Peningkatan Akurasi Penerjemah Bahasa Daerah dengan Optimasi Korpus Paralel

Herry Sujaini[1]

***Abstract*—Statistical Machine Translation (SMT) quality is influenced by several factors. The most fundamental factor is quantity of corpus used as base material for building translational and language model in SMT. Quantity of corpus is a major factor in ensuring quality of the translation, but quality of corpus can not be ignored either. Checking the source and translation sentences manually in a parallel corpus of course will be very difficult and require a lot of resources. This paper reports the experimental results using a quality improvement strategy of Indonesian-Malay and Indonesia-Javanesse corpus without having to examine and correct the sentences that exist on the corpus. The filter used is the minimum value of each sentence tested by the Bilingual Evaluation Understudy (BLEU) method. Experimental results show that parallel corpus optimization can improve the level of accuracy of Indonesian-Malay translation by 6.97% and Indonesian-Javanesse translation by 5.55%.**

***Intisari*—Kualitas terjemahan dari mesin penerjemah berbasis statistik (MPS) dipengaruhi oleh beberapa faktor. Faktor yang paling mendasar adalah kuantitas korpus yang digunakan sebagai bahan dasar untuk membangun model translasi dan model bahasa yang digunakan pada MPS. Selain kuantitas korpus yang menjadi faktor utama untuk menjamin kualitas hasil terjemahan, kualitas korpus juga tidak bisa diabaikan. Pemeriksaan kalimat-kalimat sumber dan terjemahannya secara manual dalam sebuah korpus paralel tentu saja sangat sulit dan memerlukan sumber daya yang besar. Makalah ini melaporkan hasil eksperimen menggunakan strategi perbaikan kualitas korpus bahasa Indonesia-Melayu dan bahasa Indonesia-Jawa tanpa harus memeriksa dan memperbaiki kalimat-kalimat yang ada pada korpus. Filter yang digunakan adalah nilai minimal setiap kalimat yang diuji dengan metode *Bilingual Evaluation Understudy* (BLEU). Hasil eksperimen menunjukkan bahwa optimasi korpus paralel dapat meningkatkan tingkat akurasi terjemahan bahasa Indonesia-Melayu sebesar 6,97% dan bahasa Indonesia-Jawa sebesar 5,55%.**

***Kata Kunci*—mesin penerjemah statistik, optimasi korpus, korpus paralel, bahasa Indonesia-Melayu**

I. PENDAHULUAN

Mesin penerjemah merupakan alat yang dapat dipergunakan untuk melakukan proses penerjemahan teks dari satu bahasa ke bahasa lainnya secara otomatis. Mesin penerjemah statistik (MPS) adalah sebuah pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan berdasarkan model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel, atau biasa disebut dengan korpus bilingual dari dua bahasa yang berbeda.

Penelitian dalam bidang MPS di Indonesia, terutama untuk mesin penerjemah bahasa Indonesia ke bahasa daerah, sudah mulai banyak dilakukan, di antaranya adalah penelitian akurasi penerjemahan bahasa Indonesia-Jawa menggunakan metode statistik berbasis frasa [1], penelitian tentang pengaruh kuantitas korpus terhadap akurasi MPS bahasa Bugis Wajo ke bahasa Indonesia [2], penelitian dengan memperbaiki probabilitas *Lexical Model* untuk meningkatkan akurasi MPS bahasa Indonesia-Jawa [3], penandaan kata dasar dan imbuhan pada korpus paralel untuk memperbaiki akurasi penerjemahan bahasa Indonesia–Dayak Taman [4], penelitian *tuning for quality* untuk uji akurasi MPS bahasa Indonesia-Dayak Kanayatn [5], serta penelitian tentang sistem penerjemah bahasa Jawa-aksara Jawa berbasis *finite state automata* [6].

Sebuah penelitian mengimplementasikan MPS pada sistem penerjemah situs dari bahasa Indonesia ke bahasa Melayu Pontianak. Sistem tersebut menghasilkan halaman situs baru yang persis sama dengan halaman situs aslinya, dengan kalimat-kalimat yang baru berbahasa Melayu Pontianak. Meskipun sistem telah dapat menghasilkan halaman situs terjemahan, tetapi hasil terjemahannya sendiri masih belum mencapai tingkat akurasi yang memuaskan, yaitu sebesar 63,02% [7].

