ANALISIS DAN IMPLEMENTASI *CROSS-LINGUAL SEMANTIC SIMILARITY* ANTAR KATA DENGAN METODE POINTWISE MUTUAL INFORMATION

ANALYSIS AND IMPLEMENTATION OF CROSS-LINGUAL SEMANTIC SIMILARITY WORDS USING POINTWISE MUTUAL INFORMATION METHOD

Sri Reski Anita Muhsini
Universitas Telkom, Bandung, Indonesia
srireskianita@gmail.com

ABSTRAK

Implementasi pengukuran kesamaan semantik memiliki peran yang sangat penting dalam beberapa bidang Natural Language Processing (NLP), dimana hasilnya seringkali dijadikan dasar dalam melakukan task NLP yang lebih lanjut. Salah satu penerapannya yaitu dengan melakukan pengukuran kesamaan semantik multibahasa antar kata. Pengukuran ini dilatarbelakangi oleh suatu masalah dimana saat ini banyak sistem pencarian informasi yang harus berurusan dengan teks atau dokumen multibahasa. Sepasang kata dinyatakan memiliki kesamaan semantik jika pasangan kata tersebut memiliki kesamaan dari sisi makna atau konsep. Pada penelitian ini, diimplementasikan perhitungan kesamaan semantik antar kata pada bahasa yang berbeda yaitu bahasa Inggris dan bahasa Spanyol. Korpus yang digunakan pada penelitian ini yakni Europarl Parallel Corpus pada bahasa Inggris dan bahasa Spanyol. Konteks kata bersumber dari Swadesh list, serta hasil dari kesamaan semantiknya dibandingkan dengan *datasetGold Standard* SemEval 2017 *Crosslingual Semantic Similarity* untuk diukur nilai korelasinya. Hasil pengujian yang didapat terlihat bahwa pengukuran metode PMI mampu menghasilkan korelasi sebesar 0,5781 untuk korelasi Pearson dan 0.5762 untuk korelasi Spearman. Dari hasil penelitian dapat disimpulkan bahwa Implementasi pengukuran Crosslingual Semantic Similarity menggunakan metode Pointwise Mutual Information (PMI) mampu menghasilkan korelasi terbaik. Peneliti merekomendasikan pada penelitian selanjutnya dapat dilakukan dengan menggunakan dataset lain untuk membuktikan seberapa efektif metode pengukuran Poitnwise Mutual Information (PMI) dalam mengukur Crosslingual Semantic Similarity antar kata.

Kata kunci:Kesamaan Semantik*, Crosslingual Semantic Similarity, Pointwise Mutual Information*

ABSTRACT

The implementation of semantic equality measurement has a very important role in some areas of NLP, where the results are often used as the basis for performing further NLP tasks. One of its application is by doing the measurement of multilingual semantic similaritiesbetween words. This measurement is motivated by a problem where many information search systems now have to deal with text or multilingual documents. A pair of words are said to have a semantic similarity if the word that pair has a similarity meaning or concept. In this study, the calculation of semantic similarity is implemented between words in different languages namely English and Spanish. The corpus used in this study is Europarl Parallel Corpus in English and Spanish. The word context is sourced from the Swadesh list, as well as the results of its semantic similarities compared to the Gold Standard SemEval 2017 Crosslingual Semantic Similarity dataset for measured the correlation values. The result shows that the measurement of PMI method yields a correlation 0.5781 for Pearson correlation and 0.5762 for Spearman correlation. From the result of the research, it can be concluded that Implementation of Crosslingual Semantic Similarity measurement using Pointwise Mutual Information (PMI) method can produce the best correlation. The researchers recommend that further research can be use other datasets to demonstrate how effective the method of measuring Poitnwise Mutual Information (PMI) in measuring crosslingual semantic similarity between words.

Keywords*:semantic similarity, crosslingual semantic similarity, pointwise mutual information*

ISSN 1412-565 X
e-ISSN 2541-4135

PENDAHULUAN

Kesamaan Semantik merupakan suatu pengukuran yang menghasilkan nilai dimana menyatakan tingkat kesamaan atau kedekatan secara semantik antarkata, kalimat, atau dokumen. Sepasang teks dinyatakan memiliki kesamaan semantik jika pasangan teks tersebut memiliki kesamaan dari sisi makna atau konsep (Palmer, 1976). Pengukuran kesamaan semantik dapat dilakukan terhadap pasangan teks dalam bahasa yang sama, maupun terhadap pasangan teks dalam bahasa yang berbeda (*Cross-lingual*). Kesamaan Semantik lintas bahasa (*Cross-lingual*) adalah tingkat kesamaan atau kedekatan antar teks dalam dua bahasa dari sisi makna atau konsep. Implementasi dari pengukuran kesamaan semantik sudah sejak lama diterapkan pada aplikasi Natural Languange Processing dan beberapa bidang terkait seperti data mining, information retrival, machine translation, dan lain sebagainya.

