Implementasi dan Analisis Kesamaan Semantik Antar Kata Berbahasa Inggris dengan Metode Second Order Co-occurrence Pointwise Mutual Information

Implementation and Analysis Semantic Similarity Between Words in English with the Method of Second Order Co-occurrence Pointwise Mutual Information

I Komang Resnawan Tri Putra[1], Ir. M. Arif Bijaksana, M.Tech.,Ph.D.[2], Mohamad Syahrul Mubarok[3]

Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

[1]resnaone.resnawan@gmail.com, [2]arifbijaksana@telkomuniversity.ac.id, [3]msyahrulmubarok@telkomuniversity.ac.id

Abstrak

Keterkaitan semantik mengacu pada sejauh mana dua konsep atau kata-kata yang terkait (atau tidak), sedangkan kesamaan semantik adalah kasus khusus atau bagian dari keterkaitan semantik. Kesamaan kata (*word similarity*) adalah pengukuran seberapa mirip sebuah pasangan kata secara semantik, dengan adanya hubungan sinonim maka pasangan kata tersebut memiliki nilai tertinggi. *Pointwise Mutual Information* (PMI) merupakan salah satu pengukuran secara statistik untuk keterkaitan semantik dan kesamaan semantik yang telah banyak digunakan. Salah satu varian pada PMI ialah *Second Order Co-ocurrence Pointwise Mutual Information* (SOC-PMI). Hasil dari penelitian pada tugas akhir ini merupakan nilai korelasi antara skor kesamaan yang dihasilkan sistem dengan *gold standard SimLex-999, WordSim353* dan *Miller and Charles*. Nilai korelasi tertinggi yaitu 0,2881 dengan menggunakan *window size* = 33 dan nilai δ = 6,5. Parameter yang menyebabkan korelasi terbaik dengan metode SOC-PMI ini ialah konteks katanya antara pasangan kata yang dibandingkan.

Kata kunci : Kesamaan semantik, Pointwise Mutual Information, Second Order Co-occurrence Pointwise Mutual Information.

Abstract

Semantic relatedness refers to the degree to which the two concepts or words associated (or not), while the semantic similarity is a special case or subset of a semantic relatedness. The similarity of words (word similarity) is a measurement of how closely a pair of words semantically, if a word pair have synonyms relationship then they has the highest value. Pointwise Mutual Information (PMI) is a statistical measurement of the semantic relatedness and semantic similarity that has been widely used. One variant of the PMI is Second Order Co-ocurrence pointwise Mutual Information (SOC-PMI). The results on this research is the correlation between similarity scores generated by the gold standard system SimLex-999, WordSim353 and Miller and Charles. The highest correlation value is 0.2881 by using window size = 33 and a value of δ = 6.5. The highest correlation value is 0.2881 by using window size = 33 and a value of δ = 6.5. The parameters that cause the best correlation with PMI-SOC method is term-context between words that compared.

***Keyword : Semantic Similarity, Pointwise Mutual Information,* Second Order Co-occurrence Pointwise Mutual Information.**

1. Pendahuluan

Keterkaitan semantik mengacu pada sejauh mana dua konsep atau kata-kata yang terkait (atau tidak), sedangkan kesamaan semantik adalah kasus khusus atau bagian dari keterkaitan semantik [1]. Kesamaan kata (*word similarity*) adalah pengukuran seberapa mirip sebuah pasangan kata secara semantik, dengan adanya hubungan sinonim pasangan kata tersebut akan memiliki nilai tertinggi [2]. Contoh kata "baju" dan "kain" dibandingkan dengan kata "baju" dan "celana", kata baju dan kain memiliki keterkaitan satu sama lain, sedangkan kata baju dan celana merupakan dua kata yang merujuk pada pakaian.

Kesamaan kata biasanya digunakan dalam Pemrosesan Bahasa Alami (PBA) atau sering disebut *Natural Language Processing* (NLP), pencarian informasi (*information retrieval*), dan kecerdasan buatan (*artificial intelligence*) [2]. Pendekatan yang berlaku pada komputasi *word similarity* ada dua, salah satunya ialah berbasis *thesaurus* atau statistik dari kumpulan data tulisan (*corpus*) yang besar. Pengukuran kesamaan kata yang telah banyak diketahui biasanya berdasarkan pada WordNet [3] dan sebagian besar aplikasi semantik mengandalkan taksonomi dari WordNet tersebut. Pengukuran dengan menggunakan WordNet ini secara khusus bergantung pada informasi dari definisi yang tersedia pada kata benda namun tidak lengkap untuk kata kerja dan sangat kurang