Kualitas terjemahan MPS dapat ditingkatkan dengan memperbanyak kuantitas dan memperbaiki kualitas korpus. Selain itu, kualitas terjemahan juga dapat ditingkatkan dengan menambahkan informasi linguistik pada tingkat kata pada korpus. Perbaikan atau pengembangan algoritme-algoritme yang digunakan baik pada praproses atau algoritme pada proses penerjemahan juga dapat memperbaiki kualitas terjemahan.

Metode untuk memperbaiki kualitas korpus dengan cara memfilter kalimat-kalimat yang berkualitas dari sebuah korpus paralel bahasa Inggris-Indonesia telah diusulkan [8]. Pemilihan kalimat-kalimat berkualitas pada korpus paralel sedianya dilakukan secara manual, tetapi tentu saja memerlukan waktu dan tingkat ketelitian yang tinggi, selain memerlukan sumber daya manusia yang besar dan memadai. Makalah ini menggunakan metode tersebut untuk meningkatkan akurasi penerjemah bahasa Indonesia-Melayu dan Indonesia-Jawa. Bahasa Melayu yang digunakan adalah bahasa melayu Pontianak, sedangkan bahasa Jawa yang digunakan adalah bahasa Jawa Kromo.

Mesin penerjemah merupakan mesin yang dapat melakukan penerjemahan secara otomatis yang berarti sebuah komputer mengambil alih semua pekerjaan penerjemahan. Komputer tentu saja akan bekerja lebih cepat dan lebih murah daripada manusia. Dalam dua dekade terakhir ini, penelitian pada bidang mesin penerjemah mengarah pada model

[1]*Dosen, Program Studi Teknik Informatika, Universitas Tanjungpura, Jalan Prof. Hadari Nawawi Pontianak 78124 Indonesia (tlp: 0561-740 186; fax: 0561-739630; e-mail: hs@untan.ac.id)*

penerjemahan yang dibangun secara otomatis dari korpus. Model yang biasanya disebut MPS ini menggunakan pendekatan teknik statistik.

Pada awalnya, penelitian tentang MPS dimulai dengan model berbasis kata yang diterjemahkan kata demi kata [9]. Model yang dikembangkan ini sebagian besar telah diganti oleh model yang lebih kompleks, tetapi tetap digunakan sebagai dasar untuk model kelanjutannya, yaitu penyelarasan kata (*word alignment*) [10]. Beberapa peneliti mengusulkan model berbasis frasa yang dapat menerjemahkan kalimat berdasarkan kata-kata yang berurutan dalam kalimat sumber untuk kata yang bersesuaian pada bahasa target [11], [12]. Ungkapan istilah frasa disini hanya berarti kata-kata yang berdekatan, bukan frasa sebenarnya dalam istilah tata bahasa. Awalnya, model berbasis frasa berakar dari hasil penelitian oleh beberapa penelitian sebelumnya [13]--[15]. Selain itu, penerjemahan dengan penggunaan frasa juga diusulkan oleh beberapa peneliti [16]--[18]. Kemudian, terdapat pula penelitian yang mengusulkan penggunaan frasa dalam model kata berbasis *decoding* [19], dilanjutkan penggunaan model *log-linier* [20].

MPS dapat didefinisikan sebagai proses memaksimalkan probabilitas kalimat dalam bahasa sumber, *s*, yang sesuai dengan kalimat terjemahannya dalam bahasa target, *t*. Dengan kata lain, "dengan memberikan kalimat *s* dalam bahasa sumber, kemudian dicari kalimat *t* dalam bahasa target sehingga memaksimalkan *P(t|s)* yang disebut probabilitas bersyarat atau kesempatan dari terjadinya *t* jika diberikan *s*" [21].

Hal tersebut juga dapat disebut dengan terjemahan yang memiliki probabilitas terbaik, secara formal ditulis seperti yang ditunjukkan pada (1).

