan dalam proses pengukuran kesamaan semantik diantara pasangan teks maupun kata dalam sekala tertentu. Pada penelitian ini, SimLex999 dijadikan sebagai acuan *gold standard*. SimLex999 merupakan suatu set data untuk mengevaluasi suatu model yang mempelajari makna kata dan konsep [2]. SimLex999 mengukur seberapa baik model tersebut menghitung kesamaan antara dua konsep. SimLex999 yang dijadikan sebagai acuan yaitu berupa skor kesamaan semantik pasangan kata. Skor tersebut akan dibandingkan dan dihitung nilai korelasinya dengan keluaran dari sistem yang telah dibangun.

2.3 *Cosine Similarity*

Cosine similarity yaitu sebuah metode untuk menghitung kesamaan antara dua vektor (atau dokumen pada ruang vektor) . Perhitungan dilakukan dengan menghitung kosinus sudut diantara keduanya. Persamaan untuk menghitung *cosine similarity* antara dua vektor yaitu seperti pada persamaan 2 berikut:

$$Sim = cos(\theta) = \frac{A.B}{||A||||B||} \tag{2}$$

Dimana A adalah vektor kata input pertama dan B adalah vektor kata input kedua.

2.4 Korelasi Pearson

Korelasi Pearson merupakan salah satu ukuran korelasi yang digunakan untuk mengukur hubungan linier antara dua variabel yang didefinisikan sebagai kovarian dari variabel dibagi dengan standar deviasinya[7]. Perhitungan korelasi menggunakan Korelasi Pearson umumnya digunakan pada data yang bersifaat kuantitatif (data berskala interval atau rasio) dan kedua variabel merupakan bivariat yang berdistribusi normal [7]. Adapun formula yang digunakan dalam perhitungan korelasi pada Korelasi Pearson menggunakan persamaan 3 berikut:

$$Pearson(x,y) = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{N\sum x^2 - (\sum x)^2 N\sum y^2 - (\sum y)^2}} \quad (3)$$

Dengan:
N = jumlah pasangan kata
x = nilai dari sistem
y = nilai dari gold standard

Nilai X dan Y dapat diabaikan apabila salah satu dari nilai variabel null(tidak memiliki nilai). Terdapat tiga kategori korelasi berdasarkan apabila nilai variabel pertama meningkat apakah variabel yang kedua turut meningkat atau tidak [7] :

- Korelasi Positif : nilai variabel kedua cenderung meninkat.
- Korelasi Negatif : nilai variabel kedua cenderung menurun.
- Tidak ada Korelasi: nilai variabel kedua tidak memiliki kecenderungan apakah meningkat atau menurun.

Adapun kriteria keterhubungan dapat dilihat pada tabel 1

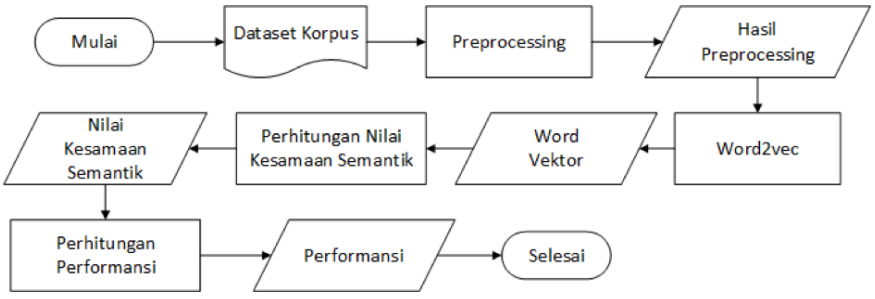
Tabel 1. Korelasi hubungan

Range Keterhubungan	Kriteria Hubungan
0	Tidak ada korelasi
0 - 0.5	Korelasi lemah
0.5 - 0.8	Korelasi sedang
0.8 - 1	Korelasi kuat/erat
1	Korelasi Sempurna

