Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics
Vol 1, No 2
September
2018
e-ISSN :
2621-9336

Perbaikan Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik (Studi Kasus : Bahasa Indonesia – Jawa Krama)

Muhammad Gerdy Asparilla[#1], Herry Sujaini[#2], Rudy Dwi Nyoto[#3]

[#]*Program Studi Informatika, Universitas Tanjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak 78124

[1]muhammadgerdy@gmail.com
[2]rudydn@informatika.untan.ac.id
[3]herry_sujaini@yahoo.com

Abstract— **Language is a communication tool that is used as a means to interact with the surrounding community. The ability to master many languages will certainly make it easier to interact with other people from different regions. Therefore, translators are needed to increase knowledge of various languages. Statistical Machine Translation (Statistical Machine Translation) is a machine translation approach with translation results produced on the basis of statistical models whose parameters are taken from the results of parallel corpus analysis. Parallel body is a pair of corpus containing sentences in a language and translation. One feature that is used to improve the quality of translation results is with corpus optimization. The aim to be achieved in this study is to look at the influence of the quality of the corpus by filtering out pairs of sentences with quality translation. The filter used is the minimum value of each sentence that is tested by the Bilingual Evaluation Understudy (BLEU) method. Testing is done by comparing the accuracy of the results of the translation before and after corpus optimization. From the results of the research, the use of corpus optimization can improve the quality of translation for Indonesian translation machines to Javanese manners. This can be seen from the results of testing by adding corpus optimization to 15 test sentences outside the corpus, there is an average increase in BLEU values of 10.53% and by using 100 test sentences derived from corpus optimization there is an average increase in BLEU values of 11.63% in automated testing and 0.03% on testing by linguists. Based on this, the machine translating Indonesian statistics into Javanese language using the corpus optimization feature can increase the accuracy of the translation results.**

Keywords - statistical translator machine, parallel corpus, quality of translation results, indonesian, javanese

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 kualitas hasil terjemahan adalah dengan optimasi korpus. Tujuan yang ingin dicapai dalam penelitian ini adalah melakukan untuk melihat pengaruh kualitas korpus dengan memfilter pasangan kalimat-kalimat dengan terjemahan berkualitas. Filter yang digunakan adalah nilai minimal setiap kalimat yang di uji dengan metode Bilingual Evaluation Understudy (BLEU). Pengujian dilakukan dengan membandingkan nilai akurasi hasil terjemahan sebelum dan setelah optimasi korpus. Dari hasil penelitian, penggunaan optimasi korpus dapat meningkatkan kualitas terjemahan untuk mesin penerjemah bahasa Indonesia ke bahasa Jawa krama. Hal itu terlihat dari hasil pengujian dengan menambahkan optimasi korpus pada 15 kalimat uji diluar korpus terdapat peningkatan rata - rata nilai BLEU sebesar 10.53% dan dengan menggunakan 100 kalimat uji yang berasal dari korpus optimasi terdapat peningkatan rata-rata nilai BLEU sebesar 11.63% pada pengujian otomatis serta 0.03% pada pengujian oleh ahli bahasa. Berdasarkan hal tersebut, mesin penerjemah statistik bahasa Indonesia ke bahasa Jawa krama dengan penggunaan fitur optimasi korpus dapat meningkatkan nilai akurasi hasil terjemahan.**

Kata Kunci - mesin penerjemah statistik, korpus paralel, kualitas hasil terjemahan, Indonesia, Jawa Krama.

I. PENDAHULUAN

Mesin penerjemah (MP) adalah mesin yang dapat melakukan translasi dari suatu bahasa ke bahasa yang lain secara otomatis. MP memiliki kegunaan yang praktis dan jelas karena dapat membantu manusia untuk

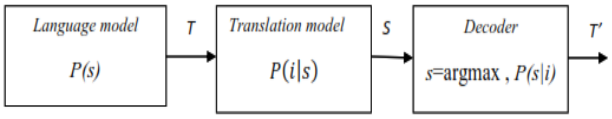
berkomunikasi dengan manusia lainnya yang memiliki bahasa yang berbeda. Masalah ini menjadi lebih penting pada era globalisasi, saat penerjemahan manual oleh manusia yang memiliki sumber daya terbatas dan mahal, MP dapat meningkatkan efesiensinya berdasarkan kualitas mesin penerjemah yang dibuat[1].