lengkap untuk kata sifat dan kata keterangan/tambahan [2]. Konsekuensinya, performanya tidak baik dan akurasi yang dihasilkan sangat kecil (tidak sampai 25%) ini terjadi dalam kasus menjawab pertanyaan sinonim TOEFL tujuannya ialah memilih kata mana yang merupakan sinonim pada empat kandidat pilihan terhadap kata yang diberikan. Beberapa pendekatan *corpus-based* mencapai akurasi yang lebih tinggi dalam kasus yang sama (diatas 80%) [2]. PMI (*Pointwise Mutual Information*) telah dimunculkan sebagai salah satu pengukuran statistik *word similarity* yang tidak berdasarkan pada hipotesis distribusional. Menghitung PMI hanya memerlukan statistik yang sederhana dari dua kata, *marginal frequencies* pasangan kata tersebut dan *co-occurrence frequency* dalam kumpulan data tulisan (*corpus*) [2].

Pada jurnal ini penulis mengimplementasikan pendekatan SOC-PMI (*Second Order Co-occurrence Pointwise Mutual Information*) dalam bentuk aplikasi untuk mengukur kesamaan kata antar kata dengan beberapa konteks katanya menggunakan dataset *Brown Corpus* dan skor yang dihasilkan akan dihitung nilai korelasinya dengan *Gold Standard* (*SimLex-999, Wordsim353 similarity* dan *Miller and Charles*) sehingga kedepannya dapat digunakan untuk analisis teks dan pencarian informasi.

2. Dasar Teori

2.1 PMI (Pointwise Mutual Information)

PMI (*Pointwise Mutual information*) adalah ukuran asosiasi yang digunakan dalam teori informasi dan statistik. Dalam komputasi linguistik, PMI untuk dua istilah yang diberikan menunjukkan kemungkinan dalam menemukan satu istilah dalam teks dokumen yang mengandung istilah lainnya [4]. Rumus penghitungan dengan PMI adalah sebagai berikut.

$$\text{PMI}(w1, w2) = \log \frac{P(t1, t2)}{P(t1) \times P(w2)} \quad (1)$$

$P(t_1,t_2)$ adalah probabilitas bahwa konsep t_1 dan t_2 yang terjadi bersamaan dalam dokumen yang sama, $P(t_1)$ dan $P(t_2)$ untuk t_1 dan t_2 masing - masing merupakan probabilitas kemunculan dalam sebuah dokumen [4]. Hasil pengukuran PMI menggambarkan bagaimana hubungan terminologi t_1 dan t_2, dengan mempertimbangkan frekuensi dari t_1 dan t_2 secara individu (*marginal probability*) dan frekuensi dari t_1 dan t_2 yang terjadi bersamaan (*joint probability*). Hal ini sangat memungkinkan t_1 dan t_2 terjadi bersamaan dalam sebuah konteks dengan cukup rendah namun t_1 dan t_2 dapat berhubungan dan deskriptif satu sama lain [5].

2.2 SOC-PMI (Second Order Co-occurrence Pointwise Mutual Information)

Metode ini juga terkait dengan literatur tentang *text mining* dan *data mining*, di dalamnya menyajjikan pendekatan metodis untuk mengekstraksi informasi relasional yang menarik dari korpus. *Second Order Co-occurrence* PMI dapat membantu sebagai alat untuk membantu dalam pembangunan otomatis dari sinonim sebuah kata. Pertama perlu dipilah daftar kata-kata yang signifikan berdasarkan pada nilai PMI untuk kata (x). jika terdapat n signifikan kata dalam daftar kata, maka terapkan SOC-PMI untuk setiap kemungkinan ppasangan kata pada pemetaan x ke n. Berdalih menentukan nilai kesamaan kata, pertimbangkan semua *second order co-occurrence* konteks dan urutkan konteks tersebut berdasarkan nilai PMI. Konteks kata yang memiliki nilai PMI tertingi bisa menjadi kandidat untuk sinonim kata.

