JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 6
No. 3
Desember
2020

Analisis Akurasi Algoritma Extended Word Similarity Based Clustering (EWSB) pada Mesin Penerjemah Bahasa Indonesia-Minang

Hendro Priyatman[*1], Muhammad Saleh[*2], Herry Sujaini[#3],

[*]*Jurusan Teknik Elektro, Fakultas Teknik, Universitas Tanjungpura*

[1]hendro.priyatman@ee.untan.ac.id

[2]muhammad.saleh@ee.untan.ac.id

[#]*Jurusan Informatika, Fakultas Teknik, Universitas Tanjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak

[3]hs@untan.ac.id

***Abstrak*— Extended Word Similarity Based (EWSB) Clustering adalah algoritma pengklasteran kata berdasarkan nilai kemiripan kata yang didapat dari hasil komputasi terhadap sebuah korpus. Salah satu manfaat dari hasil pengklasteran dengan algoritma ini adalah untuk meningkatkan kualitas output dari sebuah mesin penerjemah berbasis statistik (MPS). Dari hasil penelitian sebelumnya, hasil pengklasteran dengan algoritma EWSB terbukti memperbaiki akurasi mesin penerjemah bahasa Inggris sebagai Bahasa asal ke bahasa Indonesia sebagai Bahasa target, dimana algoritma tersebut diaplikasikan pada bahasa Indonesia sebagai bahasa target. Paper ini mendiskusikan hasil penelitian penggunaan EWSB pada MPS dari bahasa Indonesia ke bahasa Minang, dimana algoritma tersebut diaplikasikan pada bahasa Minang sebagai bahasa target. Penelitian yang dilakukan memperoleh hasil bahwa algoritma EWSB cukup efektif jika digunakan pada bahasa Minang sebagai bahasa target. Hasil penelitian ini menunjukkan bahwa penggunaan algoritma EWSB dapat meningkatkan tingkat akurasi terjemahan sebesar 6,36%.**

***Kata kunci*— Mesin penerjemah statistik, Extended Word Similarity Based (EWSB) Clustering, bahasa Indonesia-Minang**

I. PENDAHULUAN

Mesin penerjemah semakin hari semakin baik dengan banyak usaha para peneliti di bidang komputasi linguistik. Berbagai fitur-fitur dimanfaatkan untuk terus memperbaiki hasil dari terjemahan. Fitur-fitur seperti lemma dan *part-of-speech* diinduksi pada proses pelatihan data atau dapat juga digunakan sebagai informasi linguistik pada korpus. Pada penelitian-penelitian sebelumnya, Koehn dan H. Hoang [1] memaparkan bahwa dengan menmabhakan penandaan (*tag*) PoS pada mesin penterjemah English–German dapat meningkatkan keakuratan hasil terjemahan dari 18,04% menjadi 18,15%. Eksperimen pada bahasa English–Spanish (40.000 kalimat) dihasilkan 23,41% tanpa penambahan faktor, meningkat menjadi 24,25% dengan penambahan faktor morfologi dan PoS. Penelitian lain menambahkan PoS pada sistem penterjemah berbasis statistik, untuk sistem penerjemah English–Arabic yang dapat memperbaiki output terjemahan dari 60,95 % menjadi 63,94 %[2]. Razavian dkk. [3] menambahkan faktor linguistik pada mesin penterjemah berbasis statistik, untuk sistem penterjemah English–Iraqi (650.000 kalimat) memperbaiki output terjemahan dari 15,62% menjadi 16,41%, untuk sistem penterjemah Spanish-English (1.200.000 kalimat) dapat meningkatkan keakuratan hasil terjemahan dari 32,53% menjadi 32,84%, dan untuk sistem penterjemah Arabic-English (3.800.000 kalimat) dapat memperbaiki output terjemahan dari 41,70% menjadi 42,74%.

