JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989
Vol. 7, No. 1, Januari 2019

Pengaruh Kuantitas Korpus Monolingual Terhadap Akurasi Mesin Penerjemah Statistik

Mirda Wahyuni[#1], Herry Sujaini[#2], Hafiz Muhardi[#3]

[#]*Program Studi Informatika Universitas Tanjungpura*

Jl. Prof. Dr. H. Hadari Nawawi, Kota Pontianak, Kalimantan Barat, 78115

[1]mirdaw13@gmail.com

[2]hs@untan.ac.id

[3]hafiz.muhardi@informatika.untan.ac.id

Abstrak*— Untuk mencapai hasil terjemahan yang optimal mesin penerjemah statistik membutuhkan korpus paralel dalam jumlah yang besar dimana korpus tersebut berisi salinan teks bahasa sumber dan bahasa target yang sejajar. Ketersediaan korpus paralel menjadi salah satu permasalahan karena sumber yang menyediakan dokumen korpus paralel sulit ditemukan. Tidak seperti data paralel, korpus monolingual yang berisi teks hanya dalam satu bahasa dapat mempermudah pembuatan korpus (terutama bahasa target), karena dokumen teks monolingual tersedia secara luas sehingga tidak diperlukan usaha lebih untuk menerjemahkan teks korpus bilingual. Tujuan yang ingin dicapai dalam penelitian ini adalah untuk mengetahui seberapa besar pengaruh kuantitas korpus monolingual terhadap nilai akurasi hasil terjemahan pada mesin penerjemah statistik Bahasa Inggris ke Bahasa Indonesia. Pengujian otomatis menggunakan BLEU dilakukan secara bertahap terhadap 2000 kalimat uji dengan menambahkan korpus monolingual bahasa target dengan jumlah yang sama pada setiap mesinnya yaitu sebanyak 6000 hingga mencapai jumlah 60000 kalimat dan didapatkan peningkatan akurasi sebesar 10,13%. Pengujian manual dilakukan oleh seorang ahli Bahasa Inggris dengan korpus uji sebanyak 100 kalimat dengan peningkatan akurasi sebesar 10,07%. Penggunaan korpus monolingual dapat mempermudah penyediaan sumber data pada mesin penerjemah statistik namun karena peningkatan akurasinya yang terbilang cukup kecil maka dibutuhkan jumlah korpus yang sangat besar sehingga penambahan korpus monolingual ini kurang efisien untuk meningkatkan akurasi terjemahan di atas 30%.

***Kata kunci*— mesin penerjemah statistik, BLEU, korpus paralel, korpus monolingual**

I. PENDAHULUAN

Hasil terjemahan yang optimal dapat diperoleh dengan menggunakan konsep penerjemahan statistik [1]. Mesin penerjemah statistik adalah salah satu jenis mesin penerjemah dimana hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel. Sumber data utama yang digunakan dalam mesin penerjemah statistik adalah korpus paralel dan korpus monolingual [2]. Korpus paralel merupakan dua dokumen teks yang saling berhubungan dimana dokumen teks pertama berisi kumpulan kalimat sumber dan dokumen teks kedua berisi kumpulan kalimat terjemahannya, sedangkan korpus monolingual adalah kumpulan teks korpus yang hanya terdiri dari satu bahasa saja.

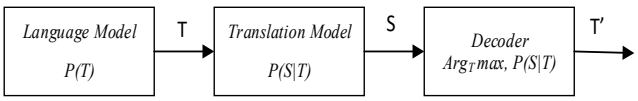
Model mesin penerjemah statistik biasanya membutuhkan sumber data korpus paralel yang cukup besar dimana korpus tersebut berisi salinan teks bahasa sumber dan bahasa target. Tidak seperti data paralel, korpus monolingual yang berisi teks hanya dalam satu bahasa (terutama bahasa target) dapat mempermudah pembuatan korpus, karena dokumen teks monolingual tersedia secara luas sehingga tidak diperlukan usaha lebih untuk menerjemahkan teks korpus bilingual. Selain itu, aksesibilitas dari korpus monolingual khusus yang memiliki beberapa sub-korpus berbeda yang mengandung sejumlah besar kumpulan teks alami dalam berbagai topik dapat memperbaiki kinerja sebuah mesin penerjemah ke tingkat yang lebih tinggi [3].

Berdasarkan uraian tersebut, maka akan dilakukan penelitian untuk mengetahui seberapa besar pengaruh kuantitas korpus monolingual terhadap akurasi hasil terjemahan pada MPS Bahasa Inggris – Bahasa Indonesia.

II. TINJAUAN PUSTAKA

A. Mesin Penerjemah Statistik

Mesin penerjemah statistik merupakan salah satu jenis mesin penerjemah dengan menggunakan pendekatan statistik [4]. Pendekatan statistik yang digunakan adalah konsep probabilitas. Setiap pasangan kalimat (S,T) akan diberikan sebuah P(T|S) yang diinterpretasikan sebagai distribusi probabilitas dimana sebuah penerjemah akan menghasilkan T dalam bahasa sasaran ketika diberikan S dalam bahasa sumber[5].