Misalkan W_1 dan W_2 menjadi dua kata yang kita butuhkan untuk menentukan kesamaan semantik dan $C = \{c_1, c_2, \ldots, c_m\}$ menunjukkan korpus teks besar (setelah dilakukan *preprocessing* contohnya *stop words elimination* dan *lemmatization*) yang mengandung m kata (*token*). Misalkan juga $T = \{t_1, t_2, \ldots, t_m\}$ menjadi kumpulan semua kata unik (*types*) yang muncul pada korpus C. T berbeda dengan korpus C, yang merupakan kumpulan urutan yang mengandung banyak kemunculan kata yang sama, T merupakan kumpulan kata yang tidak diulang [1].

Sebuah riset SOC-PMI menetapkan parameter α, yang menentukan berapa banyak kata sebelum dan sesudah target kata W, yang akan dimasukkan dalam *context window*. *Context window* juga mengandung target kata W itu sendiri, sehingga ukuran *window* $2\alpha + 1$ kata. Langkah-langkah dalam menentukan kesamaan semantik melibatkan pemindaian korpus dan ekstraksi beberapa fungsi yang terkait ke jumlah frekuensi.

Berikut ini definisi fungsi *type frequency*,

$$f(w_i) = |\{c_j = w_i\}|, \text{ dimana } i = 1, 2, \ldots, n \quad (2)$$

yang menginformasikan seberapa banyak *type* w_i muncul pada seluruh korpus. Misalkan

$$f(w_i, c_j) = |\{c_j = \text{context}(w_i) = w_i\}|, \quad (3)$$

dimana $i = 1, 2, \ldots, n$ dan $-c \le i \le c$ menjadi fungsi *bigram frequency*. $f(w_i, w)$ yang menginformasikan seberapa sering kata w_i muncul dengan kata W dalam *window size* kata $2c + 1$.

Lalu berikut ini definisi fungsi *pointwise mutual information* hanya untuk kata yang memiliki $f(w_i, w) > 0$,

$$pmi(w_i, w) = \log_2 \frac{f^{\beta}(i, W) \times m}{f(w_i) f(w)} \quad (4)$$

dimana $f(w_i) f(w) > 0$ dan m merupakan jumlah *token* dalam korpus C, seperti yang telah dikatakan sebelumnya. Sekarang untuk kata w definisikan sebuah kumpulan kata, X, di sortir dari urutan yang tertinggi hingga yang terendah berdasarkan nilai PMI dengan w dan mengambil kata p tertinggi yang memiliki $pmi(w_i, w) > 0$.

$X = \{x_i\}$ dimana $i = 1, 2, \ldots, p$

dan $pmi(w_1, w) \ge pmi(w_2, w) \ge \cdots pmi(w_{p-1}, w) \ge pmi(w_p, w)$

Dengan cara yang sama, untuk kata w definisikan sebuah kumpulan kata, Y, di sortir dari urutan yang tertinggi hingga yang terendah berdasarkan nilai PMI dengan w dan mengambil kata q tertinggi yang memiliki $pmi(w_i, w) > 0$.

$Y = \{y_i\}$ dimana $i = 1, 2, \ldots, q$

dan $pmi(w_1, w) \ge pmi(w_2, w) \ge \cdots pmi(w_{q-1}, w) \ge pmi(w_q, w)$

Catatan bahwa belum ditentukannya nilai β (p atau q), nilai β sebenarnya bergantung pada kata W dan jumlah *type* dalam korpus.

Selanjutnya, definisikan fungsi penjumlahan β-PMI. Untuk kata w fungsi penjumlahan β-PMI adalah :

$$p\beta(w) = \sum_{i=1}^{p} (pmi(w_i, w))^{\beta} \quad (5)$$

dimana, $pmi(w_i, w) > 0$ dan $pmi(w_i, w) > 0$

yang menjumlahkan semua nilai PMI positif dari kumpulan kata-kata Y juga umum untuk kata-kata di kumpulan X. Dengan kata lain, fungsi ini sesungguhnya kumpulan nilai PMI positif dari semua kata yang dekat secara semantik dari w yang juga umum pada w. Ini dapat disebut sebagai *semantically-close* karena semua kata ini memiliki nilai PMI yang tinggi dengan w dan ini tidak menjamin kedekatan dengan patuh kepada jarak dalam *window size*.

Dengan cara yang sama, untuk kata w fungsi penjumlahan β-PMI adalah :

$$q\beta(w) = \sum_{i=1}^{q} (pmi(w_i, w))^{\beta} \quad (6)$$

dimana, $pmi(w_i, w) > 0$ dan $pmi(w_i, w) > 0$

yang menjumlahkan semua nilai PMI positif dari kumpulan kata-kata X juga umum untuk kata-kata di kumpulan Y. Dengan kata lain, fungsi ini sesungguhnya kumpulan nilai PMI positif dari semua kata yang dekat secara semantik dari w yang juga umum pada w. Disini belum didiskusikan kriteria untuk memilih parameter γ eksponensial [1].