Untuk bahasa Indonesia, penelitian yang telah dilakukan oleh Sujaini dkk [4] memperlihatkan bahwa penggunaan Algoritma *Extended Word Similarity Based* (EWSB) Clustering pada mesin penerjemah statistik dapat meningkatkan akurasi sebesar 2,07% pada penerjemahan bahasa Inggris ke bahasa Indonesia. Sebaliknya, Sujaini dan Bijaksana [5] memperlihatkan bahwa penggunaan Algoritma EWSB pada mesin penerjemah statistik Bahasa Indonesia ke Bahasa Inggris malah menurunkan tingkat akurasi terjemahan sebesar 0,42%. Karena EWSB digunakan pada bahasa target, hal tersebut membuktikan bahwa EWSB hanya efektif untuk bahasa-bahasa yang menggunakan aturan "menerangkan-diterangkan" (MD). Bahasa daerah di Indonesia pada umumnya menggunakan aturan MD seperti bahasa Indonesia. Bahasa Minangkabau (bahasa Minang), Salah satu bahasa Melayu yang merupakan bahasa asli di Indonesia adalah Bahasa Minang. Berdasarkan ciri-ciri sosial, budaya maupun geografis, suku bangsa Minangkabau berdiam di kawasan Provinsi Sumatera Barat, kecuali Kepulauan Mentawai.

Provinsi Sumatera Barat kawasannya meliputi belahan barat bagian tengah pulau Sumatera dan kawasan Kepulauan Mentawai [6]. Belum banyak publikasi penelitian tentang mesin penerjemah bahasa Indonesia-Minang, salah satunya baru membahas tentang aplikasi kamus bahasa Minang [7]. Berdasarkan hal tersebut, penelitian ini menggunakan EWSB untuk meningkatkan akurasi mesin penerjemah bahasa Indonesia-Minang.

A. Mesin Penerjemah Statistik

Dalam sistem Mesin Penerjemah, basis pendekatan yang digunakan terdiri atas : (1) *Rule based* yang terdiri dari *Literal translation method, Transfer-based method,* dan *Interlingua-based method* serta (2) Corpus Based yang terdiri dari *Statistic-based method* dan *Case-Based method* [8]. Salah satu pendekatan yang popular pada mesin penerjemah yaitu pendekatan statistik dengan konsep probabilitas. Untuk setiap pasangan kalimat dalam dua bahasa yang berbeda, diberikan sebuah probabilitas hasil terjemahan dari Bahasa asal ke sumber yang diinterpretasikan sebagai distribusi probabilitas dalam bahasa targetnya saat diberikan kalimat input dalam bahasa sumber [9].

B. Kemiripan Kata (Word Similarity)

Kita selalu perlu menghitung kesamaan makna antar teks. Misalnya, mesin pencari perlu memodelkan relevansi dokumen dengan kueri, di luar tumpang tindih kata-kata di antara keduanya [10]. Contoh lain, situs tanya jawab perlu menentukan apakah sebuah pertanyaan telah ditanyakan sebelumnya. Dalam masalah hukum, tugas kesamaan teks memungkinkan untuk memitigasi risiko pada kontrak baru, berdasarkan asumsi bahwa jika kontrak baru serupa dengan yang sudah ada yang terbukti tangguh, risiko kontrak baru ini menjadi penyebab finansial. kerugian diminimalkan. Inilah prinsip prinsip Hukum Kasus. Penautan otomatis dokumen terkait memastikan bahwa situasi yang identik diperlakukan serupa di setiap kasus. Kesamaan teks menumbuhkan keadilan dan kesetaraan. Prioritas pengambilan dokumen hukum merupakan tugas pencarian informasi untuk mengambil dokumen kasus sebelumnya yang terkait dengan dokumen kasus tertentu.

Dalam layanan pelanggan, sistem kecerdasan buatan harus dapat memahami kueri yang serupa secara semantik dari pengguna dan memberikan respons yang seragam. Penekanan pada kemiripan semantik bertujuan untuk menciptakan sistem yang mengenali bahasa dan pola kata untuk menyusun tanggapan yang mirip dengan cara kerja percakapan manusia.