$$\arg\max P(t \mid s) \quad (1)$$

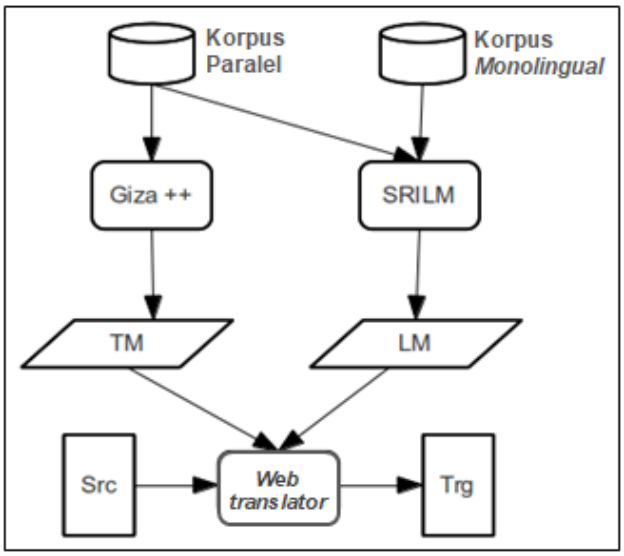
Dengan menggunakan Aturan Bayes dari (2), rumusan pada (1) untuk terjemahan yang paling mungkin terjadi dapat dituliskan seperti yang ditunjukkan pada (3).

$$P(t|s) = P(t) * P(s|t) = P(s) \quad (2)$$

$$\arg\max P(t|s) = \arg\max P(t) * P(s|t) \quad (3)$$

dengan *t* merupakan kalimat target dan *s* merupakan kalimat sumber. *P(t)* merupakan model bahasa (*language model*/LM) dan *P(s|t)* merupakan model translasi (*translation model*/TM). Operasi argmax adalah proses pencarian yang dilakukan oleh *decoder* yang merupakan bagian dari sistem MPS.

Arsitektur MPS, khususnya *decoder* Moses [21], dapat dijelaskan seperti pada Gbr. 1. Sumber data utama MPS adalah korpus paralel dan korpus *monolingual*. Korpus paralel dilatih menggunakan GIZA++ sehingga menghasilkan model translasi. Proses pelatihan terhadap bahasa target pada korpus paralel ditambah dengan korpus *monolingual* bahasa target menggunakan SRILM menghasilkan model bahasa. MODEL TRANSLASIdan model bahasa hasil proses tersebut digunakan untuk menghasilkan *decoder* Moses. Selanjutnya, Moses digunakan sebagai *decoder* mesin penerjemah untuk menghasilkan bahasa target dari masukan kalimat dalam bahasa sumber.

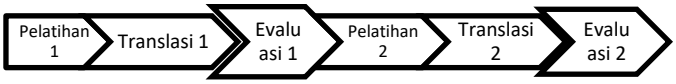


Gbr. 1 Arsitektur mesin penerjemah statistik [8].

Dari Gbr. 1 terlihat jelas bahwa bahan baku yang digunakan untuk menghasilkan model-model pada MPS adalah korpus paralel. Korpus *monolingual* dapat diperoleh dari korpus paralel pada sisi bahasa target meskipun biasanya diperbanyak lagi dari sumber-sumber lainnya. Penelitian terhadap pengaruh kuantitas dan kualitas korpus telah dilakukan pada beberapa pasangan bahasa, di antaranya untuk bahasa Inggris-Turki [22], bahasa Inggris-Estonia [23], bahasa Inggris-Hindi [24], dan Indonesia-Inggris [8]. Pada makalah ini, penelitian menggunakan bahasa Indonesia sebagai bahasa sumber, sedangkan bahasa target yang digunakan adalah bahasa Melayu dan Jawa.

II. METODOLOGI

Korpus merupakan kumpulan kalimat-kalimat dalam bahasa tertentu. Eksperimen yang dilakukan menggunakan korpus paralel bahasa Indonesia-Melayu sebesar 12.000 kalimat dan korpus *monolingual* bahasa Melayu dan Jawa masing-masing sebesar 50.000 kalimat. Sedangkan untuk korpus paralel bahasa Indonesia-Jawa, digunakan 5.100 kalimat korpus paralel dan 10.000 kalimat korpus *monolingual*. Mesin pertama sebagai *baseline* menggunakan korpus asal yang belum dioptimasi dibandingkan dengan mesin kedua yang menggunakan korpus hasil optimasi.



Gbr. 2 Tahapan metode eksperimen.

Instrumen penelitian yang digunakan adalah sebagai berikut:

1. Moses: digunakan sebagai mesin penerjemah,
2. SRILM: digunakan untuk membangun model bahasa,
3. Giza++: digunakan untuk proses penyelarasan kata, dan
4. BLEU: digunakan untuk penilaian hasil translasi [25].

Metode yang digunakan untuk menghasilkan MPS dengan korpus yang lebih berkualitas sampai pada evaluasi hasil

translasi MPS yang telah dioptimasi adalah seperti ditunjukkan pada Gbr. 2.

