PENGGUNAAN KESAMAAN SEMANTIK PADA PENENTUAN CALON SINONIM VERBA BAHASA INDONESIA

Shiva Dwi Samara Tungga[1] dan Totok Suhardijanto[2]
Universitas Indonesia
shivadsamara@gmail.com; suhardiyanto@gmail.com

ABSTRAK

Ketika kita ingin mengganti suatu kata dengan kata lain dengan makna yang sama, maka sinonim dari kata yang akan diganti itulah yang akhirnya dipilih dan digunakan. Penggunaan sinonim tersebut menjadikan pilihan kata yang kita gunakan menjadi lebih variatif dan tidak monoton. Berkaitan dengan hal tersebut, daftar sinonim yang telah ada seperti tesaurus menjadi salah satu pilihan untuk menemukan sinonim. Namun, seiring dengan semakin berkembangnya ilmu pengetahuan dan teknologi yang dapat diimplementasikan dalam ranah penelitian bahasa, memberikan pandangan serta cara baru dalam menentukan apakah suatu kata memiliki makna yang dekat dengan kata lain. Cara tersebut biasa disebut dengan kesamaan semantik (semantic similarity). Istilah kesamaan semantik tersebut merujuk pada pengukuran kesamaan dengan menggunakan perhitungan matematis. Apa yang dihitung dalam kesamaan semantik ini berdasarkan prinsip co-occurence, sehingga dalam melakukan perhitungan ini membutuhkan korpus bahasa yang selanjutnya diubah ke dalam bentuk vektor. Vektor-vektor yang terbentuk dari kata-kata di dalam korpus itulah yang selanjutnya diukur kedekatannya. Dengan adanya kesamaan semantik tersebut, maka penelitian ini membahas tentang bagaimana penggunaan kesamaan semantik dalam menentukan calon sinonim. Hal tersebut bertujuan untuk memberi gambaran bagaimana suatu pengukuran kesamaan semantik dapat diimplementasikan dalam menentukan calon sinonim. Penelitian ini menggunakan kesamaan semantik dengan metode berbasis korpus yang memanfaatkan vektor semantik serta pengukuran kesamaan menggunakan kosinus. Melakukan pengukuran untuk mendapatkan pasangan kata bersinonim dalam suatu korpus akan sulit jika dilakukan secara manual. Oleh sebab itu, perlu adanya cara lain untuk memudahkan penentuan calon-calon sinonim, seperti dengan bantuan pemrograman komputer. Mulai dari mengolah kata menjadi vektor hingga melakukan pengukuran kesamaan semantik pada penelitian ini menggunakan bahasa pemrograman Python yang telah dipasang kepustakaan kode (library) Gensim dan Word2vec. Selain itu, data berupa kata yang digunakan dalam penelitian ini untuk dilihat calon sinonimnya adalah verba bahasa Indonesia. Dengan demikian, hasil dari penelitian ini berupa calon-calon sinonim dari verba bahasa Indonesia yang dapat dijadikan sebagai penelitian awal dalam menentukan sinonim verba bahasa Indonesia.
Kata kunci: *Kesamaan Semantik, Pengukuran Semantis, Kesamaan Kosinus, Sinonim Verba*

PENDAHULUAN

Salah satu pengukuran semantik yang biasa dilakukan untuk melihat bagaimana hubungan kedekatan antar kata, kalimat, atau teks besar adalah kesamaan semantik (*semantic similarity*). Pada awalnya, kesamaan semantik merujuk pada istilah yang digunakan untuk mengatakan *similarity in meaning* (Miller & Charles, 1991). Namun, Miller dan Charles melakukan pertimbangan untuk menguji apakah suatu kemiripan semantik itu berkorelasi dengan konteks seperti yang diungkapkan oleh Rubenstein & Goodenough (1965). Istilah serta penggunaan kesamaan semantik ikut berkembang seperti apa yang dilakukan oleh Resnik (1995) dalam mengukur kesamaan semantik. Oleh sebab itu, dapat dikatakan bahwa kesamaan semantik merupakan suatu cara yang digunakan untuk mengukur derajat kesamaan dari dua objek (Lin, 1998). Selanjutnya, secara praktis kesamaan semantik digunakan sebagai istilah untuk merujuk pada pengukuran semantik yang biasanya muncul dalam Pemrosesan Bahasa Alami dengan menggunakan perhitungan matematis.