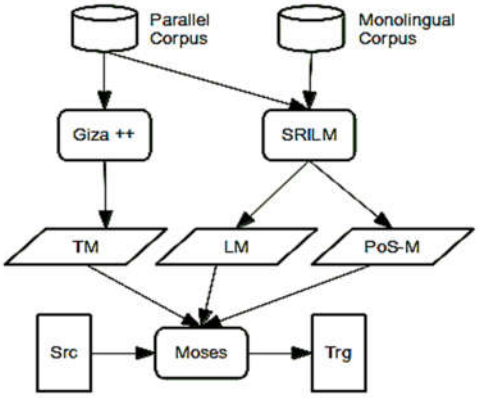


Gambar 1. Komponen Mesin Penerjemah Statistik [6]

Submitted 23-08-2018; Revised 15-10-2018; Accepted 17-10-2018

Pada Gambar 1, *language mode*l menghasilkan kalimat bahasa T, *translation model* mengirimkan kalimat bahasa T sebagai kalimat bahasa S. Sedangkan d*ecoder* mencari kalimat bahasa T' yang paling mungkin yang telah menimbulkan kalimat S.

Secara umum, arsitektur mesin penerjemah statistik Moses ditunjukkan pada Gambar 2.



Gambar 2. Arsitektur Mesin Penerjemah Statistik [2]

Sumber data utama yang dipergunakan adalah parallel corpus dan monolingual corpus. Proses training terhadap parallel corpus menggunakan GIZA++ menghasilkan *translation model* (TM). Proses training terhadap Bahasa target pada parallel corpus ditambah dengan monolingual corpus bahasa target menggunakan SRILM menghasilkan *language model* (LM), sedangkan PoS model (PoS-M) dihasilkan dari bahasa target pada parallel corpus yang setiap katanya sudah ditandai dengan PoS. TM, LM dan PoS-M digunakan untuk menghasilkan *decoder* Moses. Selanjutnya Moses digunakan sebagai mesin penerjemah untuk menghasilkan bahasa target dari input kalimat dalam bahasa sumber.

B. Korpus

Korpus didefinisikan sebagai koleksi atau sekumpulan contoh teks tulis atau lisan dalam bentuk data yang dapat dibaca dengan menggunakan seperangkat mesin dan dapat diberi catatan berupa berbagai bentuk informasi linguistik [7]. Untuk memperbaiki korpus dapat dilakukan dengan memfilter kalimat-kalimat yang berkualitas dari sebuah korpus parallel [8], menambah kuantitas kalimat pada korpus [9] [10], atau perbaikan proses cleaning [11].

Korpus monolingual adalah kumpulan teks korpus yang hanya terdiri dalam satu bahasa saja, biasanya disimpan dan diproses secara elektronik. Korpus monolingual merupakan salah satu komponen penting dalam sistem mesin penerjemah statistik untuk membangun *Language Model* dalam menetapkan probabilitas untuk menargetkan urutan kata. Sedangkan korpus paralel adalah dua atau lebih korpus yang sama dalam bahasa yang berbeda. Masing-masing korpus memuat teks yang telah diterjemahkan dari satu bahasa ke bahasa lain. Korpus ini dapat digunakan penerjemah untuk melihat adanya persamaan ekspresi di masing-masing bahasa atau melihat perbedaan yang ada di antara dua Bahasa.

C. Language Model (LM)

Language model adalah sumber pengetahuan yang penting dalam mesin penerjemah statistik. Dalam *language model* statistik, bagian-bagian yang merupakan elemen kunci adalah probabilitas dari rangkaian-rangkaian kata yang dituliskan sebagai P(w, w ,…,wn) atau P(w, n). *Language model* menetapkan probabilitas P($w1$, n) ke serangkaian n kata dengan means sebuah distribusi probabilitas. Rangkaian-rangkaian tersebut bisa berupa frase-frase atau kalimat-kalimat dan probabilitasnya dapat diperkirakan dari korpus dokumen-dokumen yang besar. Salah satu contoh pendekatan *language model* adalah n-gram model. Model bahasa n-gram merupakan jenis probalilistik *language model* untuk memprediksi item berikutnya dalam urutan tersebut dalam bentuk (n-1) [12].

Berikut merupakan contoh model bahasa n-gram,yaitu :

- Unigram (1-gram): P(w_1), $P(w_2)$…P(w_n)
- Bigram (2-gram): P(w_1), $P(w_2|w_1)$, … $P(w_n|w_{(n-1)})$
- Trigram (3-gram): P($w_{1,n}$) = $P(w_1), P(w_2|w_1)$, $P(w_3|w_{1,2})$…P($w_n|w_{n-2,n-1}$)

D. Translation Model (TM)

Translation model digunakan untuk memasangkan teks input dalam bahasa sumber dengan teks output dalam bahasa sasaran. Dalam mesin penerjemah statistik terdapat dua model penerjemahan, yaitu *word-based translation model* (model translasi berbasis kata) dan *pharase-based translation model* (model translasi berbasis frase) [5].