Proses pelatihan 1 melakukan pelatihan untuk memperoleh model bahasa dan MODEL TRANSLASI dengan menggunakan korpus awal. Selanjutnya, model-model tersebut digunakan pada *decoder* (Translasi 1). Dari sumber korpus paralel, semua kalimat yang ada pada bahasa sumber digunakan sebagai masukan pada Translasi 1 dengan bahasa target sebagai referensi. Selanjutnya, pada proses Evaluasi 1, masing-masing nilai BLEU kalimat terjemahannya dihitung terhadap kalimat referensi. Semua kalimat yang memiliki nilai BLEU yang kurang dari n% dieliminasi, sisanya digunakan sebagai korpus paralel untuk mesin yang baru. Berdasarkan hasil penelitian sebelumnya, ambang batas nilai BLEU yang digunakan, *n*, untuk memfilter kalimat-kalimat dalam korpus paralel adalah sebesar 10% [8].

Proses pelatihan 2 melakukan pelatihan untuk memperoleh model bahasa dan MODEL TRANSLASI dengan menggunakan korpus hasil optimasi. Model-model tersebut digunakan pada *decoder* (Translasi 2). Selanjutnya, dilakukan proses Evaluasi 2 untuk mengukur nilai BLEU dari mesin yang baru. Akurasi hasil terjemahan sistem Indonesia-Melayu diukur dengan menggunakan 12.000 kalimat yang dibagi atas enam *fold*, yaitu: *fold* 1: kalimat nomor 1--2.000, *fold* 2: kalimat nomor 2.001--4.000, *fold* 3: kalimat nomor 4.001--6.000, *fold* 4: kalimat nomor 6.001--8.000, *fold* 5: kalimat nomor 8.001--10.000, dan *fold* 6: kalimat nomor 10.001--12.000. Sedangkan sistem Indonesia-Jawa menggunakan 5.100 kalimat yang dibagi atas enam *fold*, yaitu: *fold* 1: kalimat nomor 1--850, *fold* 2: kalimat nomor 851--1.700, *fold* 3: kalimat nomor 1.701--2.550, *fold* 4: kalimat nomor 2.551--3.400, *fold* 5: kalimat nomor 3.401--4.250, dan *fold* 6: kalimat nomor 4.251--5.100.

III. HASIL DAN PEMBAHASAN

Korpus paralel bahasa Indonesia-Melayu yang digunakan pada eksprimen ini dikerjakan secara manual dengan menerjemahkan buku cerita berbahasa Melayu Pontianak yang berjudul “Sepok” ditambah beberapa dokumen yang diambil dari internet. Sedangkan Korpus paralel bahasa Indonesia-Jawa dibangun dari cerita Kabayan yang diterjemahkan secara manual. Kedua korpus tersebut kemudian dibersihkan (proses *cleaning*), ditokenisasi (*tokenizing*) dan dijadikan huruf kecil semua hurufnya. Contoh kalimat dalam korpus yang telah siap digunakan untuk proses pelatihan ditunjukkan pada Gbr. 3 dan Gbr. 4.

Setelah melewati tahap praproses, dilakukan pelatihan terhadap korpus paralel. Model bahasa Melayu dan Jawa yang dilatih menggunakan SRILM memberikan hasil yang potongannya ditunjukkan pada Gbr. 5. Sebagai contoh, probabilitas kemunculan urutan kata “dah” diikuti “tahu semue” dalam bahasa Melayu adalah 10^(-0,7733381) = 0,169, adapun kemunculan urutan kata “duwe” diikuti “dhuwit kanggo” dalam bahasa Jawa memiliki probabilitas 10^(-0,8332769) = 0,147. Token <s> diartikan sebagai awal kalimat, sedangkan </s> diartikan sebagai akhir kalimat.

aku tidak mau	→	aku tadak maok
aku takut	→	takot aku
mereka berdua memaksa	→	budak duak tu makse
kalau bermain lumpur aku sudah puas waktu kecil	→	kalok maen lumpor aku ni puas agek kecik
eh, ya sudahlah, kutebalkan muka ku saja	→	eh, dahlah, kutebalkan muke jaklah

Gbr. 3 Contoh kalimat dalam korpus paralel Indonesia-Melayu.