Kesamaan teks harus menentukan seberapa 'mirip' dua bagian teks dalam kesamaan leksikal dan makna (semantik) [11]. Kesamaan leksikal hanya mempertimbangkan kesamaan tingkat kata, kedua frasa ini tampak sangat mirip karena 3 dari 4 kata unik saling tumpang tindih. Ini biasanya tidak memperhitungkan arti sebenarnya di balik kata-kata atau seluruh frasa dalam konteks. Alih-alih melakukan perbandingan kata demi kata, kita juga perlu memperhatikan konteks untuk menangkap lebih banyak semantik. Untuk mempertimbangkan kemiripan semantik kita perlu fokus pada tingkat frase / paragraf (atau tingkat rantai leksikal) di mana sepotong teks dipecah menjadi sekelompok kata terkait yang relevan sebelum menghitung kesamaan. Meskipun kata-kata tersebut sangat tumpang tindih, kedua frasa ini sebenarnya memiliki arti yang berbeda.

Menggabungkan struktur taksonomi leksikal dengan informasi statistik korpus sehingga jarak semantik antar node dalam ruang semantik yang dibangun oleh taksonomi dapat dikuantifikasi dengan lebih baik dengan bukti komputasi yang berasal dari analisis distribusi data korpus [12]. Penelitian ini menggunakan metode distribusional untuk menentukan nilai kemiripan kata karena pertimbangan belum adanya sistem *thesaurus* yang memadai untuk Bahasa Indonesia.

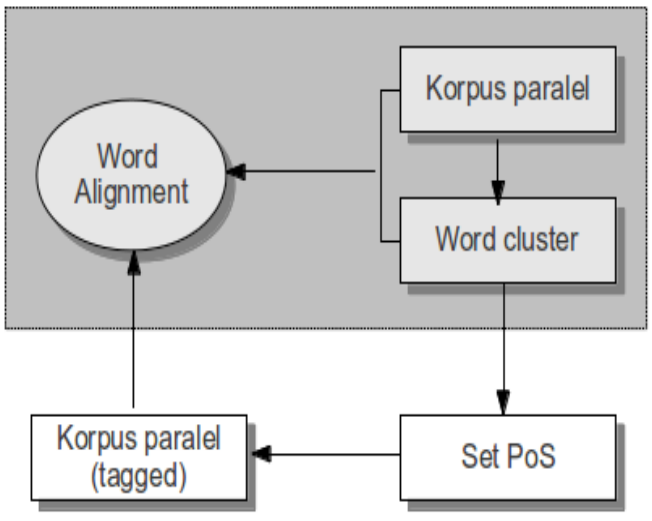
II. METODE PENELITIAN

Penelitian ini mempergunakan data korpus bilingual bahasa Indonesia-Minang sebesar 3.000 kalimat dan korpus monolingual bahasa Minang sebesar 3.000 kalimat yang diambil dari bahasa target korus paralel. Proses pengklasteran dengan algoritma EWSB juga dilakukan terhadap 3.000 kalimat bahasa Minang yang diambil dari korpus paralel.

Pemrosesan data dan pengujian system menggunakan beberapa sistem, yaitu :

1. Moses : digunakan sebagai mesin penerjemah,
2. SRILM : digunakan untuk membangun language model,
3. Giza++ : digunakan untuk proses penyelarasan kata,
4. BLEU : digunakan untuk penilaian hasil translasi, dan
5. Perl : digunakan untuk membangun program dari algoritma EWSB.

Posisi eksperimen induksi kelas kata tanpa supervisi pada penelitian ini seperti terlihat pada Gambar 2.



