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

Peningkatan Mesin Penerjemah Statistik dengan Menambah Kuantitas Korpus Monolingual (Studi Kasus : Bahasa Indonesia – Sunda)

Robby Darwis[#1], Herry Sujaini[#2], Rudy Dwi Nyoto[#3]

[#]*Program Studi Informatika Universitas Tanjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak 78124
[1]robbygluemantik@gmail.com
[2]hs@untan.ac.id
[3]rudydn@informatika.untan.ac.id

***Abstrak*— Bahasa merupakan alat komunikasi yang dijadikan sarana untuk berinteraksi dengan masyarakat sekitar. Kemampuan akan penguasaan banyak bahasa tentunya akan mempermudah untuk berinteraksi dengan orang lain dari berbagai daerah yang berbeda. Oleh karena itu, diperlukan penerjemah untuk menambah pengetahuan akan berbagai bahasa yang ada. Mesin Penerjemah Statistik *(Statistical Machine Translation)* merupakan sebuah pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel. Korpus paralel adalah pasangan korpus yang berisi kalimat-kalimat dalam suatu bahasa dan terjemahannya. Salah satu fitur yang digunakan untuk meningkatkan akurasi hasil terjemahan adalah dengan fitur Menambah Kuantitas Korpus Monolingual.Tujuan penelitian ini adalah melakukan penggunaan fitur Menambah Kuantitas Korpus Monolingual pada mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda untuk meningkatkan nilai akurasi hasil terjemahan dan mengetahui seberapa besar pengaruh dari penambahan kuantitas korpus monolingual terhadap hasil terjemahan. Pengujian dilakukan dengan membandingkan nilai akurasi hasil terjemahan sebelum dan setelah Menambah Kuantitas Korpus Monolingual. Penelitian menggunakan korpus paralel sebanyak 3000 korpus. Pengujian dilakukan dengan dua cara, yaitu pengujian otomatis menggunakan *Bilingual Evaluation Understudy* (BLEU) dan pengujian oleh ahli bahasa Sunda. Dari hasil penelitian, penggunaan Menambah Kuantitas Korpus Monolingual dapat meningkatkan kualitas terjemahan untuk mesin penerjemah bahasa Indonesia ke bahasa Sunda. Hal itu terlihat dari hasil pengujian dengan menambahkan fitur Menambah Kuantitas Korpus Monolingual terdapat peningkatan nilai BLEU dengan korpus 3400 sebesar 0.32%, 4400 korpus sebesar 0.51%, 5400 korpus sebesar 0,42%, 6400 korpus sebesar 0.51%, 7400 korpus sebesar 0.87%, 8400 korpus sebesar 1.04%, 9400 korpus sebesar 1.79% pada pengujian otomatis dan 19.13% pada pengujian oleh ahli bahasa. Berdasarkan hal tersebut, mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda dengan penggunaan fitur Menambah Kuantitas Korpus Monolingual dapat meningkatkan nilai akurasi hasil terjemahan.**

Kata Kunci: Menambah Kuantitas Korpus Monolingual, mesin penerjemah statistik, korpus paralel, BLEU *score,* Indonesia, Sunda.

I. PENDAHULUAN

Bahasa merupakan alat yang digunakan untuk berkomunikasi. Tidak hanya untuk melakukan komunikasi antara manusia dengan manusia yang lainnya, namun dalam hal ini bahasa juga menjembatani komunikasi antara manusia dengan komputer. Bahasa yang digunakan manusia untuk berkomunikasi dengan komputer dikenal dengan bahasa pemrograman. Untuk mengolah bahasa dari manusia dan computer maka diperlukan sistem Natural Languange Processing (NLP).

Dengan pesatnya perkembangan teknologi, saat ini sedang dikembangkan mesin penerjemah untuk mengatasi masalah penerjemahan bahasa. Mesin penerjemah (MP) merupakan mesin yang dapat melakukan proses penerjemahan dari satu bahasa ke bahasa lainnya secara otomatis. MP memiliki kegunaan praktis karena dapat membantu manusia untuk berkomunikasi satu sama lainnya yang memiliki bahasa yang berbeda [1]. Mesin penerjemah statistik merupakan sebuah pendekatan mesin penerjemah dengan hasil terjemahan dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus teks bilingual (korpus paralel) [2].

