Implementasi dan Analisis Keterkaitan Semantik Antar Kata Menggunakan Pointwise Mutual Information$_{max}$ dengan Estimasi dari Kata Polisemi

Implementation and Analysis of Semantic Relatedness to Words Pair Using Pointwise Mutual Information$_{max}$ with Estimates of Word Polysemy

I Made Darma Yoga[1], Ir. M. Arif Bijaksana, M.Tech.,Ph.D.[2], Mohamad Syahrul Mubarok[3]

Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

[1]adedarmayoga77@gmail.com , [2]arifbijaksana@telkomuniversity.ac.id,
[3]msyahrulmubarok@telkomuniversity.ac.id

Abstrak

Natural Language Processing atau premrosesan bahasa alami merupakan sebuah disiplin ilmu yang khusus mengolah teks yang ditulis langsung oleh manusia yang bersifat tidak terstruktur. Pengukuran *semantic similarity* antar kata merupakan salah satu tugas penerapan dari *Natural Language Proccessing* yang intinya adalah mencari skor *semantic similarity* antar kata. Skor tersebut menunjukkan seberapa erat tingkat kesamaan antar dua kata. Salah satu metode untuk menghitung *semantic similarity* adalah PMI$_{max}$ (Pointwise Mutual Information$_{max}$). PMI$_{max}$ mengestimasi korelasi maksimum antara dua kata dan korelasi antara makna terdekat kedua kata tersebut karena sebuah kata seringkali memiliki banyak makna atau bisa disebut dengan kata Polisemi.

Pada tugas akhir ini, diimplementasikan penghitungan *semantic similarity* antar kata menggunakan PMI$_{max}$ dengan menggunakan estimasi dari kata polisemi. konteks kata bersumber dari dataset Brown Corpus dan dataset Gutenberg. Hasil dari keterkaitannya dibandingkan dengan dataset *Gold Standard* WordSim-353 *semantic relatedness*, *semantic similarity*, Miller Charles dan Simlex-999.

Hasil penelitian yang didapat terlihat bahwa dengan menggunakan PMI$_{max}$ didapatkan korelasi terbaik yaitu 66,5% dengan dataset *gold standard* WordSim-353 *semantic similarity* menggunakan korelasi Pearson dan dengan menggunakan nilai sense hasil analisis variabel p dan q. Nilai *semantic similarity* setiap pasang kata sangat dipengaruhi oleh nilai *Co-Occurence* sepasang kata teresebut, semakin tinggi nilai *Co-Occurence* suatu pasangan maka akan menghasilkan skor *semantic similarity* yang tinggi.

Kata kunci : Kesamaan semantik, Pointwise Mutual Information, kata polisemi.

Abstract

Natural Languange Processing is one of science disciplines which focusing to generate knowledge from human written text which is not structured. The measuring of Semantic Similarity to word pair is one of the task in Natural Language Processing that the main idea is to find a semantic similarity score to word pair. This score is representing how similar the word pair is. One of methods for measuring semantic similarity is Pointwise Mutual Information$_{max}$ (PMI$_{max}$). PMImax estimate the maximum correlation to word pair and the closest sense between those two words because a word often has multiple sense or can be called with Word Polysemy.

In this final project is implemented semantic similarity measure to word pair using PMI$_{max}$ with estimates of word polysemy. The context of word sourced from Brown Corpus and Gutenberg dataset. The result of the score compared to gold standard dataset WordSim-353, Miller Charles, and Simlex-999.

Research results obtained by using PMImax shows that the best correlation is 66,5% with WordSim-353 semantic similarity dataset using Pearson correlation and the value of sense on the analysis of variables p and q. The semantic similarity score for each word pair is depend on Co-Occurence value, high Co-Occurence value will produce high semantic similarity score.

Keyword : Semantic Similarity, Pointwise Mutual Information, Word Polysemy.

1. Pendahuluan

Pada kemajuan zaman seperti sekarang ini, informasi merupakan suatu hal yang menjadi sangat berharga keberadaannya. Banyak contoh sumber informasi yang bisa didapatkan seperti dari televisi, radio, surat kabar, internet dan lain-lain. Sekarang ini sudah banyak orang yang mulai mayoritas menggunakan internet sebagai sumber informasi bahkan internet dapat menggantikan fungsi sumber informasi lainnya. *Natural Language Processing* merupakan cabang ilmu yang mempelajari dan mencari informasi yang berharga dari teks yang dibuat langsung oleh manusia. Dalam *Natural Language Processing* terdapat *task* penting yaitu *Semantic Similarity*.