Gambar. 2 Posisi eksperimen induksi kelas kata

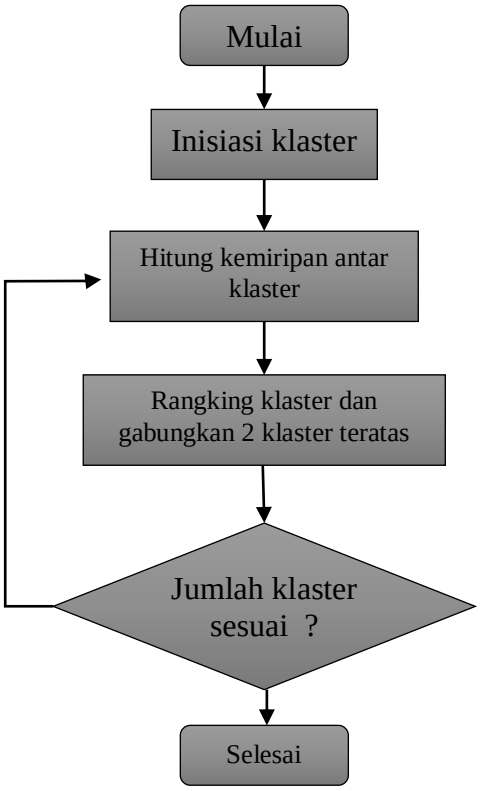
Eksperimen ini menggunakan pendekatan pengklasteran secara otomatis yang dihasilkan dari korpus untuk melihat efektifitas algoritma EWSB pada bahasa Minang sebagai bahasa target.

Algoritma yang dipergunakan untuk pengklasteran kata mengacu pada algoritma-algoritma pengklasteran (*clustering*) secara umum seperti: *connectivity based clustering (hierarchical clustering), distribution-based clustering, centroid-based clustering, density-based clustering* dan lain-lain [13-15]. Secara umum ada 2 pendekatan pengklasteran kata dengan algoritma *hierarchical clustering,* yaitu: *Agglomerative* dan *Divisive.*

Pengklasteran kata pada penelitian ini menggunakan pendekatan *Agglomerative* dengan algoritma *hierarchical clustering* untuk keperluan pengklasteran kata seperti yang ditunjukkan oleh Sujaini dkk. [4] sebagai berikut:

1. Tetapkan setiap kata unik sebagai satu klaster
2. Kalkulasikan nilai kemiripan antara dua klaster
3. Urutkan semua pasangan klaster berdasarkan nilai kemiripannya, kemudian gabungkan dua klaster teratas.
4. Berhenti sampai pada jumlah klaster yang diinginkan, jika belum, kembali ke langkah 2.

Untuk menghitung kemiripan antara dua klaster pada langkah 2, digunakan metode *average linkage clustering*. Flowchart algoritma pengklasteran dapat dilihat pada Gambar 3.



Gambar. 3 Flowchart algoritma Hierarchical Clustering

Kualitas output hasil terjemahan sistem diukur dengan menggunakan metode BLEU. Validasi asil pada eksperimen ini menggunakan *cross fold validation*, dimana 3.000 kalimat pada data korpus dibagi atas 6 fold yaitu :

fold 1 : kalimat no 1-500
fold 2 : kalimat no 501-1000,
fold 3 : kalimat no 1001-1500,
fold 4 : kalimat no 1501-2000,
fold 5 : kalimat no 2001-2500, dan
fold 6 : kalimat no 2501-3000.

Masing-masing fold diterapkan pada dua mesin, yaitu MPS dengan menggunakan mkcls sebagai baseline dan MPS yang menggunakan algoritma EWSB.

III. HASIL DAN PEMBAHASAN

Dari eksperimen yang telah dilakukan, didapat hasil pengujian terhadap masing-masing grup uji dapat dilihat pada Tabel 1. Sistem yang menggunakan mkcls sebagai algoritma acuan dari GIZA++ menghasilkan rata-rata nilai BLEU sebesar 63,25%, sedangkan penggunaan algoritma EWSB menghasilkan rata-rata nilai BLEU sebesar 67,27 %. Hal ini menunjukkan bahwa penggunaan algoritma EWSB dapat meningkatkan tingkat akurasi terjemahan sebesar ((67,27-63,25)/ /63,25) * 100% = 6,36%. Hasil penelitian menunjukkan bahwa algoritma EWSB cukup efektif jika digunakan pada bahasa Minang sebagai bahasa target.

TABEL I
NILAI BLEU HASIL PENGUJIAN AKURASI TERJEMAHAN

Grup Uji	Korpus (fold)	Kaliamt Uji (fold)	BLEU Score (%)	
			MKCLS	EWSB
A	2,3,4,5,6	1	63,44	67,98
B	1,3,4,5,6	2	61,21	65,82
C	1,2,4,5,6	3	62,13	65,91
D	1,2,3,5,6	4	63,90	66,67
E	1,2,3,4,6	5	64,79	68,72
F	1,2,3,4,5	6	64,02	68,54

Beberapa contoh hasil terjemahan dengan menggunakan mkcls sebagai perbandingan dan algoritma EWSB dapat dilihat pada Tabel 2.