Natural Language Processing (NLP) adalah salah satu bidang ilmu komputer, kecerdasan buatan, dan bahasa (linguistik) yang berkaitan dengan interaksi antara komputer dan bahasa alami manusia, seperti bahasa Indonesia atau bahasa Lainnya. Tujuan utama dari studi NLP adalah membuat mesin yang mampu mengerti dan memahami makna bahasa manusia lalu memberikan respon yang sesuai.

Sesuai dengan studi kasus diatas, ada beberapa metode yang dilakukan untuk membuat mesin penerjemah yang mampu menerjemah bahasa Indonesia ke bahasa Sunda. Salah satu cara untuk meningkatkan akurasi mesin penerjemah stastik (MPS) adalah dengan menambah kuantitas korpus monolingual.

Submitted 24-08-2018; Revised 11-10-2018; Accepted 17-10-2018

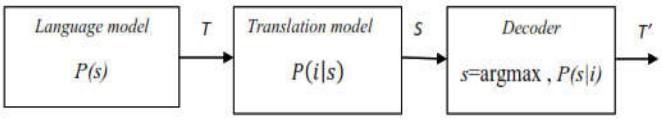
Oleh karena itu, untuk menyelesaikan permasalahan yang ada maka diperlukan penelitian tentang peningkatan akurasi mesin penerjemah statistik tersebut dengan menambah kuantitas korpus monolingual.

Berdasarkan faktor-faktor yang telah dijabarkan, untuk menghindari terjadinya kemerosotan penggunaan bahasa daerah, salah satu caranya adalah dengan mesin penerjemah statistik. Mesin penerjemah statistik (*Statistical Machine Translation*) merupakan sebuah pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel [3]

II. URAIAN PENELITIAN

A. Mesin Penerjemah Statistik

Mesin penerjemah statistik merupakan salah satu jenis mesin penerjemah dengan menggunakan pendekatan statistik. Menurut Christoper D Manning dan Hinrich Schutze, dalam *statistical machine translation* terdapat tiga buah komponen yang terlibat dalam proses penerjemahan kalimat dari suatu bahasa ke bahasa lain, yaitu *language model*, *translation model*, dan *decoder* seperti yang tertera pada Gambar 1 [4].



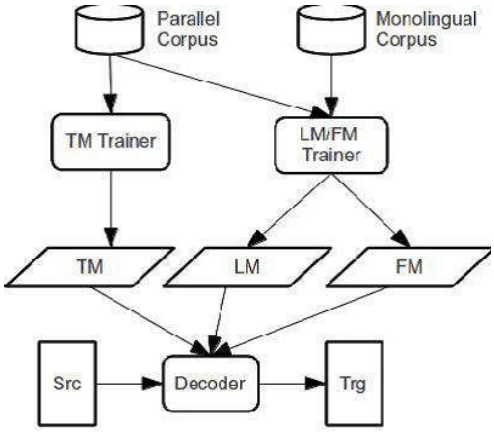
Gambar 1. Komponen mesin penerjemah statistik [4]

Mesin penerjemah statistik merupakan salah satu jenis mesin penerjemah dengan menggunakan pendekatan statistik. 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].

Language model digunakan pada aplikasi *Natural Language Processing* seperti *speech recognition*, *part-of-speech tagging* dan *syntactic parsing*. *Language model* statistik menetapkan probabilitas $P(W_{1,n})$ ke serangkaian *n* kata dengan *means* sebuah distribusi probabilitas.

Translation model merupakan salah satu komponen penting pada *statistical machine translation* dalam proses penerjemahan yang membagi kalimat bahasa asal menjadi barisan frase, menerjmahkan setiap frase ke tujuan, dan *reordering* [6].

Komponen terakhir dari mesin penerjemah statistik adalah *decoder* yang berfungsi untuk mencari teks dalam Bahasa tujuan yang memiliki probabilitas paling besar dengan pertimbangan *translation model* dan *language model* [7]



Gambar 2. Arsitektur mesin pnerjemah statistik Moses [8]

Gambar 2 merupakan arsitektur sistem dari mesin penerjemah statistik Moses. Menurut Herry Sujaini, 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 corpu*s 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 [8].