3. Sistem yang Dibangun

3.1 Gambaran Umum Sistem

Sistem yang dibangun bertujuan untuk menghasilkan nilai kesamaan semantik antar dua pasanagan kata. Nilai kesamaan semantik diperoleh berdasarkan pengimplementasian metode Word2vec. Evaluasi yang dilakukan yaitu dengan menggunakan perhitungan korelasi sebagai tolok ukur kemiripan sistem yang di bangun dengan nilai *gold standar*. Gambaran umum sistem yang dibangun dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur Skip-gram

Dari Gambar 2 diatas maka tahapan gambaran umum sistem adalah sebagai berikut:

1. Sistem membaca data *input* berupa *dataset* korpus.
2. Sistem melakukan tahapan *preprocessing*, yaitu dengan melakukan tahapan *case folding*, pembersihan data, tokenisasi dan *stopword removal* terhadap setiap file *input*.
3. Sistem melakukan perhitungan Word2vec yang diawali dengan penentuan *window size* dan dimensi vektor. Setelah itu melatih pembuatan model untuk setiap kata pada korpus.
4. 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 format file (.csv).
5. Sistem melakukan perhitungan korelasi dari nilai kesamaan semantik Word2vec dibandingkan dengan nilai *gold standard*.

3.2 Perancangan Sistem

3.2.1 *Dataset*

Dataset korpus yang digunakan pada penelitian ini yaitu menggunakan *dataset* korpus Brow memuat sekitar satu juta kata yang dibangun di *Brown University* pada tahun 1964 [6] korpus ini adalah korpus yang paling besar diantara dua korpus yang lain. Korpus selanjutnya adalah *dataset* Harry Potter dimana didapatkan dari novel Harry Potter The Sorcerer's Stone karya JK Rowling [14]. Korpus terakhir yaitu *dataset* korpus Berita yang didapatkan dari artikel berita dari web Pemprov DKI Jakarta yaitu Berita Jakarta yang dimuat sejak bulan November 2012 hingga Apil 2013 [3]

3.2.2 *Preprocessing*

- *Case Folding*
  Data pada keseluruhan teks dalam korpus diubah menjadi suatu bentuk standar (biasanya huruf kecil).
  Masukan : From those 43 teachers, only 11 of them have the status of PNS, while the rests are private teachers taken from Surya Institute with payroll Rp 2.06 billion.
  Keluaran : from those 43 teachers, only 11 of them have the status of pns, while the rests are private teachers taken from surya institute with payroll rp 2.06 billion.

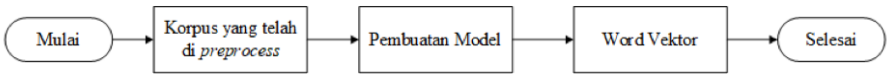
- Pembersihan Data
  Data akan dibersihkan dari seluruh atribut, sehingga data menjadi tanpa atribut seperti angka dan tanda baca.
  Masukan: from those 43 teachers, only 11 of them have the status of pns, while the rests are private teachers taken from surya institute with payroll rp 2.06 billion.
  Keluaran : from those teachers only of them have the status of pns while the rests are private teachers taken from surya institute with payroll rp billion

- Tokenisasi
  Tokenisasi dilakukan untuk memotong setiap kalimat menjadi beberapa token/bagian. Sehingga kalimat *input* menjadi kumpulan kata dalam list.
  Masukan : from those teachers only of them have the status of pns while the rests are private teachers taken from surya institute with payroll rp billion
  Keluaran : ['from', 'those', 'teachers', 'only', 'of', 'them', 'have', 'status', 'of', 'pns', 'while', 'the', 'rests', 'are', 'private', 'teachers', 'taken', 'from', 'surya', 'institute', 'with', 'payroll', 'rp', 'billion']

- *Stopword Removal*
  Setelah melakukan tokenisasi, langkah selanjutnya yaitu menghilangkan kata yang tidak perlu atau disebut dengan stopword removal, misalnya the, of, dsb. Tahap ini dilakukan supaya mengurangi waktu training, karena kata-kata tersebut memiliki jumlah yang sangat banyak pada dokumen.
  Masukan : ['from', 'those', 'teachers', 'only', 'of', 'them', 'have', 'status', 'of', 'pns', 'while', 'the', 'rests', 'are', 'private', 'teachers', 'taken', 'from', 'surya', 'institute', 'with', 'payroll', 'rp', 'billion']
  Keluaran : ['teachers', 'status', 'pns', 'rests', 'private', 'teachers', 'taken', 'surya', 'institute', 'payroll', 'rp', 'billion']