Permasalahan dalam bidang information retrieval adalah untuk menemukan dokumen yang relevan dalam kumpulan dokumen berdasarkan pada beberapa kata kunci yang menggambarkan kebutuhan informasi atau contoh dokumen yang relevan. Memperkirakan kesamaan semantik tepatnya antara kata sangat penting untuk menilai jika dokumen relevan dengan kebutuhan pengguna. Banyak sistem pencarian informasi, seperti sistem katalog perpustakaan online, serta mesin pencariweb, semua harus berurusan dengan dokumen multibahasa danmengukur kesamaan semantik antar kata dari berbagai bahasa (Wu & Chen, 2015). Oleh sebab itu, mengukur kesamaan semantik memiliki peranan yang penting, yang seringkali dijadikan sebagai dasar dalam melakukan tugas-tugas pemrosesan bahasa alami yang lebih lanjut atau lebih kompleks.

Pada penelitian ini akan diimplementasikan metode Pointwise Mutual Information (PMI) dalam mengukur kesamaan semantik untuk mengetahui keakuratan kesamaan antar kata dalam pasangan bahasa yang terdapat pada *dataset* SemEval 2017 *task cross-lingual*. Pengukuran PMI diobservasi menggunakan pasangan kata bahasa Inggris dan Spanyol pada *dataset* SemEval 2017 *task cross-lingual*. Korpus yang digunakan yaitu Europarl Parallel Corpus (Koehn, 2005) pada bahasa Inggris dan bahasa Spanyol, serta Konteks kata berasal dari *dataset* Swadesh list (Swadesh, 1950). Skor yang dihasilkan akan dihitung nilai korelasinya dengan *Gold Standard* sehingga kedepannya diharapkan dapat menghasilkan performansi yang lebih baik dibandingkan pengukuran menggunakan metode yang lain serta sistem yang dibangun dapat digunakan untuk analisis teks dan pencarian informasi.

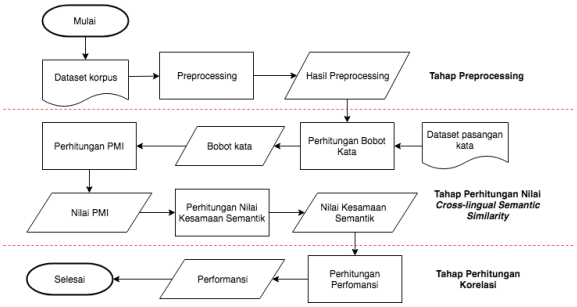
HASIL DAN PEMBAHASAN

1. Perancangan Sistem

Sistem yang dibangun dapat menghitung *cross-lingual semantic similarity* antar kata menggunakan Pointwise Mutual Information (PMI) seperti pada gambar 1. Pointwise Mutual Information (PMI) adalah metode teoritis informasi berbasis korpus sederhana untuk menemukan hubungan antara pasangan kata dengan menggunakan hipotesis distribusi, yangmenyatakan bahwa hubungan antara kata-kata bergantung pada

ISSN 1412-565 X
e-ISSN 2541-4135

Co-Occurence kata-kata dalam korpus (Jurafsky, 2000).



Gambar 1
Gambaran Umum Sistem

Dalam linguistik, dua istilah diukur mengunakan PMI untuk menunjukan kemungkinan menemukan satu istilah dalam dokumen yang mengandung istilah lainnya. untuk menghitung nilai PMI antara dua kata dapat dilihat pada persamaan berikut (Jurafsky, 2000).

$$PMI(a,b) = log_2 \frac{P_{(a \wedge b)}}{P(a).P(b)} \quad (1)$$

Dengan, P adalah kemunculan a dan b secara bersamaan di dokumen (korpus) yang sama, sedangkan P(a) dan P(b) adalah probabilitas kemunculan masing-masing di dokumen (korpus). Jadi PMI adalah log dari rasio frekuensi co-occurrence ke frekuensi kemunculan kata secara individu.