E. Automatic Evaluation (BLEU)

BLEU *(Bilingual Evaluation Understudy)* adalah sebuah algoritma yang berfungsi untuk mengevaluasi kualitas dari sebuah mesin terjemahan yang telah diterjemahkan oleh mesin dari satu bahasa alami ke bahasa lain. Ide utama dibalik ini adalah "semakin dekat terjemahan sebuah mesin dengan terjemahan manusia, maka akan semakin baik" [13]. Nilai BLEU didapat dari hasil perkalian antara brevity penalty dengan rata-rata geometri dari *modified precision score*. Semakin tinggi nilai BLEU, maka semakin akurat dengan rujukan. Nilai dari BLEU berada pada rentang 0 sampai 1. Suatu terjemahan akan mencapai nilai 1 jika terjemahan tersebut identik dengan terjemahan rujukan. Sangat semakin banyak terjemahan rujukan per kalimatnya, maka akan semakin tinggi nilainya. Untuk menghasilkan nilai BLEU yang tinggi, panjang kalimat hasil terjemahan harus mendekati panjang dari kalimat referensi dan kalimat hasil terjemahan harus memiliki kata dan urutan yang sama dengan kalimat referensi. Rumus BLEU dapat dilihat pada persamaan berikut [14].

$$BP_{BLEU} = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad (1)$$

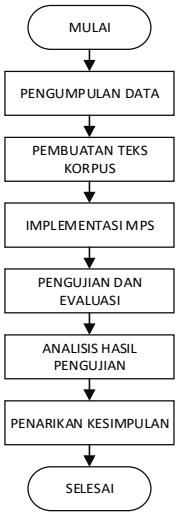
$$P_n = \frac{\sum_{Ce\ corpus\ n-gram \in C} \sum count_{clip}(n-gram)}{\sum_{Ce\ corpus\ n-gram \in C} \sum count\ (n-gram)} \quad (2)$$

$$\text{BLEU} = BP_{BLEU} \cdot e^{\sum_{n=1}^{N} w_n \log p_n} \quad (3)$$

Keterangan:
BP = brevity penalty
c = jumlah kata dari hasil terjemahan otomatis
r = jumlah kata rujukan
n = modified precission score
wn = 1/N (standar nilai N untuk BLEU adalah 4)
pn = jumlah *n-gram* hasil terjemahan yang sesuai dengan rujukan dibagi jumlah *n-gram* hasil terjemahan.

III. METODOLOGI PENELITIAN

Metodologi penelitian yang dilakukan dijelaskan pada Gambar 3.



Gambar 3. Diagram Alir Penelitian

A. Pengumpulan Data

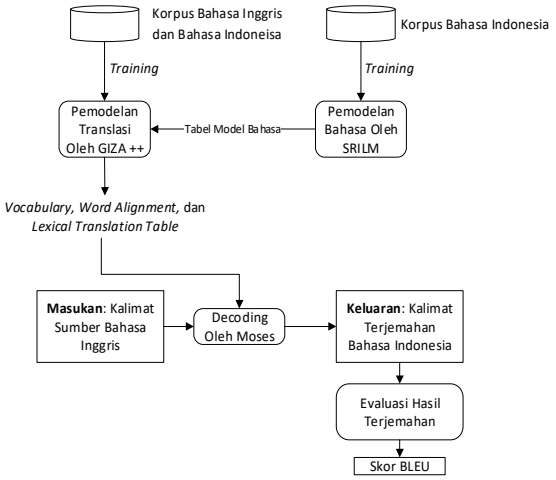
Data yang digunakan dalam penelitian ini berupa dokumen teks Bahasa Inggris dan Bahasa Indonesia yang bersumber dari situs berita dua bahasa (http://berita2bahasa.com/) dan sumber lainnya yang kemudian akan diolah menjadi teks korpus.

B. Pembuatan Teks Korpus

Dokumen teks yang telah dikumpulkan selanjutnya dibuat menjadi korpus paralel dan monolingual, yang terdiri dari 6000 pasang teks korpus paralel Bahasa Inggris – Bahasa Indonesia dan 54000 kalimat korpus monolingual Bahasa Indonesia. Korpus tersebut kemudian disimpan dengan format .en untuk korpus bahasa Inggris dan .id untuk korpus bahasa Indonesia.

C. Implementasi Mesin Penerjemah Statistik

Arsitektur sistem mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia diperlihatkan pada Gambar 4.