Lahirnya Itok.	→	Laire Itok.
Setelah Nyi Iteung hamil, orang serumah semua direpotkan.	→	Sawise Nyi Iteung mbobot, wong saomah kabeh direpotake.
Maklum namanya baru hamil muda, ada-ada saja yang diminta dan yang aneh-aneh.	→	Maklum jenenge lagi ngandheg enom, ana-ana wae sing dijaluk lan sing aneh-aneh.
Hal ini tentu saja membuat bingung orang serumah.	→	Bab iki mesthi wae gawe bingunge wong saomah.
Si Kabayan bingung sekali menghadapi sikap dan permintaan mainan	→	Si Kabayan bingung banget ngadhepi sikep lan panjaluke bojone.
Sedang Abah dan Ambu yang sesudah wanita pengalaman bisa mengerti hal itu.	→	Dene Abah lan Ambu sing wis duwe pengalaman bisa ngerteni bab iku.

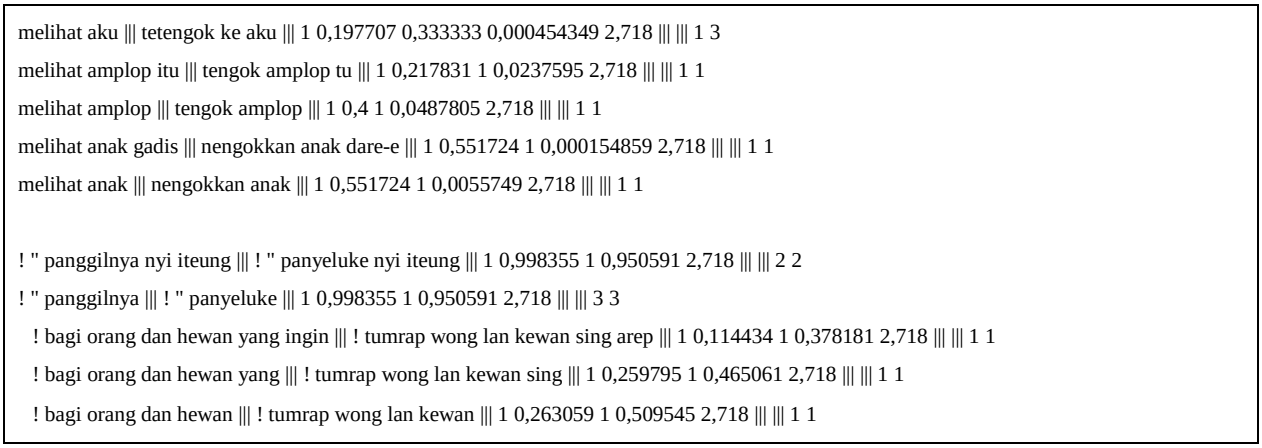
Gbr. 4 Contoh kalimat dalam korpus paralel Indonesia-Jawa.

-1,417884	→	<s> tapi tak
-0,7733381	→	dah tau semue
-0,5972468	→	die tau gak
-0,8982768	→	kitak tau ndak
-1,500337	→	tak tau </s>
-0,8332769	→	duwe dhuwit kanggo
-0,9124582	→	ing dhuwur panggung
-0,4353369	→	pantes dianggo nggeret
-0,4353369	→	wis dienteni abah
-0,3746373	→	<s> dina esuke

Gbr. 5 Potongan model bahasa Melayu dan Jawa.

Potongan model translasi yang dilatih dengan GIZA++, baik untuk Indonesia-Melayu maupun Indonesia-Jawa, ditunjukkan pada Gbr. 6. Model translasi yang diperoleh merupakan potongan-potongan frasa yang dikomputasi dari korpus paralel masing-masing pasangan bahasa. Sebagai contoh, model translasi bahasa Indonesia-Melayu memperlihatkan bahwa probabilitas terjemahan dari frasa “melihat aku” menjadi “tetengok ke aku” adalah sebesar 0,3333 atau 33,33%. Model bahasa dan model translasi tersebut digunakan pada *decoder* MOSES sebagai mesin penerjemah.

Dari eksperimen yang telah dilakukan terhadap bahasa Indonesia-Melayu, diperoleh hasil pengujian terhadap masing-masing grup uji seperti yang disajikan pada Tabel I, sedangkan untuk bahasa Indonesia-Jawa ditunjukkan pada Tabel II.



Gbr. 6 Potongan Model Translasi Indonesia-Melayu dan Indonesia-Jawa.