Peneltian dalam bidang MPS di Indonesia, terutama untuk mesin penerjemah bahasa Indonesia ke bahasa daerah, sudah mulai banyak dilakukan, diantaranya adalah penelitian akurasi penerjemahan bahasa Indonesia-Jawa menggunakan metode statistik frasa [2], penelitian tentang pengaruh kuantitas korpus terhadap akurasi MPS bahasa Bugis Wajo ke bahasa Indonesia [3], penelitian dengan memperbaiki probabilitas *Lexical Model* untuk meningkatkan akurasi MPS bahasa Indonesia–Jawa [4], penandaan kata dasar dan imbuhan pada korpus paralel untuk memperbaiki akurasi penerjemahan bahasa Indonesia-Dayak Taman [5], penelitian *tuning for quaility* untuk uji akurasi MPS bahasa Indonesia-Dayak Kanayatn [6], serta penelitian tentang sistem penerjemah bahasa Jawa-aksara Jawa berbasis *finite state automata* [7].

Sudah banyak metode yang dilakukan untuk meningkatkan kualitas mesin penerjemah. Salah satu cara untuk meningkatkan kualitas mesin penerjemah statistik (MPS) adalah dengan perbaikan kualitas korpus. Metode untuk memperbaiki kualitas korpus ini yaitu dengan cara memfilter kalimat-kalimat yang berkualitas dari sebuah korpus paralel. Pada penelitian ini, korpus menggunakan bahasa Indonesia sebagai bahasa sumber, sedangkan bahasa target yang digunakan adalah bahasa Jawa karma.

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[8].

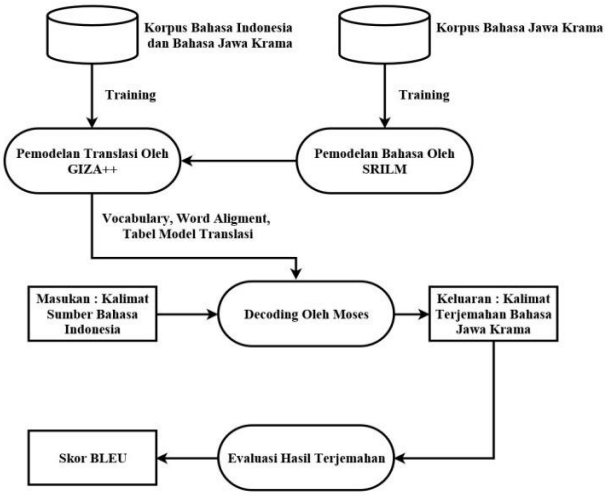


Gambar. 1 Komponen mesin penerjemah statistik [4]

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*.

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*.



Gambar.2 Arsitektur mesin penerjemah statistik Moses bahasa Indonesia ke Jawa krama

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 [9].

B. Moses

Moses adalah sebuah *software* gratis yang merupakan implementasi dari Mesin Penerjemah Statistik.Moses digunakan untuk melatih model statistik teks terjemahan dari bahasa sumber ke bahasa sasaran.Dalam menerjemahkan bahasa, Moses membutuhkan korpus dalam dua bahasa, bahasa sumber dan bahasa sasaran.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 adalah kumpulan teks alami, baik bahasa lisan maupun bahasa tulis, yang disusun secara sistematis. Dikatakan alami karena teks yang dikumpulkan merupakan teks yang diproduksi dan digunakan secara wajar dan tidak dibuat-buat[11].Korpus dapat diklasifikasikan ke dalam enam jenis, yaitu korpus umum, korpus histori, korpus regional, korpus pemelajar, korpus multibahasa, korpus lisan [11]. Korpus paralel adalah dua atau lebih korpus dalam bahasa yang berbeda. Masing-masing korpus memuat teks yang telah diterjemahkan dari satu bahasa ke bahasa lain.

II. METODOLOGI PENELITIAN

A. Data Penelitian

Data penelitian berupa buku novel kesusastraan yang berasal dari Jawa. Dokumen beserta cerita tersebut selanjutnya diolah menjadi korpus teks paralel bahasa Indonesia dan bahasa Jawa krama. Adapun jumlahnya yaitu 5000 pasangan kalimat korpus paralel bahasa Indonesia dan bahasa Jawa krama.