Terakhir, kita definisikan fungsi *semantic* PMI *similarity* antara 2 kata, w_1 dan w_2

$$SimPMI(w_1, w_2) = \frac{f^{\beta}(W_1)}{p} + \frac{f^{\beta}(W_2)}{p} \quad (7)$$

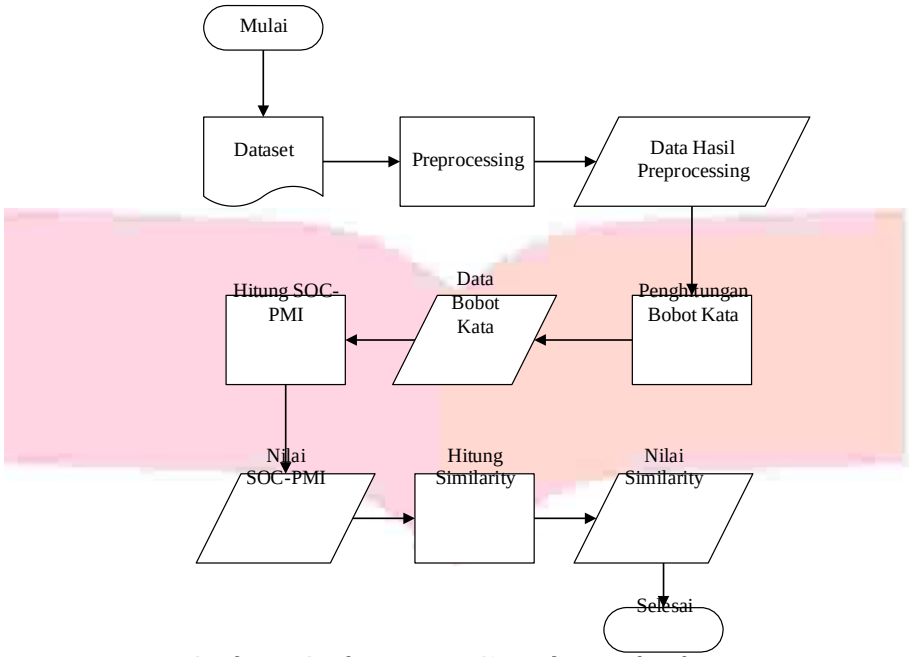
2.2.1. Memilih nilai β dan γ

Nilai dari β terkait kepada seberapa sering kata W muncul dalam korpus, dalam contoh, frekuensi W maupun jumlah *type* dalam korpus. Disni nilai β didefinisikan sebagai

$\beta_i = (\log(f(w)))^2 \frac{(\log_2(p))}{p}$, dimana i = 1, 2 (2.14)

Dimana δ adalah konstan dan dalam semua percobaan yang dilakukan oleh Md. Aminul Islam and Diana Inkpen dalam papernya yang berjudul SOC-PMI ini, digunakan nilai δ = 6.5. Nilai δ tergantung pada ukuran korpus yang digunakan. Semakin kecil korpus yang digunakan, semakin kecil nilai dari δ yang harus ditentukan. Jika nilai β diturunkan maka akan ada beberapa kata penting / menarik yang hilang, dan jika ditingkatkan maka anggapannya akan banyak kata umum diantara dan ini secara signifikan dapat merendahkan hasilnya. γ harus memiliki nilai lebih besar dari 1. Semakin besar nilai yang ditentukan untuk γ, maka semakin besar tekanan pada kata yang memiliki nilai PMI yang tinggi dengan W. Untuk semua eksperimen digunakan nilai γ = 3 [1].

Setelah mengetahui bagaimana proses penghitungan menggunakan SOC-PMI, berikut ini ialah gambaran umu sistem yang dibangun penulis (Gambar 1).