Berbagai macam pendekatan dapat diimplementasikan dalam kesamaan semantik, salah satunya yang sering digunakan adalah dengan memanfaatkan vektor semantik. Vektor semantik ini kemudian dipadukan dengan pengukuran nilai kosinus atau yang biasa disebut dengan pengukuran terhadap sudut kata-kata yang telah direpresentasikan ke dalam vektor. Seperti apa yang telah dilakukan oleh Huerta (2008), pendekatan vektor digunakan untuk melakukan pengukuran kesamaan semantik pada tingkat kalimat. Selain Huerta, penelitian yang dilakukan oleh Rani, dkk. (2017) dan Hermawan (2017) juga memanfaatkan penggunaan vektor untuk mengukur kesamaan semantik.

Penelitian lain terkait kesamaan semantik juga dilakukan oleh Islam dan Inkpen (2008), Rahutomo dkk. (2012), Slimani (2013), Guessoum dkk. (2016), Sravanthi dan Srinivasu (2017), Ali dkk (2018), serta Pawar dan Mago (2018). Apa yang telah mereka lakukan berkaitan dengan penggunaan kesamaan semantik, evaluasi, maupun bagaimana kesamaan semantik bekerja dalam tingkat kata maupun kalimat.

Selain itu, seperti apa yang telah disebutkan sebelumnya, penerapan penggunaan kesamaan semantik pada bahasa Indonesia jarang sekali dilakukan. Sebut saja salah satunya yang telah dilakukan oleh Hermawan

dkk. (2017). Akan tetapi, penggunaannya masih berfokus pada nilai statistik serta matematis dalam menunjukkan sejauh mana kesamaan semantik dapat diimplementasikan pada bahasa Indonesia, sehingga belum ada implementasi lanjutan yang dapat secara langsung dimanfaatkan dalam bidang penelitian bahasa, seperti misalnya, pada penentuan sinonim.

Jika berbicara tentang penelitian yang berkaitan dengan sinonim, terutama sinonim bahasa Indonesia, sebagian besar dilakukan untuk melihat nuansa makna yang muncul pada suatu sinonim, seperti apa yang dilakukan oleh Permatasari (2018), Oktami, dkk. (2019), dan Permatasari, dkk. (2019). Ketiga penelitian itu berfokus pada analisis komponen makna. Selain itu, penelitian terkait sinonim bahasa Indonesia juga dilakukan oleh Arifin (2015). Pada kajiannya tersebut ia menjelaskan bagaimana klasifikasi kesinoniman berdasarkan Cruse (1986) diimplementasikan dalam Bahasa Indonesia. Hal tersebut menunjukkan bahwa penelitian terkait sinonim bahasa Indonesia cenderung fokus kepada sinonim yang telah ada yang kemudian dilihat bagaimana nuansa maknanya, sehingga penelitian tentang bagaimana suatu kata memiliki kedekatan dengan kata lain dan menjadi sebuah sinonim belum dilakukan, apalagi yang berkaitan dengan kesamaan semantik. Oleh sebab itu, penelitian ini membahas tentang pemanfaatan kesamaan semantik dalam membentuk calon sinonim.