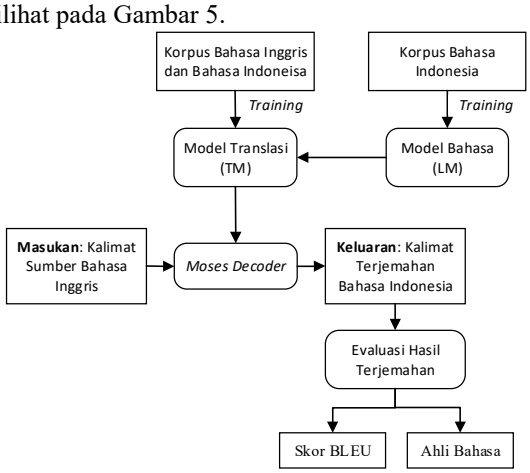
Gambar 4. Arsitektur Mesin Penerjemah Statistik

Gambar 4 merupakan perancangan arsitektur sistem mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia yang terdiri dari beberapa tahapan, yaitu tahap awal persiapan korpus paralel, pemodelan bahasa, pemodelan translasi, proses *decoding*, dan tahap evaluasi hasil terjemahan.

D. Pengujian dan Evaluasi Hasil Terjemahan

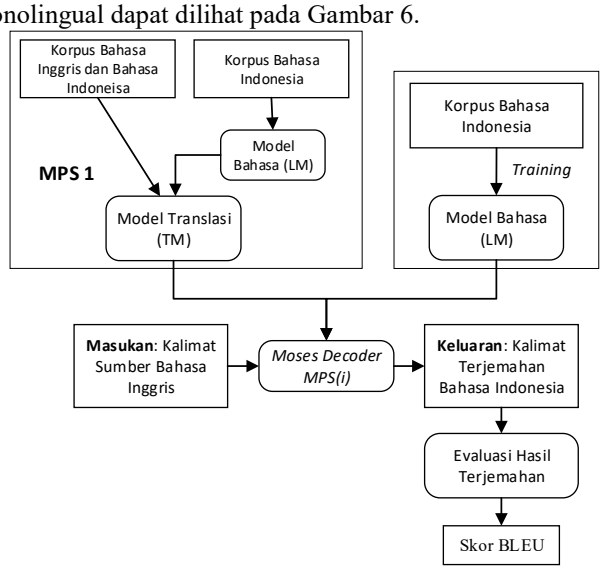
Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi antara mesin penerjemah statistik sebelum ditambah korpus monolingual dengan mesin penerjemah statistik setelah ditambah dengan korpus monolingual. Pengujian dan evaluasi ini akan dilakukan dengan dua cara yaitu secara manual oleh ahli Bahasa Inggris dan secara otomatis menggunakan BLEU.

Pengujian mesin penerjemah statistik tahap pertama dilakukan dengan membangun mesin penerjemah dengan menggunakan pasangan korpus paralel Bahasa Inggris dan Bahasa Indonesia dengan masing-masing jumlah korpus sebanyak 6000 kalimat yang kemudian akan diuji akurasinya dengan menggunakan BLEU dan pengujian manual dengan ahli bahasa. Adapun perancangan mesin penerjemah statistik dapat dilihat pada Gambar 5.



Gambar 5. Perancangan Mesin Penerjemah Statistik

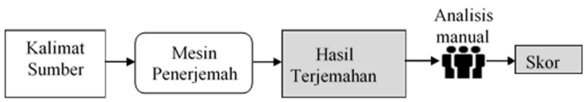
Tahap kedua dilakukan dengan membangun mesin penerjemah baru dengan sumber data berupa korpus paralel dan *translation model* (TM) yang diambil dari mesin pertama sebagai baseline dan kemudian ditambah dengan korpus monolingual bahasa target yaitu bahasa Indonesia. Penambahan korpus monolingual akan dilakukan secara bertahap dengan jumlah kelipatan korpus yang sama pada setiap mesinnya yaitu sebanyak 6000 kalimat. Setiap penambahan korpus monolingual tersebut akan dibangun mesin penerjemah baru dan akan dilakukan pengujian secara otomatis dengan menggunakan BLEU. Perancangan mesin penerjemah statistik dengan menambahkan korpus monolingual dapat dilihat pada Gambar 6.



Gambar 6. Perancangan Mesin Penerjemah Statistik dengan Menambahkan Korpus Monolingual pada Bahasa Target

Pada pengujian dengan menggunakan BLEU, akan digunakan 2000 pasang korpus uji yang kemudian akan diterjemahkan secara otomatis oleh mesin translasi dan akan menghasilkan *output* berupa dokumen korpus dalam bahasa target.

Pengujian manual akan dilakukan oleh ahli Bahasa Inggris dengan menggunakan 100 kalimat sampel Bahasa Inggris dari 2000 korpus uji otomatis dengan BLEU sebagai bahasa sumber dan Bahasa Indonesia sebagai bahasa target. perancangan proses penilaian secara manual dapat di lihat pada Gambar 7.



Gambar 7. Proses Evaluasi Secara Manual

E. Analisis Hasil Pengujian

Analisis hasil pengujian dilakukan untuk mengetahui karakteristik mesin penerjemah statistik dan mengidentifikasi apakah sudah sesuai dengan kebutuhan berdasarkan pada hasil uji akurasi mesin penerjemah statistik Bahasa Inggris – Bahasa Indonesia.