Gambar 1: Gambaran Umum Sistem dengan Flowchart

3. Pembahasan

3.1 Perbandingan Nilai Korelasi Berdasarkan Nilai Kesamaan (*similarity*) yang dihasilkan Sistem

Analisis perbandingan nilai ini dilakukan dengan membandingkan metode SOC-PMI dengan metode lainnya, pengujian perlu dilakukan dengan menggunakan tiga dataset *gold standard* yaitu Simlex-999, Wordsim-353 dan *Miller and Charles*. Penulis menggunakan sebuah *library* WS4J pada aplikasi java untuk mendapatkan skor similarity dari metode Path-based, Lin, dan Resnik sehingga dapat diukur nilai korelasinya dengan *gold standard* Simlex-999 dan Wordsim-353 menggunakan korelasi Pearson. WS4J pada java ini secara default mengambil POS-tag dan *sense* yang sering digunakan (*most frequent sense*) dari kata yang ingin kita ukur nilai kesamaannya. Nilai korelasi masing-masing metode ini dapat dilihat pada Tabel 1, Tabel 2 dan Tabel 3.

Tabel 1: Nilai Korelasi dengan Gold Standard Simlex-999

Nama Metode	Nilai Korelasi
Path	0,329
Lin	0,300
Resnik	0,186
SOC-PMI sistem	0,0032

Nilai korelasi tertinggi dengan menggunakan *gold standard* Simlex-999 dimiliki oleh metode Path-based, sedangkan SOC-PMI memiliki nilai terendah dari ketiga metode tersebut.

Tabel 2: Nilai Korelasi dengan Gold Standard Wordsim-353 Similarity

Nama Metode	Nilai Korelasi
Lin	0,469
Resnik	0,441
Path	0,413
SOC-PMI sistem	-0,0112

Berbeda dengan nilai korelasi menggunakan *gold standard* Wordsim-353, korelasi tertinggi dimiliki oleh metode Lin, SOC-PMI tetap memiliki nilai terendah dari ketiga metode tersebut.

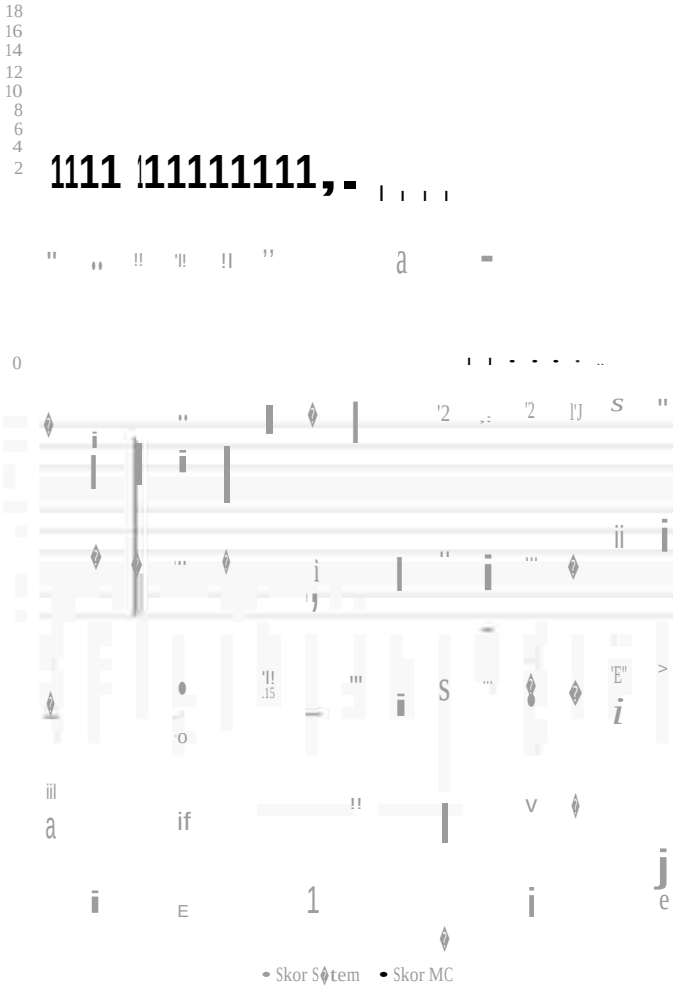
Tabel 3: Nilai Korelasi dengan Gold Standard Miller and Charles

Nama Metode	Nilai Korelasi
Lin	0,5308
Path	0,5101
Resnik	0,5063
SOC-PMI sistem	0,2281

Nilai korelasi tertinggi menggunakan *gold standard Miller and Charles* dihasilkan oleh metode Lin, SOC-PMI tetap memiliki nilai terendah dari ketiga metode tersebut.