3.2.3 Word2vec

Word2vec digunakan untuk mengubah kata-kata menjadi bentuk vektor dengan tujuan untuk mencari nilai kedekatan vektor antar kata. Urutan proses Word2vec dapat dilihat pada Gambar 3 dibawah ini.



Gambar 3. Gambaran Pemodelan Word2vec

1. Membaca Korpus Sistem membaca seluruh isi dari data korpus yang sudah dilakukan proses preprocessing. Dimana data yang dibaca yaitu berupa kata-kata pada suatu kalimat yg telah diubah kedalam bentuk array.
2. Pembuatan Model
   (a) Membangun konteks pasangan kata dari data korpus dengan berdasarkan jumlah *window size*. Pada penelitian seblumnya [10] menyatakan bahwa *window size* 5 memiliki hasil yang optimal. Untuk penulis menetapkan *window size* sebesar 7, 9, dan 11, untuk mengetahui nilai kesamaan semantik yang dihasilkan apabila *window size* yang digunakan semakin besar. Apabila ditemukan konteks pasangan kata pada window size tersebut maka frekuensi kata ditambah 1. Sistem hanya akan menghitung konteks pasangan kata yang memiliki frekuensi kemunculan minimal sebanyak tiga kali.
   (b) Setelah itu melakukan training untuk mengubah data menjadi bentuk *one-hot-vector*. Hal ini dilakukan untuk mengubah bentuk dari setiap kata pada dataset menjadi bentuk *binary vector*.
   (c) Langkah selanjutnya yaitu sistem melatih model untuk memprediksi vektor kata input berdasarkan konteks kata disekitarnya dengan satu *hidden layer* dan dimensi vektor 300 karena pada penelitian sebelumnya [10] dikatakan bahwa dimensi vektor tersebut memiliki hasil yang optimal.
   (d) Dari *hidden layer* dihasilkan matriks *output*, kemudian matriks tersebut diubah dengan *Softmax function* untuk mendapatkan *Word Vector*.
3. Word Vector
   Setelah proses pembuatan model selesai, maka sistem menghasilkan vektor-vektor dari setiap kata dari data korpus. Di dalam Word2vec, setiap satu kata bisa memiliki lebih dari satu vektor hal ini dikarenakan setiap kata pada sebuah kalimat memiliki konteks yang berbeda.

3.2.4 Perhitungan Kesamaan Semantik

Setelah dilakukan proses perhitungan vektor dengan Word2vec, maka sistem akan membaca vektor dua buah kata input lalu melakukan perhitungan kesamaan semantik dengan menggunakan rumus *cosine similarity*.

3.2.5 Perhitungan Performansi

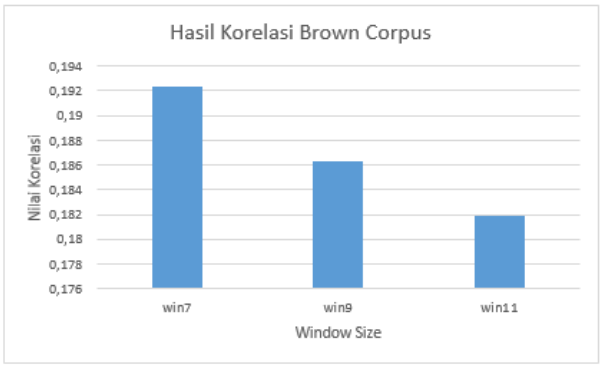
Setelah mendapatkan hasil dari perhitungan kesamaan semantik dari sistem, maka dilakukan perhitungan performansi dengan membandingkan nilai tersebut terhadap nilai *gold standard* penggunakan korelasi Pearson.