B. Moses

Moses adalah salah satu Mesin Penerjemah Statistik yang memungkinkan untuk menerjemahkan secara otomatis setiap pasangan bahasa. Moses digunakan untuk melatih model statistik teks terjemahan dari bahasa sumber ke bahasa target. Saat melakukan penerjemahkan bahasa, Moses membutuhkan korpus dalam dua bahasa, bahasa sumber dan bahasa target [9]. Moses dirilis di bawah lisensi LGPL (Lesser General Public License) dan tersedia sebagai kode sumber dan binari untuk Windows dan Linux. Perkembangannya didukung oleh proyek EuroMatrix, dengan pendanaan oleh European Commission [10].

C. 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 [11]. Korpus dapat diklasifikasikan ke dalam delapan jenis, yaitu korpus khusus (*specialised corpus*), korpus umum (*general corpus*), korpus komparatif (*comparable corpus*), korpus paralel (*parallel corpus*), korpus pemelajar (*learner corpus*), korpus pedagogis (*pedagogic corpus*), korpus historis atau diakronis (*historical or diachronic corpus*), dan korpus monitor (*monitor corpus*) [12]. Berdasarkan jenis korpus tersebut, untuk penelitian ini penulis akan fokus pada korpus paralel.

D. Definisi Penerjemahan

Dalam Kamus Besar Bahasa Indonesia (KBBI) kata "terjemah/menerjemahkan" merupakan menyalin (memindahakan) suatu bahasa ke bahasa lain atau mengalihbahasakan. Selain itu, penerjemahan adalah kegiatan mengalihkan secara tertulis pesan dari teks suatu bahasa (misalnya bahasa Inggris) ke dalam teks bahasa lain (misalnya bahasa Indonesia) [13]. Penerjemahan adalah pengalihan pikiran atau gagasan dari suatu bahasa sumber ke dalam bahasa yang lain. Penerjemahan adalah mengubah teks bahasa sumber ke dalam teks bahasa sasaran dengan mempertimbangkan makna kedua bahasa sehingga diusahakan semirip-miripnya, yang tak kalah pentingnya adalah terjemahan harus mengikuti kaidah-kaidah yang berlaku dalam bahasa sasaran [14].

E. Proses Penerjemahan

Proses penerjemahan terdiri dari 3 tahap yaitu *analysis, transfer* dan *restructuring*. Dalam proses *analysis*, penerjemah menganalisis isi pesan bahasa sumber berdasarkan gramatika dan makna. Pada tahap ini kalimat-kalimat bahasa sumber dipecah-pecah menjadi satuan-satuan gramatikal berstruktur kalimat-kalimat dasar, kata-kata dan frase-frase untuk menangkap makna yang ada dengan teknik analisis komponen. Tahap kedua*, transfer*, yaitu proses pengalihan materi-materi yang telah dianalisis dari bahasa sumber ke dalam bahasa sasaran. Tahap terakhir yaitu *restructuring*, bahwa penerjemah menyusun materi-materi yang telah dialihkan dan bertujuan untuk membuat pesan yang secara keseluruhan dapat diterima [15].

F. Automatic Evaluation

Sistem evaluasi otomatis yang populer saat ini adalah BLEU (*Bilingual Evaluation Understudy*). BLEU adalah sebuah algoritma yang berfungsi untuk mengevaluasi kualitas dari sebuah hasil terjemahan yang telah diterjemahkan oleh mesin dari satu bahasa alami ke bahasa lain. BLEU mengukur *modified n-gram precission score* antara hasil terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty.*

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. Sangat penting untuk diketahui bahwa semakin banyak terjemahan rujukan per kalimatnya, maka akan semakin tinggi nilainya. Untuk menghasikan 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 sebagai berikut [16]:

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

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

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

Keterangan:

BP = *brevity penalty*

c = jumlah kata dari hasil terjemahan otomatis

r = jumlah kata rujukan

P_n = *modified precission score*

w_n = 1/N (standar nilai N untuk BLEU adalah 4)

p_n = jumlah *n-gram* hasil terjemahan yang sesuai dengan rujukan dibagi jumlah n-gram hasil terjemahan

III. HASIL PENELITIAN

A. Data Penelitian

Data penelitian berupa buku cerita rakyat yang berasal dari Bandung. Dokumen beserta cerita tersebut selanjutnya diolah menjadi korpus teks paralel bahasa Indonesia dan bahasa Sunda. Adapun jumlahnya yaitu 3000 pasangan kalimat korpus paralel bahasa Indonesia dan bahasa Sunda.