3.2 Analisis Perbandiingan Nilai Korelasi Berdasarkan Nilai Kesamaan Semantik yang dihasilkan Sistem

Penulis melakukan analisis perbandingan nilai korelasi terbaik berdasarkan nilai kesamaan semantik yang dihasilkan sistem dengan menggunakan tiga *gold standard* yaitu Simlex-999, Wordsim-353 dan *Miller and Charles* dan menggunakan ukuran *window size* = 33 dan 41 serta nilai **δ** = 2,5 dan 6,5. Berdasarkan pengujian yang dilakukan, didapatkan korelasi terbaik menggunakan *gold standard Miller and Charles* dengan *window size* = 33 dan nilai **δ** = 6,5. Dibawah ini akan ditampilkan grafik hasil pengujian yang dilakukan Penulis.



Gambar 2: Grafik hasil pengujian MC window size = 33 dan δ =6,5

Gambar 2 diatas merupakan grafik nilai kesamaan yang dihasilkan sistem (garis abu-abu) dengan nilai kesamaan yang ada pada *gold standard Miller and Charles* (garis hitam) sangat jauh berbeda, korelasi diantara nilai yang dihasilkan sistem dengan *goldstandard Miller and Charles* ialah 0,2281. Trend skor pasangan kata "*coast*" dan "*shore*" yang dihasilkan sistem dengan *gold standard* serupa namun trend data yang lain tidak mengikuti dan rentang nilai kesamaan yang dimiliki *gold standard* MC ialah 0-4, sehingga kesamaan semantik = 0 yang dihasilkan sistem sebanyak 96% tidak jauh berbeda. Maka kekuatan korelasinya masuk pada kategori lemah. Tabel 4 dibawah ini menampung semua nilai korelasi dari pengujian yang telah dilakukan Penulis dengan menggunakan tiga dataset berbeda.

Tabel 4: Nilai Korelasi Sistem dengan Tiga Gold Standard

Dataset	Jumlah Pasangan Kata	Window Size	δ (Teta)	Nilai Korelasi

Wordsim-353	189	33	2,5	0,1947
	189	33	6,5	0,1464
	189	41	2,5	0,1880
	189	41	6,5	0,1438
Simlex-999	1000	33	2,5	-0,0020
	1000	33	6,5	-0,0227
	1000	41	2,5	-0,0061
	1000	41	6,5	-0,0209
Miller and Charles	30	33	2,5	-0,0480
	30	33	6,5	0,2281
	30	41	2,5	0,0065
	30	41	6,5	-0,0877

3.3 Analisis Parameter yang Mempengaruhi Nilai Kesamaan yang dihasilkan Sistem dan Nilai Korelasinya

Berdasarkan pengujian sebelumnya, untuk menganalisis parameter yang mempengaruhi nilai korelasi terbaik maka, lima sampel pasangan kata digunakan untuk membantu menemukan parameter tersebut karena *window size* dan nilai δ tidak terlalu signifikan menyebabkan korelasi berkriteria tinggi.

Lima sampel pasangan kata tersebut ada pada Tabel 4-8, disana ditampilkan nilai β, nilai fungsi penjumlahan β-PMI, nilai kesamaanya(skor sistem), nilai kesamaan pada *gold standard*, dan nilai PMI tertinggi untuk konteks kata 1/kata 2 sebanyak nilai β akan dilampirkan pada Lampiran 3 Konteks Kata untuk Analisis Parameter yang Mempengaruhi Nilai Kesamaan, hal ini berfungsi untuk membantu melakukan analisis. Sampel ini menggunakan *window size* = 11 dan δ = 2,5

Tabel 5: 5 Sampel Data Analisis

Kata 1	Kata 2	β kata 1	β kata 2	β-PMI 1	β-PMI 1	Skor sistem	Skor *gold standard*
tiger	cat	28,9	84,9	-	-	0	7,35
tiger	tiger	28,9	28,9	6781,7	6781,7	467,9	10
plane	car	156,1	214,7	118,8	112,9	1,28	5,77
train	car	144,8	214,7	686,4	536,6	7,23	6,31
television	radio	92,8	143,5	1900,3	1522,4	31,08	6,77