Dari beberapa contoh hasil terjemahan dengan menggunakan MKCLS dan EWSB, terlihat bahwa hasil terjemahan menunjukkan perbedaan yang cukup sifnifikan, sebagai contoh hasil terjemahan kalimat “sudah sekitar enam jam sekarang”, terjemahan referensi “alah sekitar jam anam kini ko” menjadi “alah sekitar anam jam kini” pada MKCLS terbalik antara kata “enam” dan “jam” dan diperbaiki menjadi “alah sekitar jam anam kini” pada EWSB. Hasil terjemahan kalimat “saya tidak bisa menjamin, tapi kami akan mencoba yang terbaik”, pada frase terjemahan referensi “kito akan mancubo yang tabaiek” menjadi “kami akan mancubo nan ka elok” pada

MKCLS diperbaiki menjadi “kito akan mancubo yang tabaiek” pada EWSB. Hasil terjemahan kalimat “saya sudah membayar uang muka untuk makan dan hotel”, pada frase terjemahan referensi “ambo alah membayie pitih muko” menjadi “ambo alah membayie pitih ka pitih muko” pada MKCLS diperbaiki menjadi “ambo alah membayie pitih muko” pada EWSB. Contoh lainnya adalah hasil terjemahan kalimat “kalau saya , biasanya urusan kantor , jarang ada waktu untuk bersenang-senang”, pada frase terjemahan referensi biasonyo urusan kantua” menjadi “dek biasonyo urusan kantua” pada MKCLS diperbaiki menjadi “biasonyo urusan kantua” pada EWSB.

TABEL II
PERBANDINGAN HASIL TERJEMAHAN DENGAN MENGGUNAKAN MKCLS DAN EWSB

No		Kalimat
1	Input	sudah sekitar enam jam sekarang
	Ref	alah sekitar jam anam kini ko
	MKCS	alah sekitar anam jam kini (BLEU= 0.00 %)
	EWSB	alah sekitar jam anam kini (BLEU= 81.87 %)
2	Input	saya tidak bisa menjamin , tapi kami akan mencoba yang terbaik
	Ref	ambo indak bisa menjamin , tapi kito akan mancubo yang tabaiek
	MKCS	ambo indak bisa menjamin , tapi kami akan mancubo nan ka elok (BLEU = 46,92 %)
	EWSB	ambo indak bisa menjamin , tapi kito akan mancubo yang tabaiek (BLEU = 100,00 %)
3	Input	saya sudah membayar uang muka untuk makan dan hotel
	Ref	ambo alah membayie pitih muko untuak makan dan hotel
	MKCS	ambo alah membayie pitih ka pitih muko untuak makan dan hotel (BLEU = 68,34%)
	EWSB	ambo alah membayie pitih muko untuak makan dan hotel (BLEU = 100,00%)
4	Input	kalau saya , biasanya urusan kantor , jarang ada waktu untuk bersenang-senang
	Ref	kalau ambo , biasonyo urusan kantua , jarang ado waktu untuak basanang-sanang
	MKCS	kalau ambo , dek biasonyo urusan kantua , jarang ado waktu untuak basanang-sanang (BLEU = 76,12%)
	EWSB	kok ambo , biasonyo urusan kantua , jarang ado waktu untuak basanang-sanang (BLEU = 90,36%)

Masih rendahnya nilai BLEU pada kedua mesin yang menggunakan algoritma berbeda ini adalah karena masih minimnya kalimat yang terdapat pada korpus. Masalah ini memang merupakan masalah utama untuk mengembangkan mesin penerjemah yang memiliki akurasi yang tinggi. Meskipun dengan korpus terbatas, algoritma EWSB yang digunakan untuk pengklasteran kata pada proses pra-proses pembangunan mesin penerjemah statistik sudah dapat meningkatkan akurasi penerjemahan.