B. Implementasi Mesin Penerjemah Statistik Indonesia ke Bahasa Sunda

1. Implementasi SRILM

Model bahasa digunakan sebagai sumber pengetahuan berbasis teks dengan nilai-nilai probabilistik. Penelitian ini menggunakan n-gram sebagai *language model*. Model bahasa dibangun dengan tools SRILM. Model bahasa akan menghasilkan output dengan format file *.lm. Gambar 3 merupakan tabel model bahasa yang dihasilkan oleh SRILM pada mesin penerjemah statistik bahasa Indonesia – Bahasa Sunda.

```
\data\
ngram 1=3619
ngram 2=13264
ngram 3=1019

\1-grams:
-2.81221abdi    -0.2256695
-3.475367       acan    -0.08614869
-4.016886       aeyuna  -0.08614868
.....
\2-grams
-2.210994       abdi gaduh      -0.1470668
-0.811752       albeum dibuka   -0.1470667
-1.840778       ari ayeuna      -0.06591805
.....
\3-grams
-0.4813389      tos ampir sasasih
-0.5218764      ka angin peuting
-0.577717       mawa cai kopi
```

Gambar 3. Tabel model bahasa dengan bahasa Indonesia sebagai bahasa target

2. Implementasi Giza++ Untuk Pemodelan Translasi

Model 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 *lexical model table.*

Dokumen-dokumen tersebut terdapat dalam folder "train" yang didalamnya terdapat 4 file yaitu "corpus, giza.sd-id, giza.id-sd dan model".

```
1       UNK     0
2       teh     717
3       inu     493
4       nu      491
5       ka      356
6       awit    306
7       mah     298
8       ku      254
9       mila    167
10      cek     160
```

Gambar 4. Dokumen *vocabulary corpus* bahasa Sunda

Gambar 4 merupakan isi dari dokumen *vocabulary* corpus. Angka 1 sampai 10 pada dokumen vocabulary corpus merupakan *uniq id* untuk setiap data token, sedangkan angka disebelah kanan token menunjukkan frekuensi kemunculan. *Vocabulary corpus* yang dihasilkan mesin penerjemah bahasa Indonesia – bahasa Sunda.

```
# Sentence pair (55) source length 9 target length 9
alignment score : 2.99844e-05

teh lila imut , tuluy nyekel leungeun inu pageuh
NULL ({ }) tidak ({ 1 }) lama ({ 2 }) senyum
({ 3 }) , ({ 4 }) terus ({ 5 }) memegang ({ 6 })
tangan ({ 7 }) inu ({ 8 }) erat ({ 9 })
```

Gambar 5. Dokument *alignment* bahasa Indonesia - bahasa Sunda

Gambar 5 merupakan Dokumen *alignment* Bahasa Indonesia ke Sunda terdapat tiga baris kalimat. Baris pertama berisi letak kalimat target (55) dalam korpus, panjang kalimat sumber (9), panjang kalimat target (9) dan skor *alignment* 2.99844e-05. Baris kedua merupakan bahasa sumber dan baris ketiga merupakan *alignment* kalimat bahasa target terhadap kalimat bahasa sumber. Kata "lila" ({ 2 }) memiliki makna bahwa kata "lila" pada kalimat bahasa target, di-*align* ke kata keenam pada kalimat bahasa sumber yaitu "lama".

```
atas luhur 0.8333333
tertawa seuri 0.9090909
kaya beunghar 1.0000000
tidur sare 1.0000000
bukan lain 0.9574468
lama lami 0.5000000
```

Gambar 6. Tabel *lexical model* mesin penerjemah bahasa Indonesia - bahasa Sunda

Gambar 6 merupakan tampilan dari tabel *lexical model* pada mesin penerjemah statistik bahasa Indonesia - bahasa Sunda. Proses *lexical translation table* oleh Giza++ akan menghasilkan tabel translasi *lexical model* yang terdiri dari tabel kata yang berisi kosakata dari bahasa sumber yang memiliki makna pada bahasa sasaran ataupun sebaliknya (leksikal). Setiap kosakata yang dihasilkan memiliki jumlah probabilitas sebesar 1.0 yang terbagi dalam beberapa makna.