Menurut hipotesis yang dikemukakan oleh (Harris, 1954, Firth, 1957; Deerwester et al., 1990) bahwa setiap kata yang terdapat pada konteks yang sama seharusnya memiliki makna yang sama, hal tersebut dapat dijadikan patokan bahwa untuk sebuah pasangan kata yang memiliki makna yang sama nilai kemsaannya tinggi. Namun hal ini berbeda dengan nilai yang dihasilkan sistem pada Tabel 4-8, untuk pasangan kata "tiger" dan "cat" pada *gold standard* memiliki nilai 7,38 sedangkan sistem menghasilkan nilai kesamaan = 0, karena konteks kata yang diurutkan untuk pasangan kata tersebut tidak ada satupun yang sama. Untuk pasangan kata "*tiger*" dan "*tiger*" seluruh konteks yang diurutkan berdasarkan nilai PMI tertinggi untuk pasangan kata tersebut sama, sehingga nilai kesamaan yang dihasilkan sangat tinggi.

Window size yang digunakan jika diperbesar akan mempengaruhi jumlah konteks kata, semakin besar nilai *window size* maka konteks katanya semakin banyak, sehingga menambah kemungkinan adanya konteks kata yang sama antara pasangan kata yang dibandingkan. Namun pada metode SOC-PMI ini, jika konteks kata diantara pasangan kata tersebut setelah di urutkan berdasarkan nilai PMI tertingginya tidak ada yang sama satu sama lain maka nilai kesamaan katanya akan sama dengan 0.

Nilai δ akan mempengaruhi seberapa besar nilai β, semakin besar nilai δ yang digunakan maka semakin kecil nilai β dan juga kebalikannya. Berdasarkan pengujian yang dilakukan menggunakan tiga *gold standard* pada pengujian sebelumnya, nilai **δ** tidak mempunyai pengaruh besar terhadap nilai kesamaan yang dihasilkan.

Parameter yang menyebabkan nilai korelasi tertinggi juga dapat dilihat dari grafik yang dihasilkan pada percobaan sebelumnya (Gambar 4-14), skala nilai kesamaan yang ada pada *gold standard* MC ialah 0-4 dan nilai kesamaan = 0 yang dihasilkan sistem mencapai 96% sehingga perbedaan trend antara nilai kesamaan pada *gold standard* dan nilai kesamaan yang dihasilkan sistem tidak jauh berbeda.

Jadi, parameter yang sangat mempengaruhi nilai kesamaan yang dihasilkan sistem ialah konteks kata yang ada pada korpus, nilai kesamaan yang dihasilkan juga akan berpengaruh pada perubahan trend dalam grafik nilai korelasinya.

4. Kesimpulan

Berdasarkan pengujian dengan menggunakan metode SOC-PMI (*Second Order Co-occurrence Pointwise Mutual Information*) yang dilakukan pada bab sebelumnya maka dapat ditarik kesimpulan yaitu:

1. Penulis berhasil menerapkan SOC-PMI (Second Order Co-occurrence Pointwise Mutual Information) sebagai salah satu metode pengukuran kesamaan semantic.
2. Pengujian dengan menggunakan *gold standard Simlex-999* korelasi terbaik dihasilkan oleh metode *Path-based*, pengujian menggunakan *gold standard Wordsim-353 similarity* korelasi terbaik dihasilkan oleh metode Lin, dan dengan *gold standard Miller and Charles* korelasi terbaik dihasilkan oleh metode Lin.

3. Nilai korelasi terbaik yang dihasilkan sistem ialah 0,2281 dengan menggunakan *gold standard Miller and Charles*, *window size* =33 dan nilai $\delta = 6,5$.
4. Parameter yang sangat mempengaruhi nilai korelasi terbaik menggunakan metode SOC-PMI ini ialah konteks kata antara pasangan kata yang dibandingkan.

Daftar Pustaka

[1] M. A. Islam dan D. Inkpen, "Second Order Co-occurence PMI for Determining the Semantic Similarity of Words," *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1033-1038, 2006.

[2] H. Lushan, T. Finin, P. McNamee, A. Joshi dan Y. Yesha, "Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy," *IEEE Transaction on Knowledge and Data Engineering, Vol. 25, No.6, pp 1307-1321*, June 2013.

[3] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *Proc. 14th Int'l Joint Conf. Artificial Intelligence*, 1995.

[4] A. Pesaranghader, S. Muthaiyah dan A. Pesaranghader, "Improving Gloss Vector Semantic Relatedness Measure by Integrating Pointwise Mutual Information," *International Conference on Informatics and Creative Multimedia*, pp. 196-201, 2013.

[5] D. Jurafsky dan J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2009.