TABEL I
NILAI BLEU PENGUJIAN AKURASI INDONESIA-MELAYU

Grup Uji	Korpus (*fold*)	Kalimat Uji (*fold*)	Skor BLEU (%)	
			Baseline	**Korpus Baru**
A	2,3,4,5,6	1	68,78	73,40
B	1,3,4,5,6	2	66,20	71,04
C	1,2,4,5,6	3	63,45	68,70
D	1,2,3,5,6	4	64,23	68,48
E	1,2,3,4,6	5	65,42	70,22
F	1,2,3,4,5	6	67,34	71,15

TABEL II
NILAI BLEU PENGUJIAN AKURASI INDONESIA-JAWA

Grup Uji	Korpus (fold)	Kalimat Uji (*fold*)	Skor BLEU (%)	
			Baseline	**Korpus Baru**
A	2,3,4,5,6	1	46,22	49,42
B	1,3,4,5,6	2	47,01	49,88
C	1,2,4,5,6	3	45,33	46,98
D	1,2,3,5,6	4	42,06	45,08
E	1,2,3,4,6	5	43,98	46,35
F	1,2,3,4,5	6	44,45	46,28

Dari Tabel I, dapat dihitung rata-rata nilai BLEU yang merepresentasikan akurasi dari sistem penerjemah Indonesia-Melayu. Sistem *baseline* menghasilkan rata-rata nilai BLEU sebesar 65,90%, sedangkan sistem dengan korpus paralel hasil optimasi menghasilkan rata-rata nilai BLEU sebesar 70,50%. Hal ini menunjukkan bahwa optimasi korpus paralel meningkatkan akurasi terjemahan Indonesia-Melayu sebesar 6,97%. Dari data pada Tabel II dapat dihitung rata-rata nilai BLEU yang merepresentasikan akurasi sistem penerjemah Indonesia-Jawa. Sistem *baseline* menghasilkan rata-rata nilai BLEU sebesar 44,84%, sedangkan sistem dengan korpus paralel hasil optimasi menghasilkan rata-rata nilai BLEU sebesar 47,33%. Hal ini menunjukkan bahwa optimasi korpus paralel meningkatkan akurasi terjemahan sebesar 5,55%. Meskipun tidak terlalu signifikan, hasil penelitian menunjukkan bahwa optimasi korpus paralel cukup efektif jika digunakan pada MPS bahasa Indonesia ke bahasa Melayu dan bahasa Jawa. Sebagai perbandingan, penelitian sebelumnya untuk bahasa Indonesia ke bahasa Inggris meningkatkan akurasi sebesar 7,84% [8].

Beberapa contoh hasil terjemahan bahasa Indonesia-Melayu dengan menggunakan *baseline* sebagai perbandingan dan terjemahan mesin dengan menggunakan korpus hasil optimasi disajikan pada Tabel III.

Beberapa contoh hasil terjemahan memperlihatkan bahwa sistem penerjemah dengan korpus hasil optimasi berhasil memperbaiki kesalahan terjemahan dari sistem *baseline*, Pada kalimat pertama, frasa "*buah durian jatuh*" diterjemahkan menjadi "*jatok buah durian*". Kesalahan ini dapat diperbaiki oleh mesin yang menggunakan korpus hasil optimasi menjadi "*buah durian jatok*". Pada kalimat kedua, frasa "*, akhirnya*" tidak muncul pada sistem *baseline* sehingga terjemahannya menjadi kurang tepat. Sistem yang menggunakan korpus hasil optimasi dapat mengatasi kekurangan ini. Pada kalimat ketiga, frasa "*itu baru namanya pemimpin*" diterjemahkan menjadi "*itu namenye pemimpén*" pada sistem *baseline*, sedangkan sistem korpus hasil optimasi menerjemahkannya sebagai "*itu barok namenye pemimpén*" seperti pada kalimat referensi, Frasa "*sudah memberi saran*" pada kalimat keempat yang seharusnya diterjemahkan menjadi "*udah kasik saran*" diterjemahkan menjadi "*kasik udah saran*" oleh sistem *baseline*. Kesalahan ini dapat diperbaiki oleh sistem yang menggunakan korpus hasil optimasi. Pada kalimat kelima, frasa "*coba kalian bayangkan*" diterjemahkan menjadi "*cobe kalian*" oleh sistem *baseline*. Kesalahan tersebut diperbaiki oleh sistem yang menggunakan korpus hasil optimasi dengan terjemahan "*cobe kitak bayangkan*". Dari hasil eksperimen tersebut, terbukti bahwa mesin kedua dapat memperbaiki kesalahan-kesalahan yang terjadi pada sistem pertama, sehingga dapat disimpulkan bahwa optimasi korpus dengan mengeliminasi kalimat-kalimat dengan nilai BLEU kurang dari 10% dapat meningkatkan akurasi mesin penerjemah bahasa Indonesia-Melayu,