IV. KESIMPULAN

Mesin penerjemah bahasa Indonesia-Minang yang dibangun pada penelitian ini memperoleh hasil akurasi yang diwakili oleh nilai BLEU sebesar 67,27%. Nilai ini sangat dipengaruhi oleh kecilnya kuantitas korpus yang digunakan dalam penelitian ini.

Dari eksperimen yang dilakukan terhadap mesin penerjemah statistik dengan menggunakan algoritma EWSB, penggunaan algoritma tersebut dapat meningkatkan tingkat akurasi terjemahan sebesar 6,36% dibandingkan dengan menggunakan MKCLS, Hasil penelitian menunjukkan bahwa algoritma EWSB cukup efektif jika digunakan pada bahasa Minang sebagai bahasa target. Walaupun demikian, perlu dilakukan penelitian lebih jauh dengan korpus yang lebih besar, demikian juga penelitian terhadap bahasa-bahasa daerah lainnya yang digunakan di seluruh nusantara.

REFERENSI

[1] Koehn, P., dan Hoang, H., 2007. “Factored translation models”, Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

[2] Youssef, I., Sakr, M., dan M. Kouta, 1999. Linguistic factors in statistical machine translation involving arabic language, IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.11.

[3] Razavian, N. Sharif, dan Vogel, S., 2010. “Fixed length word suffix for factored statistical machine translation”, Proceedings of the ACL 2010 Conference Short Papers, Uppsala.

[4] Sujaini, H., Kuspriyanto, Arman, A.A. dan Purwarianti, A., 2013. Extended Word Similarity Based Clustering on Unsupervised PoS Induction to Improve English-Indonesian Statistical Machine Translation, 16th ORIENTAL COCOSDA/CASLRE-2013, Gurgaon, India.

[5] Sujaini H., dan Bijaksana, A., 2015. “Analisis Penggunaan Algoritma EWSB pada Mesin Penerjemah Bahasa Indonesia-Inggris”, Seminar Nasional FORTEI 2015, Pontianak.

[6] Juzrizal, 2012. “Tatabahasa Bahasa Minangkabau”. UNP Press. Padang

[7] Efendi, R., Fitri M., dan Andreswari, D., 2014. Andreswari, “Rancang Bangun Aplikasi Kamus Bahasa Indonesia-Minang, Minang-Indonesia Berbasis Android”, Jurnal Ilmiah Bidang Sains – Teknologi Murni Disiplin dan Antar Disiplin. Vol.1 No.14

[8] Peng, L. A., 2013. “Survey of Machine Translation Methods”. TELKOMNIKA Indonesian Journal of Electrical Engineering. 11(12): 7125-7130

[9] Sujaini, H., Kuspriyanto, Arman, A.A. dan Purwarianti, 2012. “Pengaruh part-of-speech pada mesin penerjemah bahasa inggris-

indonesia berbasis factored translation model", SNATI 2012, Yogyakarta.
[10] Maskur, 2014. "Relevansi Hasil Pencarian Pada Mesin Pencari Berdasarkan Kedekatan Kata Menggunakan Ontologi", Junal Gamma, Vol. 10, No. 1., 123-129
[11] Hermawan, R.F., Romadhony, A., Al-Faraby, S., 2017. Implementasi dan Analisis Kesamaan Semantik pada Bahasa Indonesia dengan Metode berbasis Vektor, e-Proceeding of Engineering, Vol.4, No.3, 4641-4649.
[12] Jiang, J.J., and Conrath, D.W. 1997, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), Taiwan.
[13] Hartuv, E., and Shamir, R., 2000. "A Clustering Algorithm Based on Graph Connectivity", Information processing letters, vol. 76, no. 4-6.
[14] Uppada, S.K. 2014. Centroid Based Clustering Algorithms - A Clarion Study, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, NO. 6, 7309-7313.
[15] Mumtaz, K.K. and Duraiswamy, K. 2010. An Analysis on Density Based Clustering of Multi Dimensional Spatial Data, Indian Journal of Computer Science and Engineering, Vol. 1, No. 1.