IV. HASIL DAN ANALISIS

A. Implementasi Mesin Penerjemah Statistik Bahasa Inggris – Bahasa Indonesia

Tahapan implementasi mesin penerjemah statistik bahasa Inggris - bahasa Indonesia terlebih dahulu korpus teks paralel yang telah dibuat dilakukan proses *cleaning*, *tokenizing*, dan *lowercase*. Proses cleaning adalah proses pencarian dan perbaikan (penghapusan) kata atau kalimat yang salah ataupun tidak sesuai [15]. Proses cleaning yang disediakan oleh mosesdecoder hanya menghapus kalimat yang terlalu panjang, serta yang kalimat kosong [16]. Fungsi dari proses *cleaning* adalah untuk menyaring data, seperti menghilangkan kata yang terlalu panjang sesuai batas yang ditentukan, menghilangkan spasi ganda, dan menghapus tanda baca titik di akhir kalimat. Sedangkan proses tokenisasi berfungsi untuk memotong *string input* berdasarkan tiap kata yang menyusunnya [17] dan menyisipkan spasi antara kata dan tanda baca. Sedangkan *lowercase* merupakan proses untuk mengubah huruf kapital yang terdapat di dalam korpus menjadi huruf kecil *(case folding)*.

B. Implementasi SRILM untuk Pemodelan Bahasa

Pemodelan bahasa dilakukan untuk mendapatkan model bahasa dari bahasa target yaitu Bahasa Indonesia. Model bahasa digunakan sebagai sumber pengetahuan berbasis teks dengan nilai-nilai probabilistik. Model bahasa yang digunakan dalam penelitian ini yaitu n-gram data. Model Bahasa dibangun dengan tools SRILM dan menghasilkan *output* dengan format file *.lm. Tabel model bahasa yang dihasilkan oleh SRILM dapat dilihat pada Gambar 8.

```
\data\
ngram 1=14143
ngram 2=79745
ngram 3=8260

\1-grams:
-3.017467       air             -0.3596329
-3.37131        bekerja         -0.4030444
-3.855984       belanja         -0.2624049
----------------------------------------------------------------
\2-grams:
-0.760619       area bekas      -0.2126758
-0.3357921      babak pertama   -0.01780334
-1.25746        beberapa hari   -0.1428827
----------------------------------------------------------------
\3-grams:
-1.218022       dan akan segera
-1.118035       pada akhir pekan
-0.8200825      pasokan air bersih
```

Gambar 8. Tabel Model Bahasa dengan Bahasa Indonesia sebagai Bahasa Target

C. Implementasi Giza++ Untuk Pemodelan Translasi

Pemodelan translasi digunakan untuk memasangkan teks *input* dalam bahasa sumber dengan teks *output* dalam Bahasa target. Model translasi dibangun dengan tools Giza++. Proses pemodelan translasi oleh Giza++ menghasilkan dokumen *vocabulary corpus, word alignment* dan tabel model translasi.

Dokumen vocabulary corpus yang dihasilkan dapat dilihat pada Gambar 9 dan Gambar 10 berikut.

1	UNK	0
2	and	3110
3	a	2237
4	is	1105
5	for	1008

Gambar 9. Dokumen *Vocabulary Corpus* Bahasa Inggris

1	UNK	0
2	dan	3233
3	di	3200
4	yang	2740
5	pada	1331

Gambar 10. Dokumen *Vocabulary Corpus* Bahasa Indonesia

Angka 1 sampai 10 pada dokumen *vocabulary corpus* merupakan *uniq id* untuk setiap data token, sedangkan angka disebelah kanan token menunjukkan frekuensi kemunculan.

Pada proses *word alignment* akan dihasilkan dokumen korpus *alignment* yaitu kalimat bahasa target dipetakan pada kalimat bahasa sumber. Dokumen alignment yang dihasilkan dapat dilihat pada Gambar 11.

```
# Sentence pair (514) source length 12 target length 14 alignment score: 4.93978e-14

according to him , intelligence cooperation with russia would provide many benefits for indonesia

NULL ({ }) menurutnya ({ 1 2 3 }) , ({ 4 }) kerjasama ({ 6 }) intelijen ({ 5 }) dengan ({ 7 }) rusia ({ 8 }) akan ({ 9 }) memberi ({ 10 }) banyak ({ 11 }) manfaat ({ 12 }) bagi ({ 13 }) indonesia ({ 14 })
```

Gambar 11. Dokumen *Alignment* Bahasa Inggris-Bahasa Indonesia

Pada dokumen hasil alignment bahasa Inggris sebagai bahasa sumber dan bahasa Indonesia sebagai bahasa target menghasilkan tiap pasangan kalimat yang disimpan dalam tiga baris. Baris pertama pada Gambar 11 berisi posisi kalimat target (514) dalam korpus, panjang kalimat sumber (12), panjang kalimat target (14), dan skor alignment (4.93978e-14). Baris kedua merupakan Bahasa sumber dan baris ketiga merupakan alignment kalimat bahasa target terhadap kalimat bahasa sumber.