METODE PENELITIAN

Penelitian ini menggunakan data berupa verba bahasa Indonesia yang diambil dari kumpulan artikel Wikipedia Bahasa Indonesia. Kumpulan artikel yang digunakan sebagai sampel data ini merupakan Dump atau arsip per Januari 2020. Kumpulan artikel tersebut dijadikan korpus untuk pemrosesan data serta untuk diambil verba yang memiliki kemunculan terbanyak sebagai sampel data. Korpus ini diperlukan sebab kesamaan semantik yang dilakukan pada penelitian ini menggunakan metode berbasis korpus (Chandrasekaran & Mago, 2020)

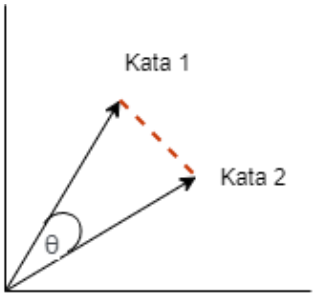
Pengumpulan data verba bahasa Indonesia ini dilakukan dengan bantuan komputer dengan bahasa pemrograman Python. Sebanyak 40 verba digunakan sebagai sampel data. 40 verba tersebut diambil dari seratus kata yang terbanyak muncul di dalam korpus. 100 kata terbanyak tersebut telah dilakukan pengecekan terhadap kelas katanya menggunakan KBBI Daring (2019).

Selain untuk melakukan pengumpulan data, bahasa pemrograman Python ini juga digunakan untuk mengolah kata menjadi vektor serta memproses pengukuran kesamaan semantik. Selain itu, proses pengolahan kata menjadi vektor tersebut juga dibantu dengan kepustakaan kode (*library*) Gensim (Řehůřek & Sojka, 2010) yang dipasang dan dijalankan menggunakan python. Proses pengolahan tersebut biasanya disebut sebagai *word embeddings.* Ketika kata-kata telah diubah menjadi vektor maka tahapan selanjutnya adalah pengukuran kesamaan semantik dengan menggunakan kosinus yang dijalankan dengan Word2vec yang dikembangkan oleh Tomas Mikolov dkk (2013). Proses tersebut menghasilkan nilai kedekatan dari verba yang diuji.

Prinsip yang digunakan dalam pengukuran kesamaan semantik ini menggunakan prinsip pengukuran kosinus, yaitu jika sudut yang dibentuk dari dua vektor 0 derajat maka hasilnya adalah 1, sedangkan jika sudut yang dibentuk dari dua vektor adalah 90 derajat, maka hasilnya adalah 0.

Gambar 1. Vektor dengan sudut 90 Derajat dan 0 Derajat

Berdasarkan prinsip pengukuran kosinus tersebut dapat dikatakan bahwa jika nilai kedekatan yang dihasilkan mendekati 1, maka kedua kata yang berpasangan tersebut memiliki tingkat kedekatan yang tinggi. Begitu juga sebaliknya, jika nilai kedekatan menjauhi nilai 1 dan justru mendekati nilai 0, maka kedua kata yang diukur tersebut memiliki kedekatan yang rendah atau semakin jauh.



Gambar 2. Dua Kata dalam Bentuk Vektor dengan Jarak Kedekatan Digambarkan dengan Garis Putus-Putus

HASIL DAN PEMBAHASAN

Pengukuran kesamaan semantik yang telah dilakukan pada empat puluh verba menghasilkan verba beserta kata terdekatnya. Hasil nilai terkecil dan terbesar dari pengukuran kesamaan semantik antara verba dan kata terdekatnya adalah 0,39 dan 0,94. Berdasarkan hasil tersebut maka perlu ditentukan ambang batas (*trace hold*) untuk menentukan pasangan verba mana saja yang akan menjadi calon sinonim. 0,6 digunakan sebagai ambang batas yang dilihat dari besarnya nilai serta ketidaksesuaian pasangan kata yang terbentuk. Dengan demikian, pasangan kata yang memiliki nilai kesamaan semantik kurang dari 0,6 tidak dapat dikategorikan menjadi calon sinonim.