C. Pengujian Hasil Terjemahan Mesin Translasi

Pengujian hasil translasi dilakukan dengan cara pengujian otomatis dari mesin penerjemah. Pengujian otomatis dari mesin penerjemah menghasilkan keluaran berupa nilai akurasi yang dihasilkan oleh BLEU (*Billingual Evaluation Understudy*). Hasil pengujian ini nantinya akan menjadi parameter untuk membandingkannya dengan hasil pengujian setelah dilakukan penambahan korpus monolingual.

Langkah pada pengujian otomatis, korpus yang akan diuji terlebih dahulu melalui langkah translasi otomatis yang akan memberikan *output* berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin. Pengujian mesin menggunakan metode *K-Fold Cross-Validation.*

Setelah membuat output berupa hasil translasi otomatis dari mesin penerjemah, langkah selanjutnya adalah mendapatkan skor dari output dengan cara membandingkan output tersebut dengan korpus manual bahasa target yang telah dibuat sebelumnya.

```
acer@acer-Aspire-4755::~/moses/M-
A$ ~/mosesdecoder/scripts/generic/multi-bleu.perl
Fold-A.ref < output.sdA
BLEU = 25.40, 55.3/33.1/20.2/12.9 (BP=0.966,
ratio=0.966, hyp_len=4823, ref_len=5001)
```

```
acer@acer-Aspire-4755::~/moses/M-
B$ ~/mosesdecoder/scripts/generic/multi-bleu.perl
Fold-B.ref < output.sdB
BLEU = 27.02, 56.7/35.1/21.3/13.2 (BP=0.986,
ratio=0.986, hyp_len=4873, ref_len=4941)
```

```
acer@acer-Aspire-4755::~/moses/M-
C$ ~/mosesdecoder/scripts/generic/multi-bleu.perl
Fold-C.ref < output.sdC
BLEU = 30.34, 57.8/37.8/24.7/15.8 (BP=0.998,
ratio=0.998, hyp_len=4400, ref_len=4409)
```

```
acer@acer-Aspire-4755::~/moses/M-
D$ ~/mosesdecoder/scripts/generic/multi-bleu.perl
Fold-D.ref < output.sdD
BLEU = 27.93, 56.9/35.3/22.4/14.3 (BP=0.986,
ratio=0.986, hyp_len=4611, ref_len=4674)
```

```
acer@acer-Aspire-4755::~/moses/M-
E$ ~/mosesdecoder/scripts/generic/multi-bleu.perl
Fold-E.ref < output.sdE
BLEU = 30.57, 57.9/37.5/24.6/16.7 (BP=0.996,
ratio=0.996, hyp_len=4623, ref_len=4643)
```

Gambar 7. Tampilan Skor BLEU

Gambar 7 merupakan nilai skor BLEU pada mesin penerjemah bahasa Indonesia ke bahasa Sunda sebelum dilakukan penambahan kuantitas korpus monolingual adalah sebesar 28.25%.

D. Penambahan Kuantitas Korpus Monolingual

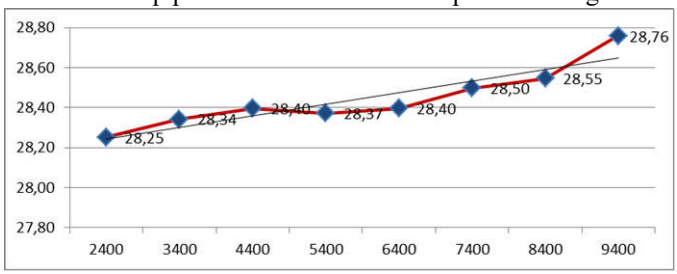
Setelah mendapatkan nilai awal dari korpus uji, maka langkah selanjutnya adalah melakukan proses penambahan kuantitas korpus monolingual pada bahasa Sunda. Proses penambahan kuantitas korpus bahasa Sunda seperti yang telah dijelasakan pada bab sebelumnya. Proses penambahan kuantitas korpus dilakukan penulis dengan menyiapkan 7 file korpus bahasa Sunda yang setiap masing-masing korpus berisi 1000 kalimat bahasa Sunda yang nantinya akan ditambahkan ke dalam korpus awal.