Persamaan joint probability konteks terhadap kata:

$$P_{(a \wedge b)} = \frac{f_{(a \wedge b)}}{N}$$

Persamaan marginal probability kata:

$$P(a) = \frac{f_{(a)}}{N}$$

$$P(b) = \frac{f_{(b)}}{N}$$

Dengan:

Dengan:

f_((a□b))= frekuensi bobot kata 1 dan kata 2

f_((a))= frekuensi bobot kata 1

f_((b))= frekuensi bobot kata 2

N = jumlah frekuensi seluruh kata terhadap konteks yang ada

Adapun *dataset* yang menjadi korpus pada sistem ini adalah Europarl Parallel Corpus. Sistem akan membaca masukkan berupa *dataset* SemEval 2017 dan *dataset* konteks kata Swadesh list sebagai pasangan kata lalu melakukan perhitungan nilai PMI pada pasangan kata tersebut. Perhitungan PMI dilakukan pada masing-masing *dataset* tiap bahasa.

Selanjutnya dilakukan perhitungan nilai *cross-lingual semantic similarity* dengan skala 0-1 dengan metode cosine similarity. Metode cosine similarity adalah metode untuk menghitung kesamaan daridua teks. Penentuan kesesuaian antar teks dipandang sebagai pengukuran (*similarity measure*) antara vektor teks pertama (A) dengan vektor teks kedua(B). Semakin sama suatuvektor teks pertama dengan vektor teks kedua makateks pertama dapat dipandang semakin sesuai dengan teks kedua (Swadesh, 1950). Persamaan digunakan untuk menghitung cosine similarity dengan tujuan untuk mengetahui angka similarity antar kata. Persamaan yang dimaksud sebagai berikut:

$$Sim = cos(\theta) = \frac{A.B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (5)$$

Dengan, A adalah vektor 1 dan B adalah vektor 2

Dalam penelitian ini, hasil korelasi akan menjadi keluaran sistem mengetahui performansi sistem yang dibangun, dimana nilai korelasi akan dihitung dengan membandingkan nilai *cross-lingual semantic similarity* yang dihasilkan oleh sistem dengan

ISSN 1412-565 X
e-ISSN 2541-4135

acuan *dataset* SemEval 2017 *task cross-lingual* sebagai *gold standard* menggunakan korelasi pearson dan korelasi spearman. Korelasi Spearman adalah pengukuran nilai korelasi yang berdasarkan tingkat (rank) keterhubungan antar dua variabel (StatisticsSolutions, 2017). Formula yang digunakan dalam mengukur korelasi dengan Korelasi Spearman yaitu pada persamaan (6)

$$\rho = 1 - \frac{6\sum_{i-1}^{n} d_i^2}{n(n^2-1)}$$

Dengan:

ρ = Spearman Correlation

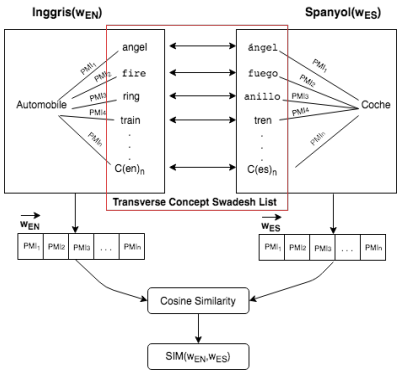
d_i= perbedaan di antara rank x dan y

n = jumlah data

Sedangkan Korelasi Pearson merupakan salah satu ukuran korelasi yang digunakan untuk mengukur hubungan linier atau antara dua variabel yang didefinisi kan sebagai covarians dari variabel dibagi dengan standar deviasinya. Dua variabel dapat dikatakan berkorelasi apabila perubahan salah satu variabel disertai dengan perubahan variabel lainnya. Koefisien Korelasi Pearson selalu berada di skala -1hingga 1. Formula yang digunakan untuk mengukur keterkaitan dua variabel menggunakan korelasi pearson yaitu pada persamaan berikut:

$$Pearson(x,y)\ \frac{N\sum xy-(\sum x)(\sum y)}{\sqrt{N\sum x^2-(\sum x)^2 N\sum y^2-(} \quad (7)$$

Dengan, N adalah jumlah pasangan kata, x adalah nilai dari sistem dan y adalah nilai dari *gold standard.*



Gambar 2
Arsitektur *Cross-lingualSemantic Similarity*

Penjelasan proses pada skema pembentukan model dan pengujian yang terdapat pada gambar 2 yaitu;