Terdapat 4 pasangan kata yang memiliki nilai pengukuran kesamaan semantik di bawah 0,6 yaitu, "bernama" dan "benama" dengan nilai kedekatan 0,45; "tergolong" dan "skombride" dengan nilai kedekatan 0,39; "mengebor" dan "larva" dengan nilai kedekatan 0,56; serta "menjadi" dan "mejadi" dengan nilai kedekatan 0,59. Jika dilihat, dari empat pasang kata tersebut, terdapat dua pasang kata yang memiliki kesalahan pengetikan dan dua pasang kata yang hubungannya kedekatannya sangat jauh, sehingga nilai yang dihasilkan pun kecil.

Berdasarkan seleksi menggunakan ambang batas tersebut, maka tersisa 36 pasangan kata dengan nilai kedekatan di atas 0,6 yang memungkinkan untuk menjadi calon sinonim. Akan tetapi, dari 36 pasangan verba tersebut masih harus dilihat apakah terdapat kata terdekat yang mengalami kesalahan pengetikan, sebab adanya kesalahan pengetikan tersebut memungkinkan dianggap sebagai kata yang sama, sehingga memiliki nilai kedekatan yang cukup tinggi. Setelah dilakukan penelusuran pada 36 pasangan verba tersebut terdapat satu pasang kata yang memiliki kesalahan pengetikan namun memiliki nilai kedekatan yang relative tinggi. Pasangan tersebut adalah pasangan kata "lihat" dan "lihar" yang memiliki nilai kedekatan sebesar 0.73. Hal tersebut menunjukkan bahwa pasangan kata "lihat" dan "lihar juga tidak akan menjadi calon sinonim meskipun nilai kedekatannya di atas 0.6. Dengan demikian, pengukuran kesamaan semantik yang dilakukan pada 40 verba bahasa Indonesia ini menghasilkan 35 calon sinonim yang memiliki nilai kedekatan di atas 0,6.

Meskipun telah didapatkan 35 calon sinonim, untuk menjadikan calon sinonim itu menjadi sinonim perlu penelusuran dan analisis lebih lanjut, sebab dari 35 calon sinonim tersebut ditemukan beberapa kasus khusus. Kasus khusus tersebut seperti adanya pasangan calon sinonim yang memiliki bentuk dasar sama. Berikut tabel pasangan verba yang memiliki bentuk dasar sama.

Tabel 1. Pasangan Verba yang Memiliki Bentuk Dasar Sama

No	Verba yang Diuji	Hasil Verba Terdekat	Nilai Kedekatan	Bentuk Dasar
1	hidup	hidupnya	0,69	hidup
2	masuk	dimasukkan	0,69	masuk
3	berdasarkan	berdasar	0,76	dasar
4	berubah	mengubah	0,77	ubah
5	muncul	dimunculkan	0,71	muncul
6	memberikan	memberi	0,92	beri

Selain pasangan verba dengan bentuk dasar sama, calon sinonim yang dihasilkan dari pengukuran kesamaan semantik ini juga menghasilkan satu pasangan calon sinonim yang justru memiliki makna berlawanan, yaitu pasangan verba "berhasil" dan "gagal". Meskipun memiliki makna berlawanan,

pasangan verba tersebut justru memiliki nilai kedekatan yang cukup tinggi yaitu, 0,71. Kasus khusus yang dihasilkan dari pengukuran kesamaan semantik tersebut tidak akan dibahas pada penelitian ini. Oleh sebab itu, calon sinonim yang dihasilkan pada penelitian ini dapat dilanjutkan analisisnya pada penelitian-penelitian selanjutnya agar dapat menjadi pasangan sinonim sekaligus juga dapat dilakukan pembahasan yang lebih dalam terhadap kasus khusus yang muncul pada pengukuran kesamaan semantik.

KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan, dapat disimpulkan bahwa pengukuran kesamaan semantik dapat dimanfaatkan dalam menentukan calon sinonim dari verba bahasa Indonesia dengan beberapa syarat dan penelusuran yang lebih lanjut. Oleh sebab itu, penggunaan pengukuran kesamaan semantik ini dapat dijadikan sebagai langkah awal dalam menemukan pasangan kata yang bersinonim dalam sebuah korpus besar. Hal tersebut akan mempermudah pekerjaan para peneliti di bidang bahasa untuk melakukan penelitian, sebab tidak perlu secara manual menelusuri setiap kata-kata yang ada di dalam korpus dan mencari pasangan sinonimnya. Dengan demikian implementasi pengukuran kesamaan semantik dalam menentukan kedekatan antarverba terutama dalam hal menemukan sinonim cukup baik.

Berdasarkan simpulan yang telah disebutkan di atas, maka kita dapat melihat bagaimana proses kerja kesamaan semantik pada verba bahasa Indonesia dalam menentukan calon sinonim. Akan tetapi, ada beberapa hal yang masih perlu diperhatikan lebih lanjut seperti kemunculan kasus-kasus khusus yang dapat dibahas pada penelitian-penelitian selanjutnya

Selain itu, penelitian ini juga masih memiliki keterbatasan-keterbatasan lain, seperti belum dilakukan pemecahan masalah mengapa antonim tersebut muncul dengan nilai kedekatan yang relatif tinggi yang seolah-olah merupakan sinonim menggunakan metode lain. Oleh sebab itu, pengujian secara langsung tersebut menjadi salah satu hal yang dapat ditingkatkan dalam penelitian ini.

DAFTAR PUSTAKA

Badan Pengembangan Bahasa dan Perbukuan, Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2019, Oktober). *KBBI Daring*. Diambil kembali dari KBBI Daring: https://kbbi.kemdikbud.go.id/

Ali, A., Alfayez, F., & Alquhayz, H. (2018). SEMANTIC SIMILARITY MEASURES BETWEEN WORDS: A BRIEF SURVEY. *Sci. Int. (Lahore)*, 907-914.

Chandrasekaran, D., & Mago, V. (2020). Evolution of Semantic Similarity - A Survey. *ArXiv*(abs/2004.13820.).

Hermawan, R. F., Romadhony, A., & Faraby, S. A. (2017). Implementasi dan Analisis Kesamaan Semantik pada Bahasa Indonesia dengan Metode berbasis Vektor. *e-Proceeding of Engineering*, 4641-4649.

Huerta, J. M. (2008). Vector based Approaches. *Advances in Natural Language Processing and Applications*, 163-174.

Islam, A., & Inkpen, D. (2008). Semantic Text Similarity Using Corpus-Based. *ACM Trans. Knowl. Discov. Data*. doi:10.1145/1376815.1376819

Lin, D. (1998). *An Information-Theoretic Definition of Similarity.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems* (hal. 3111-3119). Lake Tahoe: Curran Associates Inc.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of Semantic Similarity. *Language and Cognitive Processes*, 1-228.

Pawar, A., & Mago, V. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. *arXiv*.

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic Cosine Similarity. *ICAST*.

Rani, M. M., Bijaksana, M. A., & Faraby, S. A. (2017). Analisis dan Implementasi Kesamaan Semantik Antar Teks Menggunakan Pendekatan Aligment dan Vektor Semantik pada Terjemahan Alquran. *e-Proceeding of Engineering*, 3254-3262.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New* (hal. 45-50). Valletta: ELRA.

Resnik, P. (1995). Using information Informaton Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of IJCAI-95*, (hal. 448-453). Montreal, Canada.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 627-633.

Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications*, 25-33. doi:10.5120/13897-1851

Sravanthi, P., & Srinivasu, B. (2017). SEMANTIC SIMILARITY BETWEEN SENTENCES. *International Research Journal of Engineering and Technology (IRJET)*, 156-161.

RIWAYAT HIDUP

Nama Lengkap	Institusi	Pendidikan	Minat Penelitian
Shiva Dwi Samara Tungga	Universitas Indonesia	S2	Linguistik Komputasional
Totok Suhardijanto	Universitas Indonesia	S3	Linguistik Komputasional