Proses pemodelan translasi oleh Giza++ akan menghasilkan tabel model translasi yang terdiri dari tabel kata yang berisi sekumpulan kata-kata yang telah dipadankan antara bahasa sumber dengan bahasa target yang memiliki nilai probabilitas. Tabel model translasi frasa yang dihasilkan dapat dilihat pada gambar 12.

```
family and friends ||| keluarga dan teman-teman ||| 1 0.305731 1 0.253655
people can see ||| orang-orang dapat melihat ||| 1 0.0605546 1 0.0131089
return to mexico ||| kembali ke meksiko ||| 1 0.01634 1 0.0772756
saturday morning ||| sabtu pagi ||| 1 0.470588 1 0.6
service between ||| layanan antara ||| 1 0.214286 1 0.188259
```

Gambar 12. Tabel Model Translasi

D. Decoding

Proses *decoding* digunakan untuk menemukan teks dalam bahasa target yang memiliki probabilitas paling besar dengan pertimbangan faktor *translation model* dan *language model*. *Tools* yang digunakan untuk proses *decoding* adalah Moses.

Decoder Moses akan menerjemahkan kalimat masukan berupa kalimat sumber (Bahasa Inggris). Selanjutnya kalimat masukan tersebut akan diproses oleh *decoder* moses dan akan menghasilkan kalimat keluaran berupa kalimat hasil terjemahan ke dalam bahasa target (Bahasa Indonesia).

E. Pengujian Hasil Terjemahan Secara Otomatis

Pengujian hasil terjemahan otomatis pertama dilakukan dengan membangun mesin penerjemah dengan menggunakan pasangan korpus paralel Bahasa Inggris dan Bahasa Indonesia dengan masing-masing jumlah korpus sebanyak 6000 kalimat yang kemudian akan diuji akurasinya dengan menggunakan BLEU. Korpus uji yang digunakan pada tahap ini berjumlah 2000 kalimat.

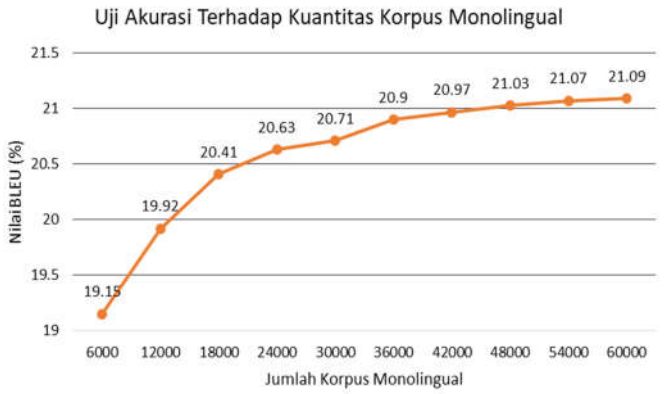
TABEL I
HASIL PENGUJIAN PENERJEMAHAN TERHADAP KUANTITAS KORPUS MONOLINGUAL

Mesin	Jumlah Korpus Monolingual	Nilai BLEU (%)
1	6000	19.15
2	12000	19.92
3	18000	20.41
4	24000	20.63
5	30000	20.71
6	36000	20.90
7	42000	20.97
8	48000	21.03
9	54000	21.07
10	60000	21.09

Grafik hasil pengujian terjemahan terhadap kuantitas korpus monolingual terdapat pada Gambar 13.

Selanjutnya dibangun mesin penerjemah baru dengan sumber data berupa korpus paralel dan translation model (TM) yang diambil dari mesin pertama sebagai baseline dan kemudian ditambah dengan korpus monolingual bahasa target yaitu bahasa Indonesia. Penambahan korpus monolingual akan dilakukan secara bertahap dengan jumlah kelipatan korpus yang sama pada setiap mesinnya yaitu berjumlah 6000 kalimat. Setiap penambahan korpus monolingual tersebut akan dibangun mesin penerjemah baru untuk kemudian dilakukan pengujian output secara otomatis dengan

menggunakan BLEU. Setiap mesin menghasilkan nilai BLEU yang berbeda. Secara umum hasil pengujian terjemahan terhadap kuantitas korpus monolingual diperlihatkan pada Tabel 1.