B. Metode Evaluasi

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 *modifiedn-gramprecission 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 [9]:

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

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

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

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 DAN ANALISIS

A. Implementasi Mesin Penerjemah Statistik Indonesia ke Bahasa Jawa Krama

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 3merupakan tabel model bahasa yang dihasilkan oleh SRILM pada mesin penerjemah statistik bahasa Indonesia–Bahasa Sunda.

```
\data\
ngram 1=22474
ngram 2=82654
ngram 3=2571

\1-grams:
-5.037271       kang    -0.06391659
-5.037271       ing     -0.06392457
\2-grams
-3.424013       kang inggih
-3.424013       kang ingucap
\3-grams
-0.5332069      ala yen
-0.9477593      angingu yen
```

Gambar. 3 Tabel model bahasa dengan bahasa Jawa krama 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       kang    2346
3       ing     2145
4       tan     1310
5       wus     815
6       lan     727
7       ingkang 708
8       yen     634
9       samya   569
10      wong    502
```

Gambar. 4 Dokumen *vocabulary corpus* bahasa Jawa krama

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 (6) source length 9 target length 7
alignment score : 2.28887e-07
Jangan merawat pagi hari, jangan tidur
NULL ({ }) sampun ({})nanggel  ({ 1 2 }) sae ({  }) benjing-enjing ({ 3 4 }), ({ 5 }) karsa ({}) paduka ({}) dugekken ({ 6 7}) lampah ({})
```

Gambar. 5 Dokumen *alignment* bahasa Indonesia ke bahasa Jawa krama

Gambar.5 merupakan dokumen *alignment* terdapat tiga baris kalimat. Baris pertama berisi letak kalimat target dalam korpus, panjang kalimat sumber, panjang kalimat target dan nila i*alignment.* Baris kedua merupakan bahasa sumber dan baris ketiga merupakan *alignment* kalimat bahasa target terhadap kalimat bahasa sumber.

Kalimat bahasa target di-*align* ke kalimat bahasa sumber. Kata sampun ({ }) memiliki makna bahwa kata "sampun" pada kalimat bahasa target, di-*align* ke kata pertama pada kalimat bahasa sumber yaitu "jangan".

```
Narajang perilaku 0.0909091
Tingkahe perilaku 0.0909091
Singgih perilaku  0.0909091
Anenggih perilaku 0.0909091
Laku perilaku     0.1818182
Uwong perilaku    0.0909091
```

Gambar. 6 Tabel *lexical model* mesin penerjemah bahasa Indonesia ke bahasaJawa krama

Gambar.6 merupakan tampilan dari tabel *lexical model* pada mesin penerjemah statistik bahasa Indonesiake bahasa Jawa krama. 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).

B. 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 proses optimasi korpus yaitu dengan memfilter kalimat – kalimat berkualitas pada korpus.

Langkah pada pengujian otomatis, korpus yang akan diuji menggunakan 5000 kalimat yang terlebih dahulu akan dibagi dalam sepuluh *fold*, yaitu: *fold* 1: kalimat nomor 1-500, *fold* 2: kalimat nomor 501-1000, *fold* 3: kalimat nomor 1001-1500, *fold* 4: kalimat nomor 1501-2000, *fold* 5: kalimat nomor 2001-2500, *fold* 6: kalimat nomor 2501-3000, *fold* 7: kalimat nomor 3001-3500, *fold* 8: kalimat nomor 3501-4000, *fold* 9: kalimat nomor 4001-4500, *fold* 10: kalimat nomor 4501-5000, lalu akan melalui langkah translasi otomatis yang akan memberikan output berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin. Kalimat uji menggunakan 15 kalimat yang berasal dari luar korpus.

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. Hasil BLEU salah satu grup uji dapat dilihat pada gambar. 7.

.

```
toshiba@toshiba-Satellite-L840:~moses$ ~/moses/mosesdecoder/scripts/generic/multi-bleu.perl ref < out BLEU = 31.96, 37.8/15.3/6.6/3.2 (BP=0.916, ratio=0.919, hyp_len=5925, ref_len=6445)
toshiba@toshiba-Satellite-L840:~moses$
```

Gambar .7 Tampilan Skor BLEU Grup Uji A

Gambar.7 merupakan nilai skor BLEU Grup Uji A pada mesin penerjemah bahasa Indonesia ke bahasa Jawa krama sebelum dilakukan optimasi adalah sebesar 31,96%. Nilai masing-masing Grup Uji sebelum optimasi dapat dilihat pada Tabel I.