- Sistem membaca *dataset* korpus yaitu Europarl Parallel Corpus dan pasangan kata yang terdapat pada *dataset* SemEval 2017 dan *dataset* Swadesh list. Sistem akan membaca tiap baris pada file input yang berisi daftar pasangan kata yang ingin dibandingkan, jika pada baris tersebut terdapat kata yang tidak ada pada korpus maka sistem membaca baris selanjutnya. Sepasang kata yang menjadi inputan pada penelitian ini ialah kata yang juga tersedia pada *gold standard*.
- Sistem akan menghitung nilai marginal frequency dan Co-Occurrence frequency. Perhitungan bobot kata dilihat dari window size, penulis menetapkan nilai window size sebesar 15, 21 dan 31, apabila kata yang dimaksud ditemukan pada window size tersebut maka frekuensi kata ditambah 1.
- Sistem akan menghitung jumlah frekuensi seluruh kata terhadap konteks yang ada.
- Sistem menghitung nilai PMI setiap kata pada dataset SemEval dengan setiap kata yang terdapat pada Swadesh list dalam masing-masing bahasa (Inggris dan

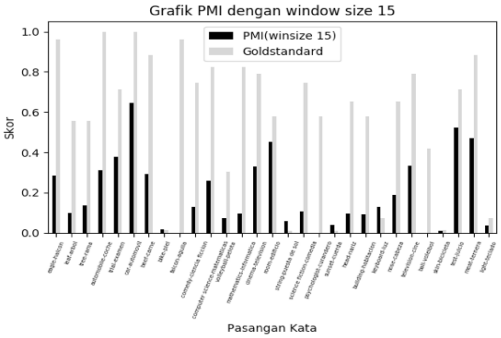
ISSN 1412-565 X
e-ISSN 2541-4135

Spanyol). Jumlah kata pada Swadesh list yang digunakan pada penelitian ini ialah sebanyak 100 kata dalam dua bahasa (Inggris dan Spanyol). Nilai PMI yang diperoleh pada masing-masing bahasa akan disimpan dalam file. Sehingga dihasilkan dua file yang berisi nilai PMI untuk bahasa Inggris dan nilai PMIuntuk bahasa Spanyol. Untuk lebih jelasnya dapat dilihat pada gambar 2.

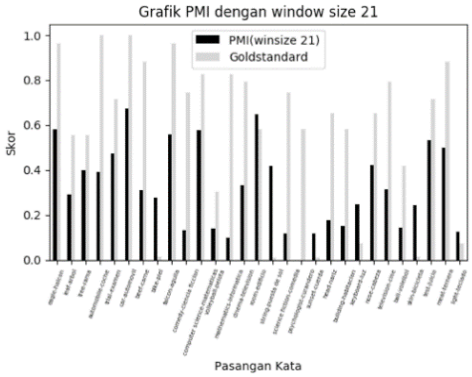
- Sistem akan membaca nilai PMI sebagai sebuah vektor. Kemudian nilai kesamaan semantik pasangan kata dua bahasa (Inggris dan Spanyol) diperoleh melalui perhitungan cosine similarity antar kedua vektor tersebut.

2. Analisis Pengujian Skenario 1

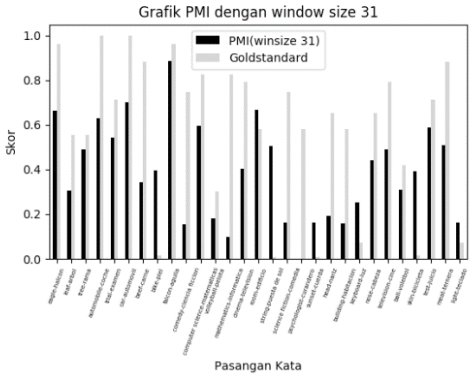
Pengujian pertama yang dilakukan yaitu menganalisis hubungan kesamaan semantik antar dua kata berdasarkan ukuran *window size*. Nilai hasil keluaran sistem terhadap seluruh pasangan kata pada *dataset* SemEval 2017 akan dibandingkan dengan nilai dari *gold standard* dari data itu sendiri dengan menggunakan perhitungan korelasi Pearson dan korelasi Spearman. Dalam pengujian ini dipisahkan ke dalam *window size* 15, *window size*, 21 dan *window size* 31. Hasil pengujian dapat dilihat pada gambar 3, gambar 4, dan gambar 5.



Gambar 3
PMI dengan *window size* 15



Gambar 4
PMI dengan *window size* 21



Gambar 5
PMI dengan *window size* 31

Dari hasil pengujian tersebut juga dapat disimpulkan *window size* mempengaruhi nilai kesamaan semantik yang diperoleh. Meningkatnya ukuran *window size* yang digunakan menyebabkan frekuensi kemunculan setiap pasangan kata pada korpus juga meningkat, sehingga kata tersebut memiliki nilai PMI, dan akan meningkatkan peluang konteks kata Swadesh list saling beririsan antar bahasa. Hal tersebut akan menyebabkan nilai kesamaan semantik yang dihasilkan juga meningkat.