Gambar 13. Grafik Hasil Pengujian Terjemahan terhadap Kuantitas Korpus Monolingual

Persentase kenaikan atau peningkatan nilai BLEU pada tiap mesin dapat dihitung dengan persamaan berikut.

$$Peningkatan = \frac{Nilai\ BLEU\ mesin_n - Nilai\ BLEU\ mesin_{n-1}}{Nilai\ BLEU\ mesin_{n-1}}\ x\ 100\% \quad (4)$$

$$Peningkatan = \frac{Nilai\ BLEU\ mesin\ 2 - Nilai\ BLEU\ mesin\ 1}{Nilai\ BLEU\ mesin\ 1}\ x\ 100\%$$

$$Peningkatan = \frac{19{,}92 - 19{,}15}{19{,}15}\ x\ 100\%$$
$$= 0{,}041\ x\ 100\%$$
$$= 4{,}1\%$$

TABEL 2
PERSENTASE PENINGKATAN NILAI BLEU PADA TIAP MESIN

Mesin	Peningkatan (%)
2	4,1
3	2,45
4	1,07
5	0,38
6	0,92
7	0,34
8	0,28
9	0,19
10	0,10
Rata-rata	**1,10**

Berdasarkan hasil perhitungan pada Tabel 2, terjadi perubahan nilai akurasi terjemahan pada mesin penerjemah statistik dengan rata-rata peningkatan sebesar 1,10% terhadap penambahan korpus monolingual sebanyak 6000 kalimat.

Peningkatan nilai akurasi dari mesin 1 ke mesin 10 (berdasarkan Tabel 1) dapat dilihat pada perhitungan berikut ini.

$$Peningkatan = \frac{21{,}09 - 19{,}15}{19{,}15}\ x\ 100\%$$
$$= 0{,}1013\ x\ 100\%$$
$$= 10{,}13\%$$

F. Pengujian Hasil Terjemahan Secara Manual

Pengujian manual dilakukan pada hasil terjemahan dari dua mesin penerjemah statistik yang berbeda yaitu mesin pertama dengan jumlah korpus paralel sebanyak 6000 pasang kalimat dan mesin terakhir yang dibangun dengan sumber data berupa korpus paralel dan translation model (TM) yang diambil dari mesin pertama sebagai baseline dan kemudian ditambah dengan korpus monolingual bahasa target sebanyak 54000 kalimat. Penilaian oleh ahli bahasa dilakukan dengan metode skoring dimana pembobotan ini diberikan pada setiap kalimat uji dengan menggunakan sampel kalimat dari korpus uji dengan jumlah 100 kalimat Bahasa Inggris sebagai kalimat bahasa sumber dan Bahasa Indonesia sebagai kalimat Bahasa target. Berikut merupakan skor pembobotan untuk setiap kalimat uji:

5 = Sangat baik
4 = Baik
3 = Cukup baik
2 = Kurang baik
1 = Tidak baik

TABEL 3
HASIL PENILAIAN MANUAL OLEH AHLI BAHASA

Mesin	Bobot					Jumlah Kalimat
	1	2	3	4	5	
6000	0	45	34	19	2	100
60000	0	33	35	25	7	100

Berdasarkan hasil penilaian yang diberikan sesuai dengan pengetahuan dan pemahaman ahli bahasa, maka dilakukan perhitungan akurasi manual dengan persamaan berikut:

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n} \quad (5)$$

dengan: $\bar{x}$ = nilai rata-rata (mean) akurasi terjemahan
$\sum x$ = total skor dari bobot penilaian
n = banyaknya data

Secara umum, hasil perhitungan penilaian manual terjemahan Bahasa Inggris ke Bahasa Indonesia berdasarkan persamaan 5 dapat dilihat pada Tabel 4.

TABEL 4
HASIL PERHITUNGAN PENGUJIAN MANUAL OLEH AHLI BAHASA

Pengujian Secara Manual	Mesin	$\sum x$, n	Nilai Akurasi
	6000	278, 100	2,78
	60000	306, 100	3,06

Dari tabel di atas diperoleh nilai akurasi hasil pengujian manual mesin penerjemah statistik Bahasa Inggris ke Bahasa Indonesia, yaitu 2,78 untuk mesin penerjemah statistik pertama dengan jumlah korpus paralel 6000 pasang kalimat dan 3,06 untuk mesin penerjemah statistik dengan jumlah korpus paralel dan monolingual sebanyak 60000 kalimat.

Berikut merupakan perhitungan untuk perbandingan hasil pengujian manual oleh ahli bahasa dari mesin penerjemah statistik 1 dengan jumlah korpus paralel 6000 dan mesin penerjemah statistik 10 dengan jumlah korpus paralel dan monolingual sebanyak 60000.

$$Peningkatan = \frac{3{,}06 - 2{,}78}{2{,}78}\ x\ 100\%$$
$$= 10{,}07\%$$

G. Analisis Hasil Pengujian

Berikut merupakan analisis terhadap hasil pengujian yang telah dilakukan.

1. Penambahan korpus monolingual secara bertahap dengan kelipatan 6000 kalimat pada tiap mesin dapat mempengaruhi nilai akurasi terjemahan dimana diperoleh peningkatan rata-rata sebesar 1,10% untuk setiap mesin, sedangkan untuk mesin 1 dan 10 terjadi peningkatan sebesar 10,13% pada pengujian otomatis dengan BLEU.
2. Pada pengujian manual yang dilakukan oleh ahli bahasa, diperoleh nilai akurasi sebesar 2,78 untuk mesin 1 (6000 pasang korpus paralel) dan 3,06 untuk mesin 10 (60000 korpus paralel dan monolingual). Dengan demikian, terjadi peningkatan akurasi terjemahan sebesar 10,07%.
3. Perkiraan jumlah korpus monolingual pada mesin penerjemah statistik Bahasa Inggris ke Bahasa Indonesia dapat dihitung berdasarkan fungsi logaritma dimana nilai tersebut diperoleh dari dari grafik hasil uji akurasi mesin penerjemah statistik terhadap kuantitas korpus monolingual.