TABEL III
PERBANDINGAN HASIL TERJEMAHAN INDONESIA-MELAYU DENGAN MENGGUNAKAN KORPUS HASIL OPTIMASI DAN *BASELINE*

No		Kalimat
1	Masukan	***kata orang, ibarat buah durian jatuh dari langit***
	Ref	*kate orang, ibarat buah durian jatok dari langét*
	BL	*kate orang, ibarat jatok buah durian dari langét (BLEU= 41,54 %)*
	KHO	*kate orang , ibarat buah durian jatok dari langét (BLEU= 100,00 %)*
2	Masukan	***angan-angannya, akhirnya datang juga***
	Ref	*angan-angan die, ahérnye tibe ugak*
	BL	*angan-angan die tibe ugak (BLEU = 0,00 %)*
	KHO	*angan-angan die, ahérnye tibe ugak (BLEU = 100,00 %)*
3	Masukan	***janganlah kamu berbicara seperti itu karena tidak baik didengar orang lain***
	Ref	*janganlah awak ngomong macam tu sebab tadak baék didengar orang laén*
	BL	*janganlah awak bicara macam tu sebab baék tak didengar orang laén (BLEU = 0,00 %)*
	KHO	*janganlah awak bicara macam tu sebab tadak baék didengar orang laén (BLEU = 74,19 %)*
4	Masukan	***itu baru namanya pemimpin, saya dengar dia pun sudah memberi saran ke SBY untuk segera bertindak***
	Ref	*itu barok namenye pemimpén , saye dengar die pon udah kasik saran ke SBY untok segere betindak*
	BL	*itu namenye pemimpén , saye dengar die pon kasik udah saran ke SBY untok segere betindak (BLEU = 66,68%)*
	KHO	*itu barok namenye pemimpén , saye dengar die pon udah kasik saran ke SBY untok segere betindak (BLEU = 100,00 %)*
5	Masukan	***coba kalian bayangkan betapa besarnya pelabuhan rotterdam itu***
	Ref	*cobe kitak bayangkan betape besaknye pelabohan rotterdam tu*
	BL	*cobe kalian betape besaknye pelabohan rotterdam tu (BLEU = 55,78 %)*
	KHO	*cobe kitak bayangkan betape besaknye pelabohan rotterdam tu (BLEU = 100,00 %)*

Catatan : BL = *Baseline* ; KHO = Korpus Hasil Optimasi

Terjadinya perbedaan nilai akurasi masing-masing mesin penerjemah disebabkan oleh kuantitas dan kualitas pada masing-masing korpus yang digunakan, begitu pula perbedaan persentase kenaikan masing-masing sistem. Kuantitas korpus yang dimaksud adalah jumlah kalimat dengan kosakata bervariasi yang terdapat pada korpus. Adapun kualitas korpus yang dimaksud adalah tingkat kebenaran hasil terjemahan yang dilakukan secara manual pada korpus paralel.

IV. KESIMPULAN

Dari eksperimen yang dilakukan terhadap MPS, diperoleh hasil bahwa optimasi korpus paralel dapat meningkatkan tingkat akurasi terjemahan mesin penerjemah bahasa Indonesia-Melayu sebesar 6,97% dan bahasa Indonesia Jawa sebesar 5,55%, dibandingkan dengan sistem *baseline* masing-masing bahasa. Jadi dapat disimpulkan bahwa penggunaan korpus hasil optimasi dapat direkomendasikan untuk digunakan pada mesin penerjemahan dari bahasa Indonesia ke bahasa Melayu dan bahasa Indonesia ke bahasa Jawa.

Untuk penelitian-penelitian selanjutnya, perlu dilakukan eksperimen lanjutan untuk melihat seberapa jauh optimasi korpus paralel dapat meningkatkan akurasi terjemahan bahasa Indonesia ke bahasa-bahasa daerah lainnya.

REFERENSI

[1] Nugroho, R.A., Adji, T.B. & Hantono, B.S., "Penerjemahan Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Berbasis Frasa", *Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015)*, 2015, hal. 51.

[2] Apriani, T., "Pengaruh Kuantitas Korpus Terhadap Akurasi Mesin Penerjemah Statistik Bahasa Bugis Wajo ke Bahasa Indonesia", *Jurnal Sistem dan Teknologi Informasi (JustIN)*, Vol. 1, No. 1, hal. 1-6, 2016.