4. Evaluasi

4.1 Hasil Pengujian

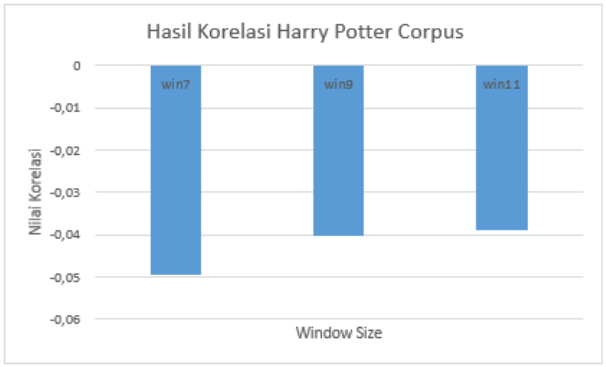
4.1.1 Hasil Pengujian Skenario 1

Pengujian pertama yang dilakukan yaitu menganalisis hubungan kesamaan semantik antar pasangan kata dengan Word2vec berdasarkan pada ukuran *window size*. Tiga *window size* yang digunakan dalam penelitian ini yaitu window size 7, window size 9, dan window size 11, serta korpus yang digunakan yaitu Brown Corpus, Harry Potter Corpus, dan Berita Corpus. Nilai hasil keluaran sistem dari masing-masing *window size* akan dibandingkan dengan nilai dari gold standard dengan perhitungan korelasi Pearson. Hasil pengujian dapat dilihat pada gambar 4, 5, 6 :



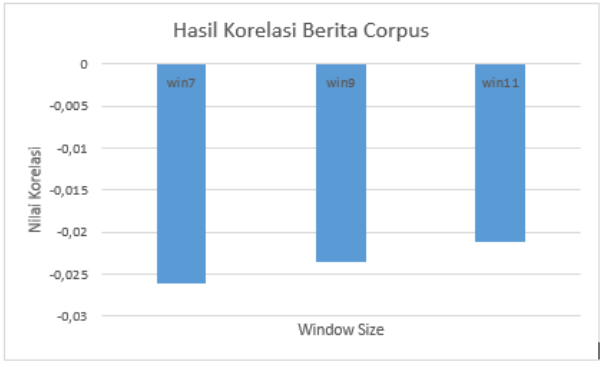
Gambar 4. Korelasi Word2vec dengan Brown Corpus

Pada gambar 4 dapat dilihat bahwa hasil pengujian menggunakan korpus Brown menunjukkan bahwa *window size* 7 memiliki nilai korelasi paling tinggi diantara dua *window size* yang lain yaitu nilai korelasi *window size* 7 sebesar 0.19235, *window size* 9 sebesar 0.18636 dan nilai korelasi yang paling kecil yaitu terdapat pada *window size* 11 sebesar 0.18196.



Gambar 5. Korelasi Word2vec dengan Harry Potter Corpus

Pada gambar 5 dapat dilihat bahwa hasil pengujian menggunakan korpus Harry Potter menunjukkan bahwa *window size* 7 memiliki nilai korelasi paling tinggi diantara dua *window size* yang lain yaitu nilai korelasi *window size* 7 sebesar -0.0495, *window size* 9 sebesar -0.0402 dan nilai korelasi yang paling kecil yaitu terdapat pada *window size* 11 sebesar -0.0388.

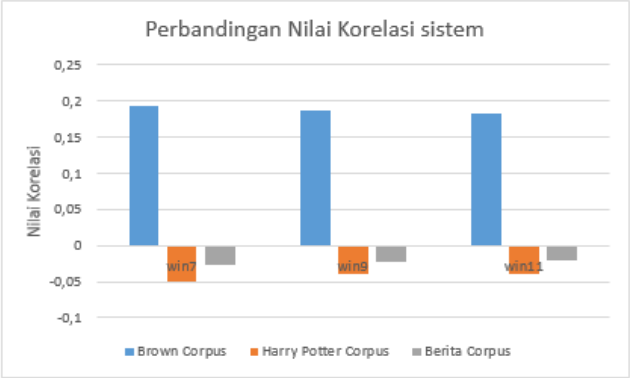


Gambar 6. Korelasi Word2vec dengan Berita Corpus

Pada gambar 6 dapat dilihat bahwa hasil pengujian menggunakan korpus Berita menunjukkan bahwa *window size* 7 memiliki nilai korelasi paling tinggi diantara dua *window size* yang lain yaitu nilai korelasi *window size* 7 sebesar -0.0261, *window size* 9 sebesar -0.0235 dan nilai korelasi yang paling kecil yaitu terdapat pada *window size* 11 sebesar -0.0211.