E. Pengujian Ulang Hasil Terjemahan Mesin Translasi

Langkah berikutnya adalah melakukan pengujian kembali hasil terjemahan mesin translasi bahasa Indonesia ke bahasa Sunda yang telah melewati proses *penambahan kuantitas korpus monolingual.* Langkah pengujian yang dilakukan sama halnya dengan langkah pengujian sebelumnya, yakni dengan cara melakukan pengujian otomatis dengan menggunakan metode *K-Fold Cross-Validation* yang akan memberikan *output* berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin dan pengujian oleh ahli bahasa.

1. Pengujian Otomatis

Pengujian dilakukan dengan cara membandingkan nilai BLEU hasil terjemahan otomatis dari mesin penerjemah bahasa Indonesia - bahasa Sunda sebelum dan setelah melewati tahap penambahan kuantitas korpus monolingual.



Gambar 10. Tampilan grafik nilai BLEU sebelum dan setelah penambahan kuantitas korpus monolingual

Gambar 10 merupakan tampilan grafik sebelum mengalami penambahan kuantitas korpus monolingual, nilai BLEU sebelum dilakukan penambahan kuantitas korpus monolingual pada korpus paralel 2.400 kalimat sebesar 28.25% dan setelah dilakukan penambahan kuantitas korpus monolingual 3.400 sebesar 28.34%, 4.400 sebesar 28.40%, 5.400 sebesar 28.37%, 6.400 sebesar 28.40%, 7.400 sebesar 28.50%, 8.400 sebesar 28.55% dan 9.400 sebesar 28.76%

2. Pengujian Ahli Bahasa

Pengujian ahli bahasa dilakukan terhadap hasil terjemahan mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda. Pengujian dilakukan dengan mengambil kalimat yang mengalami perubahan pada hasil terjemahan otomatis yang terdapat pada korpus Paralel sebelum dan sesudah dilakukan penambahan kuantitas korpus monolingual sebanyak 20 kalimat. Ahli bahasa menilai apakah hasil terjemahan lebih baik, sama, atau lebih buruk berdasarkan tingkat akurasi terjemahan kata. Perhitungan akurasi dilakukan dengan Persamaan berikut :

$$P = \frac{C}{R} 100\%$$

Keterangan:

P = Persentase akurasi

C = Jumlah kata yang diterjemahkan dengan tepat menurut penilaian dari ahli bahasa

R = Jumlah kata hasil terjemahan

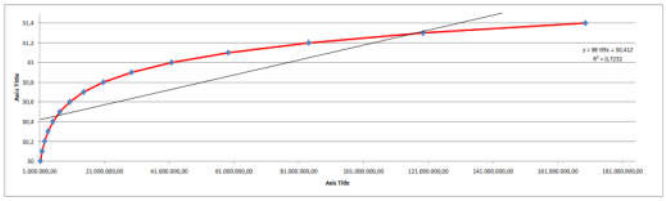
TABEL 1
TABEL AKURASI AHLI BAHASA

Kalimat Hasil Terjemahan	Ahli Bahasa	C,R	$P = \frac{C}{R} 100\%$
Sebelum Penambahan Kuantitas Korpus Monolingual	Bella Yuda	C = 115, R = 153	75.16%
Setelah Penambahan Kuantitas Korpus Monolingual	Bella Yuda	C = 137, R=153	89.54%

Tabel 1 merupakan tampilan tabel akurasi dari ahli bahasa sebelum mengalami penambahan kuantitas korpus monolingual, nilai dari ahli bahasa sebesar 75.16% dan setelah dilakukan penambahan kuantitas korpus monolingual didapat nilai dari ahli bahasa sebesar 89.54%. Terdapat peningkatan nilai BLEU sebesar 19.13% dilihat dari perbandingan sebelum dan sesudah mengalami penambahan kuantitas korpus monolingual.

F. Perkiraan Jumlah Korpus Berdasarkan Penambahan Kuantitas Korpus Monolingual

Persamaan jumlah korpus pada mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda dapat dihitung berdasarkan fungsi logaritma. Adapun nilai dari fungsi logaritma diperoleh dari grafik uji akurasi terhadap kuantitas korpus.