TABEL 5
PERKIRAAN JUMLAH KORPUS MONOLINGUAL PADA MESIN PENERJEMAH STATISTIK BAHASA INGGRIS KE BAHASA INDONESIA

y (Persentase Akurasi MPS)	***x*** (Perkiraan Jumlah Korpus Yang Diperlukan)
30%	$2{,}17 \times 10^{9}$
40%	$3{,}50 \times 10^{14}$
50%	$5{,}66 \times 10^{19}$
60%	$9{,}14 \times 10^{24}$
70%	$1{,}48 \times 10^{30}$
80%	$2{,}39 \times 10^{35}$
90%	$3{,}85 \times 10^{40}$

V. KESIMPULAN

Berdasarkan hasil analisis dan pengujian, maka kesimpulan yang dapat diambil adalah sebagai berikut.

1. Proses penambahan kuantitas korpus monolingual bahasa target dapat meningkatkan nilai akurasi terjemahan sebesar 10,13% pada pengujian otomatis oleh BLEU. Sedangkan pada pengujian manual oleh ahli Bahasa diperoleh peningkatan sebesar 10,07%.
2. Penggunaan korpus monolingual dapat mempermudah penyediaan sumber data pada mesin penerjemah statistik namun karena peningkatan akurasinya yang terbilang cukup kecil maka dibutuhkan jumlah korpus yang besar sehingga penambahan korpus monolingual ini kurang efisien untuk meningkatkan akurasi terjemahan di atas 30%.

REFERENSI

[1] Apriani, Tri., *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. 168-173, 2016.

[2] Sujaini, Herry., dan Negara, Arif Bijaksana Putra. *Analysis of Extended Word Similarity Clustering based Algorithm on Cognate Language*. Gujarat: ESRSA Publications Pvt. Ltd. 2015.

[3] Miangah, Tayebeh Mosavi., dan Khalafi, Ali Delavar. *Word Sense Disambiguation Using Target Language Corpus in a Machine Translation System*. Iran: Literary and Linguistic Computing, Vol.2, No.2, hal 237-249. 2005.

[4] Hasbiansyah, Muhammad. 2016. *Tuning For Quality Untuk Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia - Bahasa Dayak Kanayatn. Pontianak*, Jurnal Sistem dan Teknologi Informasi (JustIN), Vol. 4, No. 1, hal. 209-213, 2016.

[5] Tanuwijaya, Hansel. *Penerjemahan Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan Word Reordering dan Phrase Reordering*. Jakarta, Jurnal ilmu Komputer dan Informasi Vol 2 No 1, hal. 17-24. 2009.

[6] Manning, Christopher D. dan Schutze, Hinrich. *Foundations of Statistical Natural Language Processing*. London: The MIT Press Cambridge Massachusetts. 2000.

[7] McEnery, Tony dan Wislon, Andrew. *Corpus Linguistics*. Edinburgh: Edinburgh University Press. 1996.

[8] Sujaini, Herry. dan Arif B.P.N. *Strategi Memperbaiki Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik*. Jakarta, Seminar Nasional Teknologi Informasi XI. 2015.

[9] 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.

[10] Maheshwar, S. & Sharma, H., *Improvements in Corpus Quality for Statistical Machine Translation.* IJSRD - International Journal for Scientific Research & Development, Vol. 2, No, 5, hal. 23210613, 2014.

[11] Xu, Hainan and Koehn, Philipp (2017): *Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora.* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[12] Hadi, Ibnu. *Uji Akurasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Melayu Sambas dan Bahasa Melayu Sambas ke Bahasa Indonesia*. Pontianak, Jurnal Sistem dan Teknologi Informasi (JustIN), Vol. 3, No. 1, hal. 127-135. 2014.

[13] Papineni, Kishore; Ruokos, Salim; Ward, Todd; dan Zhu, Wei-Jing. *BLEU: a Methode For Automatic Evaluation of Machine Translation*. USA: IBM TJ Watson Research Center. 2002.

[14] Y. Jarob, H. Sujaini dan N. Safriadi. *Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan.* Jurnal Edukasi dan Penelitian Informatika (JEPIN), Vol. 2 No. 2, 2016.

[15] Devi. Sapna, dan Kalia, Arvind. 2015. *Study of Data Cleaning & Comparison of Data Cleaning Tools*. IJCSMC, Vol. 4, Issue. 3, March 2015.

[16] Koehn, Philipp. *MOSES Statistical Machine Translation User Manual dan Code Guide*. The University of Edinburgh. 2016.

[17] Triawati, Candra. *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia*. Jakarta: IT TELKOM. 2009.