4.1.2 Hasil Pengujian Skenario 2

Pengujian dilakukan untuk mengetahui perbandingan nilai korelasi yang dihasilkan sistem (Word2vec) dari tiga korpus data berbeda. Masing-masing korpus data tersebut yaitu korpus Brown, korpus Harry Potter, dan korpus Berita. Hasil pengujian dapat dilihat pada gambar 7.



Gambar 7. Perbandingan Korelasi antar korpus

Pada gambar 7 dapat dilihat nilai korelasi tertinggi dari hasil pengujian yang telah dilakukan dengan *window size* yang sama yaitu terdapat pada pengujian dengan korpus Brown, lalu diikuti oleh nilai korelasi korpus Harry Potter dan nilai terkecil yaitu terdapat pada korpus Berita.

4.1.3 Hasil Pengujian Skenario 3

Pengujian ini dilakukan untuk mengetahui perbandingan dari kesamaan kata antara keluaran sistem dari tiga korpus berbeda dengan WordNet. WordNet merupakan database leksikal bahasa Inggris yang dibangun oleh Princetion University pada tahun 1986[1]. Contoh hasil pengujian dapat dilihat pada tabel 2 berikut:

Tabel 2. Contoh perbandingan pengujian kesamaan semantik dengan WordNet

Kata Input	WordNet	Brown W2V	Berita W2V	HarryPotter W2V
car	auto, automobile, machine, motorcar, elevator car	truck, yard, ahead, bench, parked	motorcycle, box, wheel, grazed, trailer	crash, lift, parked, telephone, maze
old	erstwhile, former, onetime, one-time, quondam	young, hans, handley, comedy, ladies	replaced, age, unhealty, computers, worn	fashioned, oldest, aged, elderly, born
salary	wage, pay, earnings, remuneration, salary	estimate, finance, earnings, award, improvement	wage, minimum, paid, raised, debts	-
good	honorable, skillful, well, honest, safe	fun, luck, nice, fine, honest	spirit, goal, successful, easy, okay	nice, amusing, gracious, fair, clever

4.2 Analisis Hasil Pengujian

Proses perhitungan nilai kesamaan semantik antar kata menggunakan Word2vec dilakukan dengan korpus data dengan ukuran yang berbeda-beda, dan menggunakan ukuran window size yang berbeda. Hal ini bertujuan untuk mengetahui apa saja faktor-faktor yang dapat mempengaruhi hasil nilai dari penggunaan Word2vec untuk perhitungan kesamaan semantik antar kata. Perbandingan hasil performansi dari Word2vec dapat dilihat pada tabel 3 berikut:

Tabel 3. Hasil perbandingan korelasi Kesamaan Semantik dengan Word2vec

Korpus	Window Size	Nilai Korelasi
Brown	7	0.19235
Brown	9	0.18636
Brown	11	0.18196
Harry Potter	7	-0.0495
Harry Potter	9	-0.0402
Harry Potter	11	-0.0388
Berita	7	-0.0261
Berita	9	-0.0235
Berita	11	-0.0211

Hasil pada tabel 3 menunjukkan bahwa nilai performansi dari sistem dapat dipengaruhi oleh beberapa faktor. Berikut beberapa hal yang mempengaruhi nilai perhitungan kesamaan semantik dengan Word2vec:

1. Penggunaan window size. Hasil korelasi Word2vec menunjukkan nilai korelasi tertinggi dimiliki oleh *window size* 7 dan nilai korelasi terendah pada *window size* 11. Hal ini terjadi karena dengan pemilihan window size dapat menentukan jumlah probabilitas suatu kata berpasangan dengan kata yang lain. Sehingga nilai Word2vec dapat ditingkatkan berdasarkan penentuan jumlah window size.