Gambar 11. Tampilan grafik perkiraan jumlah korpus berdasarkan penambahan kuantitas korpus monolingual

IV. KESIMPULAN

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

1. Berdasarkan hasil penelitian, penambahan kuantitas korpus monolingual dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia – bahasa Sunda.
2. Persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia – bahasa Sunda dengan korpus 2.400 kalimat sebesar 28.25% dan setelah dilakukan penambahan kuantitas korpus monolingual 3.400 sebesar 400 sebesar 28.34% meningkat 0,32%, dengan korpus 4.400 sebesar 28.40% meningkat 0,51%, dengan korpus 5.400 sebesar 28.37% meningkat 0.42%, dengan korpus 6.400 sebesar 28.40% meningkat 0.51%, dengan korpus 7.400 sebesar 28.50% meningkat 0.87%, dengan korpus 8.400 sebesar 28.55% meningkat 1.04% dan dengan korpus 9.400 sebesar 28.76% meningkat 1.79%

3. Penilaian yang dilakukan oleh ahli bahasa menghasilkan persentase peningkatan sebelum dilakukan penambahan kuantitas korpus monolingual sebesar 83.77% dan setelah dilakukan penambahan kuantitas korpus monolingual menjadi 96.75%. Dari hasil tersebut terjadi peningkatan hasil terjemahan sebesar 19.13%.
4. Untuk mencapai nilai BLEU hingga 31,40% dibutuhkan setidaknya 169.574.400 korpus monolingual yang ditambahkan kedalam mesin penerjemah statistik bahasa Indonesia ke bahasa Sunda.
5. Penurunan atau peningkatan nilai BLEU ketika di lakukan penambahan kuantitas korpus monolingual dipengaruhi oleh kualitas korpus yang ditambahkan.
6. Perlu penambahan jumlah korpus untuk meningkatkan kualitas terjemahan mesin penerjemah statistik.
7. Perlu menggunakan korpus yang berkualitas agar mendapatkan nilai akurasi yang lebih tinggi.
8. Perlu dilakukan penelitian lanjutan untuk melakukan analisis dalam menghasilkan terjemahan bahasa Indonesia – bahasa Sunda dengan menggunakan metode penelitian yang lain.
9. Melakukan implementasi mesin penerjemah statistik ke dalam bahasa daerah yang lain dengan metode penambahan kuantitas korpus monolingual.
10. Perlu dilakukan pengecekan ulang terhadap korpus teks paralel untuk mencegah kesalahan penulisan (typo).

REFERENSI

[1] Sujaini, Herry. (2014). Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke bahasa Melayu Pontianak. Jurnal ELKHA Vol. 6. No 2

[2] Hidayat, Andri. (2015). Aplikasi Penerjemah Dua Arah Bahasa Indoensia - Bahasa Melayu Sambas Berbasis Web Dengan Menggunakan Decoder *Moses.* Pontianak: Universitas Tanjungpura

[3] Hadi, Ibnu. 2014. Uji Akurasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Melayu Sambas dan Bahasa Melayu Sambas ke Bahasa Indonesia. Pontianak: JUSTIN Vol 3 No 1.

[4] Manning, Christopher D., Schutze, Hinrich. 2000. Foundations Of Statistical Natural Language Processing. London : The MIT Press Cambridge Massachusetts.

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

[6] Indrayana, Danny. 2016. Meningkatkan Akurasi Mesin Penerjemah Bahasa Indonesia ke Bahasa Melayu Pontianak Dengan Part Of Speech. Pontianak: JUSTIN Vol 3 No 1.

[7] Mandira, Soni. 2016. Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik. Pontianak: JEPIN Vol 2 No 1.

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

[9] Hasbiansyah, Muhammad. 2016. Tuning For Quality Untuk Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia - Bahasa Dayak Kanayatn. Pontianak, JEPIN Vol 1 No 1 2016.

[10] Koehn, Philipp. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

[11] McEnery, T., et al. 2006. Corpus-Based Language Studies: An Advanced Resource Book. Oxon: Routledge.

[12] Hunston, S. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

[13] Amalia, Farida. 2009. *"Ideologi dalam Penerjemahan"*. Makalah disajikan dalam Forum Ilmiah Pengajar Bahasa Prancis Prancis se Indonesia di Bandung.

[14] Sudarno, A.P. 2011. *Penerjemahan Buku Teori dan Aplikasi.* Surakarta :UNS Press.

[15] Sheddy, N. Tjandra. 2005. *Analisis Penerjemahan.* Jakarta, library UI Vol 8 No 1.

[16] Papineni, K., et al. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Philadelphia : Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.