Semantic Similarity adalah sebuah pengukuran mengenai seberapa kemiripan antara sepasang kata secara semantik [1]. PMI merupakan salah satu metode yang digunakan dalam pengukuran dari *semantic similarity* antara

sepasang kata yang masuk kedalam distributional *similarity* yang artinya pengukuran *semantic similarity* berdasarkan statistik dari korpus besar. Seiring dengan mahalnya untuk memproduksi sebuah korpus yang lengkap dengan *sense-tagged*, banyak riset telah dilakukan mengenai PMI sehingga menghasilkan banyak variasi yang salah satunya adalah PMI_{max}. Berbeda dengan PMI yang pengukurannya mengasumsikan bahwa setiap kata hanya memiliki satu *sense*, PMI_{max} mengestimasi korelasi maksimum antara dua kata dan korelasi antara makna terdekat kedua kata tersebut karena setiap kata mungkin memiliki lebih dari satu makna atau disebut kata polisemi.

Pada jurnal ini, diimplementasikan perhitungan *semantic similarity* antar kata menggunakan PMI_{max} dengan estimasi dari kata polisemi. Konteks kata bersumber dari dataset Brown Corpus dan dataset Gutenberg. Untuk mengevaluasi sistem yang dibuat, maka nilai hasil keluaran sistem dibandingkan dengan dataset *gold standard* WordSim-353, Simlex-999, dan Miller-Charles menggunakan pengukuran statistik yaitu korelasi dan akan dicari nilai korelasi terbaik.

2. Dasar Teori

2.1 Pointwise Mutual Information

PMI adalah sebuah teknik perhitungan keterhubungan dari sebuah asosiasi yang digunakan dalam teori informasi dan statistik. Dalam perhitungan linguistik, PMI untuk dua istilah yang diberikan mengindikasikan kemungkinan untuk mencari satu istilah dalam teks dokumen yang mengandung istilah lainnya. Rumus umum dari PMI sendiri adalah :

$$PMI(c1,c2) = \log \frac{fd(c1,c2) \cdot N}{f(c1) \cdot f(c2)} \quad (1)$$

fd(c1,c2) adalah bahwa konsep c1 dan c2 terjadi secara bersamaan sedangkan fc1, fc2 adalah frekuensi kemunculan masing-masing, N merupakan total kata dalam korpus [1]. Meskipun banyak digunakan, PMI memiliki batasan umum yaitu pertama, PMI mungkin menghasilkan skor negatif atau positif yang mempersulit interpretasi dan tidak ada batasan nilai khusus. Kedua, PMI dikenal memberikan skor tinggi untuk pasangan kata yang kemunculan secara bersamaannya rendah [2].

2.2 Pointwise Mutual Information$_{max}$

PMI_{max} merupakan modifikasi dari PMI. PMI_{max} antara dua kata w_1 dan w_2 dinyatakan dengan rumus :

$$PMI_{max}(w_1,w_2) = \log \left(\frac{(fd(w1,w2) - \frac{e^k}{N}(fw1.fw2 - \frac{fw1}{yw1}.\frac{fw2}{yw2}))N}{\frac{fw1}{yw1}.\frac{fw2}{yw2}} \right) \quad (2)$$

Dimana $fd(w_1,w_2)$ adalah nilai frekuensi *co-occurrence* antara sebuah pasangan kata w_1 dan w_2, e^k nilai tetapan yaitu 30, N merupakan total jumlah kata pada korpus, fw_1 dan fw_2 adalah frekuensi kemunculan suatu kata w_1 dan w_2 dalam korpus, sedangkan yw_1 dan yw_2 adalah nilai makna(*sense*) dari kata w_1 dan w_2.

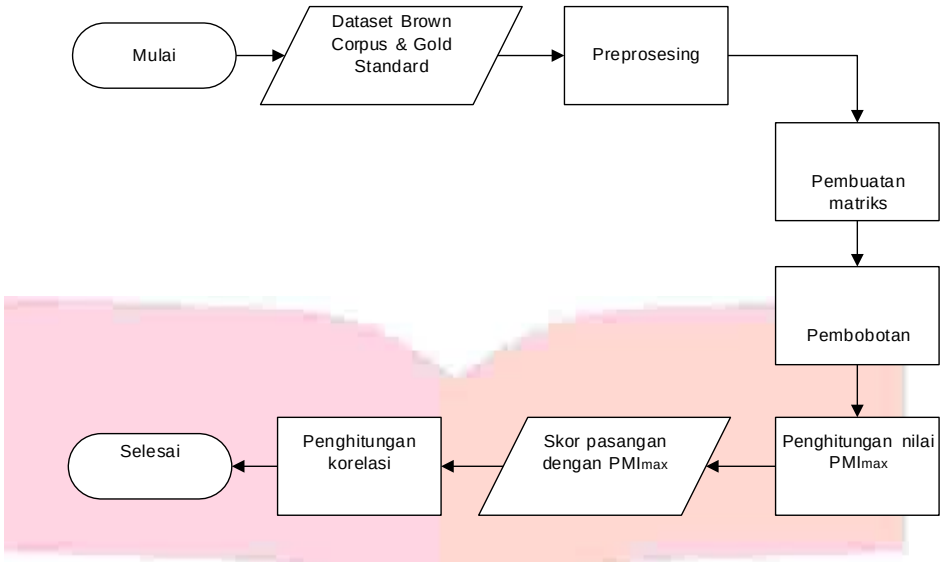
$$yw = \alpha (\log(fw) + q)^p \quad (3)$$