2. Besar ukuran dari korpus data. Apabila ukuran korpus semakin besar, maka semakin banyak kosa kata yang dimiliki korpus tersebut, sehingga akan semakin baik hasil nilai kesamaan semantik yang dihasilkan oleh sistem.

3. Kesamaan semantik yang dihasilkan oleh sistem dipengaruhi oleh kosa kata yang dimiliki oleh korpus serta kemungkinan pasangan kata yang sering muncul bersama pada korpus tersebut.

5. Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, maka dapat ditarik kesimpulan sebagai berikut:

1. Berdasarkan skor kesamaan semantik yang dihasilkan dari perhitungan menggunakan Word2vec dapat disimpulkan bahwa Word2vec tidak terlalu baik dalam menghitung kesamaan semantik antar kata karena menghasilkan nilai kesamaan semantik yang jauh berbeda apabila dibandingkan dengan nilai dari *gold standard*. Nilai korelasi yang dihasilkan oleh sistem tergolong dalam keterhubungan lemah, dimana nilai korelasi tertinggi yaitu ditemukan pada pada korpus Brown dengan *window size* 7 yaitu sebesar 0.19235.

2. Faktor-faktor yang mempengaruhi nilai kesamaan semantik antar kata menggunakan Word2vec yaitu nilai Word2vec yang dihasilkan sepasang kata (dataset SimLex999) dipengaruhi oleh jumlah kemunculan kata pada korpus dan *window size* yang digunakan. Jika *window size* yang digunakan terlalu kecil maka konteks kata yang dihasilkan juga semakin sedikit sehingga dapat mempengaruhi nilai kesamaan semantik. Semakin besar *window size* yang digunakan maka semakin bayak konteks kata yang dihasilkan, sehingga kemungkinan kemunculan pasangan kata semakin besar. Faktor lain yang mempengaruhi nilai kesamaan semantik yaitu ukuran korpus, semakin besar ukuran korpus maka semakin banyak kosa kata yang dimiliki oleh korpus tersebut sehingga lebih banyak data yang dapat dilatih.

Saran

1. Melakukan pengujian dengan korpus data yang lebih besar dan kategori yang lebih luas untuk membuktikan seberapa efektif Word2vec dalam melakukan perhitungan kesamaan semantik antar kata.

2. Melakukan pengembangan pengukuran Word2vec terkait pemilihan *window size* yang lebih optimal.

Daftar Pustaka

[1] HANDLER, A. An empirical study of semantic similarity in wordnet and word2vec.

[2] HILL, F., REICHART, R., AND KORHONEN, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics 41*, 4 (2015), 665–695.

[3] JAKARTA, P. D. Berita Jakarta. `http://www.beritajakarta.id/`.

[4] JURAFSKY, D., AND MARTIN, J. H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2009.

[5] KENTER, T., AND DE RIJKE, M. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (2015), ACM, pp. 1411–1420.

[6] KUCERA, H., AND FRANCIS, W. N. Brown corpus manual. *Revised and Amplified Version* (1979).

[7] LICENCE, C. C. Pearson's correlation. `http://www.statstutor.ac.uk/`. [Online; accessed 17-Oktober-2017].

[8] MEYER, D. How exactly does word2vec work?

[9] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[10] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[11] PALMER, F. R. Semantics.

[12] PIRRÓ, G. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering 68*, 11 (2009), 1289–1308.

[13] RESNIK, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research 11* (1999), 95–130.

[14] ROWLING, J. Harry potter and the sorcerer's stone. new york: Arthur a, 1998.

[15] STANFORD. Unsupervised Feature Learning and Deep Learning Softmax Regression. [Online; accesed 23-April-2018].

[16] WORDNET. A Lexical Database for English. `https://wordnet.princeton.edu/`. [Online; accessed 5-Mei-2018].