[3] Mandira, S., Sujaini, H. & Putra, A.B., "Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik", *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, Vol. 2, No.1, hal. 1-5, 2016.

[4] Jarob, Y., Sujaini, H. & Safriadi, N., "Uji Akurasi Penerjemahan Bahasa Indonesia–Dayak Taman Dengan Penandaan Kata Dasar Dan Imbuhan", *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, Vol. 2, No. 2, hal. 78-83, 2016.

[5] Hasbiansyah, M., "Tuning for Quality untuk Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia-Bahasa Dayak Kanayatn", *Jurnal Sistem dan Teknologi Informasi (JustIN)*, Vol. 1, No. 1, hal. 1-5, 2016.

[6] Yohanes, B.W., Robert, T., dan Nugroho, S., "Sistem Penerjemah Bahasa Jawa-Aksara Jawa Berbasis *Finite State Automata*", Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), Vol. 6, No. 2, hal. 127-132, Mei 2017.

[7] Sujaini, H., "Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke Bahasa Melayu Pontianak", *Jurnal Teknik Elektro (ELKHA)*, Vol. 6, No. 2, hal. 38-44, Oktober 2014.

[8] Sujaini, H. & Bijaksana, A., "Strategi Memperbaiki Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik", *Seminar Nasional Teknologi Informasi XI Tahun 2014*, Jakarta, Desember 2014.

[9] Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., dan Mercer, R.L., "The mathematics of statistical machine translation", *Computational Linguistics*, Vol. 19, No. 2, hal. 263-313, 1993.

[10] Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D., & Yamada, K., "Translation with Scarce Bilingual Resources", *Journal Machine Translation*, Vol. 17, No. 1, hal. 1-17, 2002.

[11] Koehn, P., Och, F.J., dan Marcu, D., "Statistical Phrase Based Translation", *Proceedings of the Joint Conference on Human Language Technologies and the An-nual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2003, hal. 48-54.

[12] Zens, R., Och, F.J., & Ney, H., "Phrase-based Statistical Machine Translation", *Proceedings of the German Conference on Artificial Intelligence (KI 2002)*, 2002, hal. 18-32.

[13] Och, F.J. & Weber, H, "Improving Statistical Natural Language Translation With Categories And Rules", *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL)*, 1998, hal. 985-989.

[14] Och, F.J., Tillmann, C., & Ney, H., "Improved Alignment Models For Statistical Machine Translation", *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, 1999, hal. 20–28.

[15] Och, F.J. & Ney, H., "The Alignment Template Approach to Statistical Machine Translation", *Journal Computational Linguistics*, Vol. 30, No. 4, hal. 417–449, 2004.

[16] Venugopal, A., Vogel, S., & Waibel, A., "Effective Phrase Translation Extraction From Alignment Models", Hinrichs, E, and Roth, D,, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, hal. 319–326.

[17] Wang, Y.Y. & Waibel, A., "Modeling With Structures In Statistical Machine Translation", *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL)*, 1998, hal. 1357-1363.

[18] Watanabe, T., Sumita, E., & Okuno, H.G., "Chunk-Based Statistical Translation", *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 2003, Vol. 1, hal. 303-310.

[19] Marcu, D., "Towards A Unified Approach To Memory And Statistical-Based Machine Translation", *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2001, hal. 378-385.

[20] Och, F.J. & Ney, H., "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation", *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2002, hal. 295-302.

[21] Koehn, P., *Statistical Machine Translation*, New York, USA: Cambridge University Press, 2010.

[22] Yıldız, E., Tantuğ, A.C., & Diri, B., "The Effect of Parallel Corpus Quality vs Size in English-to-Turkish SMT", *Sixth International Conference on Web services & Semantic Technology (WeST 2014)*, 2014, hal. 21-30.

[23] Kaalep, H.J. & Veskis, K., "Comparing Parallel Corpora and Evaluating their Quality", *Proceedings of MT Summit XI*, 1997, hal. 275-279.

[24] Maheshwar, S. & Sharma, H., "Improvements in Corpus Quality for Statistical Machine Translation", *IJSRD - International Journal for Scientific Research & Development*, Vol. 2, No, 5, hal. 2321-0613, 2014.

[25] Papineni, K., Roukos, S., Ward, T., dan Zhu, W.-J., "BLEU: A Method For Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, Pennsylvania, 2002, hal. 311-318.