Pada persamaan 7 terdapat tiga variabel yang perlu diketahui nilainya, oleh karena itu rumus tersebut perlu dirubah menjadi

$$yw = \frac{(\log(fw) + q)}{(\log(700) + q)^p} \quad (4)$$

Dimana variabel q pada persamaan 2.16 nilainya antara *range* [-6 - 10] berkelipatan 1. Sedangkan variabel p memiliki *range* [0 - 10] berkelipatan 0,5

PMI_{max} mengestimasi korelasi maksimum antara dua kata. Korelasi antara makna terdekat mereka. Dalam keadaan kita tidak mengetahui makna umum dari *sense* yang digunakan, itulah alasan untuk mengambil nilai kesamaan maksimum antara semua kemungkinan pasangan makna sebagai penghitungan kesamaan kata **[1]**. Gambaran umum mengenai perancangan sistem dapat dilihat pada gambar 1

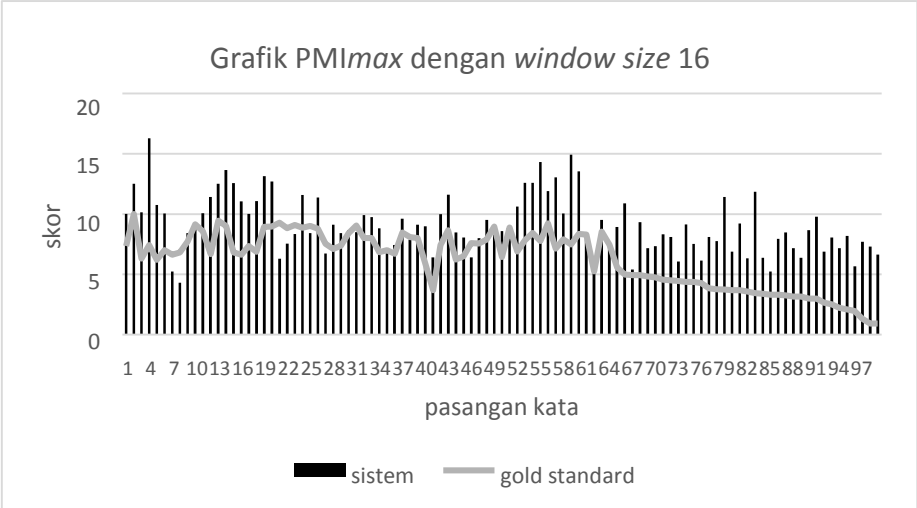


Gambar 1 Flowchart gambaran umum sistem

3. Pembahasan

3.1 Analisis Nilai *Semantic Similarity* Antar Kata PMI_{max} Berdasarkan Nilai Korelasi Terbaik

Pada analisis ini, akan dicari hubungan keterkaitan antara dua kata berdasarkan korelasi terbaik. Nilai hasil keluaran sistem terhadap seluruh pasangan kata pada dataset *gold standard* WordSim-353, Miller Charles dan Simlex-999 itu sendiri dengan menggunakan perhitungan korelasi Pearson dan korelasi Spearman. Pada pengujian ini dipisahkan ke dalam *window size* 16 dan *window size* 32. Nilai *sense* menggunakan dataset Wordnet.



Gambar 2 Grafik skor PMImax window size 16 dan skor dataset WordSim-353 semantic similarity

Dapat dilihat pada Gambar 2 yang memperlihatkan bagaimana korelasi antara sistem yang menggunakan *window size* 16 (garis berwarna hitam) dengan dataset *Gold Standard* WordSim-353 *semantic similarity* (garis berwarna abu-abu). Pada Gambar 4-1 tersebut terlihat bahwa sebagian besar pasangan kata sudah menghasilkan skor yang mendekati *gold standard*, namun ada kata yang skornya jauh dengan *gold standard* seperti pasangan kata "media-radio" yang menghasilkan skor tinggi yaitu 16,27 jika dibandingkan dengan *gold standard* yang nilainya 7,42. Nilai *Co-Occurence* untuk pasangan ini adalah 1 sedangkan frekuensi kemunculan kata "media" sebanyak 18 dan kata "radio" sebanyak 5. Dari frekuensi kemunculan kata "radio" yang hanya 5 kali, salah satunya beririsan dengan kata "media" sehingga memnghasilkan skor yang tinggi untuk pasangan kata tersebut. Jika dibandingkan dengan dataset *gold standard* dengan menggunakan korelasi maka menghasilkan korelasi sebesar 0,465 pada korelasi Pearson dan 0,477 pada korelasi Spearman.