3. Analisis Pengujian Skenario 2

Pengujian dilakukan untuk mengetahui perbandingan nilai korelasi yang dihasilkan sistem (PMI) dengan nilai korelasi yang dihasilkan menggunakan metode pengukuran kesamaan semantik Resnik, Path, dan Jiang

ISSN 1412-565 X
e-ISSN 2541-4135

& Conrath (1995). Hasil pengujian dapat dilihat pada tabel 1.

Tabel 1
Hasil perbandingan korelasi *Cross-lingual Semantic Similarity*

Nama Metode	Nilai Korelasi Pearson	Nilai Korelasi Spearman
PMI (*window size* 15)	0.5781	0.5762
Jiang & Conrath	0.4522	0.5386
Resnik	0.4653	0.4607
Path	0.4476	0.4543
PMI (*window size* 21)	0.4326	0.4447
PMI (*window size* 31)	0.4099	0.4432

Pada tabel 1, hasil korelasi PMI memiliki nilai yang tertinggi pada penggunaan *window size* 15 dan nilai korelasi terendah pada *window size* 31 dari ketiga metode pembanding yang digunakan. Hal tersebut disebabkan karena adanya faktor tambahan yang mempengaruhi nilai PMI yang tidak terdapat pada metode lain yaitu pada pemilihan *window size* yang digunakan, sehingga hasil nilai PMI bisa ditingkatkan berdasarkan penentuan *window size*.

SIMPULAN

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, maka dapat ditarik kesimpulan sebagai berikut; (1) Implementasi pengukuran *Crosslingual Semantic Similarity* menggunakan metode Pointwise Mutual Information (PMI) mampu menghasilkan korelasi terbaik diantara ketiga metode pembanding yang digunakan yaitu sebesar 0.5781 pada korelasi Pearson dan 0.5765 pada korelasi Spearman dengan menggunakan ukuran *window size* 15; (2) Faktor-faktor yang mempengaruhi nilai *Cross-lingual semantic similarity* antar kata menggunakan metode *Pointwise Mutual Information* (PMI) yaitu nilai PMI yang dihasilkan sepasang kata (*dataset* SemEval dan konteks kata Swadesh List), dan banyaknya konteks kata Swadesh list yang saling beririsan antar bahasa. Nilai PMI yang dihasilkan dipengaruhi oleh kemunculan pasangan kata (*dataset* SemEval dan kontek kata Swadesh List) tersebut pada korpus, serta ukuran *window size* yang digunakan. *Window size* yang digunakan jika diperbesar akan mempengaruhi jumlah konteks kata Swadesh list, semakin besar nilai *window size* maka konteks katanya semakin banyak,sehingga menambah kemungkinan adanya konteks kata yang sama antara pasangan kata dalam dua bahasa yang dibandingkan.

DAFTAR RUJUKAN

Church, K.W.,& Hanks, P. (1990). *Word association norms, mutual information, and lexicography*.Computational linguistics.

Jiang, J. J., & Conrath, D. W. (1997). *Semantic similarity based oncorpus statistics and lexical taxonomy*. ArXiv preprint cmp-lg/9709008

Jurafsky, D. (2000). *Speech and language processing: an introductionto natural language processing*.Computational Linguistics, and SpeechRecognition.

Koehn, P. 2005. *Europarl: A parallel corpus for statistical machine transla-tion*. InMT summit, vol. 5, pp. 79–86.

ISSN 1412-565 X
e-ISSN 2541-4135

Palmer, F. R. (1976). Semantic.*InSemantics*, pp. 5–7.

Resnik, P. (1995). *Using information content to evaluate semantic similarity ina taxonomy*. ArXiv preprint cmp-lg/9511007(1995).

SemEval2017task2. (2017). [Online]. Di akses dari http://alt.qcri.org/semeval2017/task2/index.php?id=task-details.

StatisticsSolutions. (2017). [Online]. http://www.statisticssolutions.com/spearman-rank-correlation/.

Swadesh, M. (1950). Salish internal relationships.*International Journal of American Linguistics* 16, 4, 157–167.

Swadesh, M. (1952). *Lexico-statistic dating of prehistoric ethnic contacts: withspecial reference to north american indians and eskimos*. Proceedings ofthe American philosophical society 96, 4, 452–463.

Wu, Y., Wu, S., & Chen, D. (2015). Chinese-english bilingual word seman-tic similarity based on chinese wordnet.*JSW*, 10, 1, 20–31.

Zhang, Z., Gentile, A. L., & Ciravegna, F. (2013). Recent advancesin methods of lexical semantic relatedness–a survey.*Natural Language Engineering,* 19, 04, 411–479.