Berikut hasil korelasi PMI_{max} dengan *window size 16* dan 32 pada dataset *gold standard* WordSim-353 *Semantic Relatedness, Semantic Similarity*, Simlex-999 dan Miller-Charles dalam bentuk tabel yang diperlihatkan pada Tabel 1 & 2.

Tabel 1 Hasil korelasi PMImax dengan window size 16

	Brown	Gutenberg
WS353 Semantic Relatedness(pearson)	0.353	0.234
WS353 Semantic Similarity(pearson)	0.359	0.465
SimLex-999(pearson)	0.029	-0.001
MC(pearson)	-0.161	0.472
WS353 Semantic Relatedness(spearman)	0.366	0.236
WS353 Semantic Similarity(spearman)	0.354	0.478
SimLex-999(spearman)	0.017	-0.029
MC(spearman)	-0.314	0.508

Tabel 2 Hasil korelasi PMImax dengan window size 32

	Brown	Gutenberg
WS353 Semantic Relatedness(pearson)	0.304	0.236
WS353 Semantic Similarity(pearson)	0.383	0.476
SimLex-999(pearson)	0.046	0.041
MC(pearson)	0.191	0.488
WS353 Semantic Relatedness(spearman)	0.312	0.265
WS353 Semantic Similarity(spearman)	0.365	0.494
SimLex-999(spearman)	0.032	0.016
MC(spearman)	0.035	0.546

3.2 Analisis Nilai Koefisien p dan q dalam Pencarian Sense PMImax Berdasarkan Korelasi Terbaik

Berdasarkan rumus pencarian *sense* dari suatu kata pada PMI_{max}, terdapat dua koefisien yang harus ditentukan nilainya. Pada analisis ini dicari nilai optimal dari koefisien p dan koefisien q, dimana koefisien q berada pada rentang [-6 - 10] berkelipatan 1 dan koefisien p berada pada rentang [0 - 10] berkelipatan 0,5, sehingga ditemukan nilai optimal untuk koefisien p dan koefisien q berdasarkan korelasi terbaik dengan dataset *gold standard* serta dicari nilai variabel yang berpengaruh terhadap skor *semantic similarity*.

Pasangan koefisien p dengan nilai 7,5 dan koefisien q dengan nilai 10 menghasilkan nilai korelasi terbaik yaitu 0,668 pada korelasi Pearson dan 0,628 pada korelasi Spearman.

4. Kesimpulan

Berdasarkan implementasi dan analisis pengujian yang dilakukan dapat ditarik kesimpulan sebagai berikut :

1. Sistem yang dibangun dapat mengimplementasikan perhitungan *semantic similarity* antar kata dengan metode PMI_{max} pada pasangan kata dataset *gold standard* WordSim-353, Miller-Charles, dan Simlex-999 dan memperoleh korelasi terbaik pada korelasi Pearson sebesar 0,665 dengan dataset *gold standard* WordSim-353 *semantic similarity*.
2. Parameter yang mempengaruhi nilai korelasi *semantic similarity* antar kata adalah dengan pencarian *sense* melalui analisis nilai koefisien p dengan nilai 7,5 dan q dengan nilai 10.
3. Skor *semantic similarity* antar kata sangat dipengaruhi oleh kemunculan kata tersebut pada korpus serta nilai *Co-Occurence* sepasang kata tersebut. Semakin tinggi *Co-Occurence* sepasang kata, maka akan semakin tinggi skornya.
4. Semakin tinggi ukuran *window size*, maka semakin tinggi peluang nilai *Co-Occurence* meningkat.

Daftar Pustaka

[1] L. Han, T. Finin, P. McNamee, A. Joshi dan Y. Yesha, "Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 1307-1319, 2013.

[2] F. Role dan M. Nadif, "Handling The Impact Of Low Frequency Events On Co-Occurence Based Measures Of Word Similarity (A Case Study of Pontwise Mutual Information)," pp. 1-4, 2011.

[3] P. D. Jurafsky dan J. H. Martin, Speech and Language Processing : An Introduction to natural language processing, computational linguistics, and speech recognition, 2006.