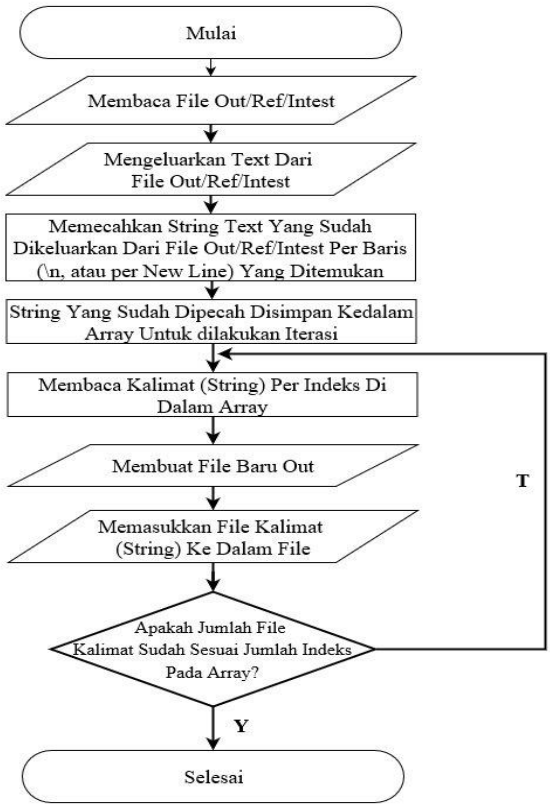
TABEL 1

NILAI BLEU SEBELUM OPTIMASI KORPUS

Grup Uji	Korpus (fold)	Skor BLEU (%)
A	2,3,4,5,6,7,8,9,10	31,96
B	1,3,4,5,6,7,8,9,10	35,04
C	1,2,4,5,6,7,8,9,10	38,38
D	1,2,3,5,6,7,8,9,10	36,82
E	1,2,3,4,6,7,8,9,10	18,07
F	1,2,3,4,5,7,8,9,10	41,95
G	1,2,3,4,5,6,8,9,10	30,43
H	1,2,3,4,5,6,7,9,10	38,21
I	1,2,3,4,5,6,7,8,10	41,52
J	1,2,3,4,5,6,7,8,9	36,74

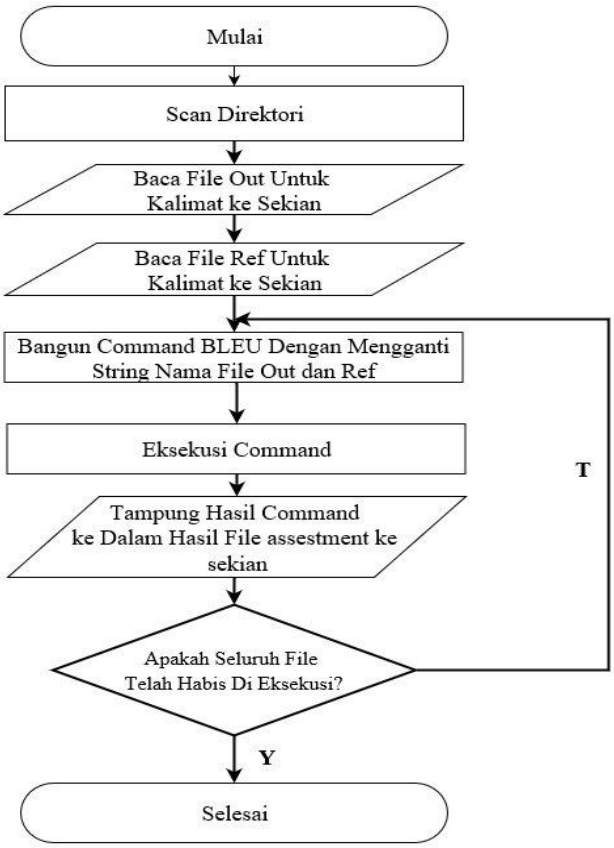
C. Optimasi Korpus

Setelah mendapatkan nilai awal dari korpus uji, maka langkah selanjutnya adalah melakukan optimasi pada korpus hasil output dari pengujian sebelumnya. Optimasi korpus dengan cara memfilter kalimat-kalimat berkualitas pada korpus. Masing-masing nilai BLEU pada kalimat terjemahan akan dihitung terhadap kalimat referensi. Semua kalimat yang memiliki nilai BLEU yang kurang dari n% akan dieleminasi secara otomatis dengan menggunakan aplikasi bantu yang telah dibuat dengan bahasa pemrograman PHP yaitu *text_divider.php*, *assetsment_single.php*, dan *calculate.php*, sisanya digunakan sebagai korpus paralel untuk mesin yang baru. Ambang batas nilai BLEU yang digunakan, n, untuk memfilter kalimat-kalimat dalam korpus paralel adalah sebesar 10%.



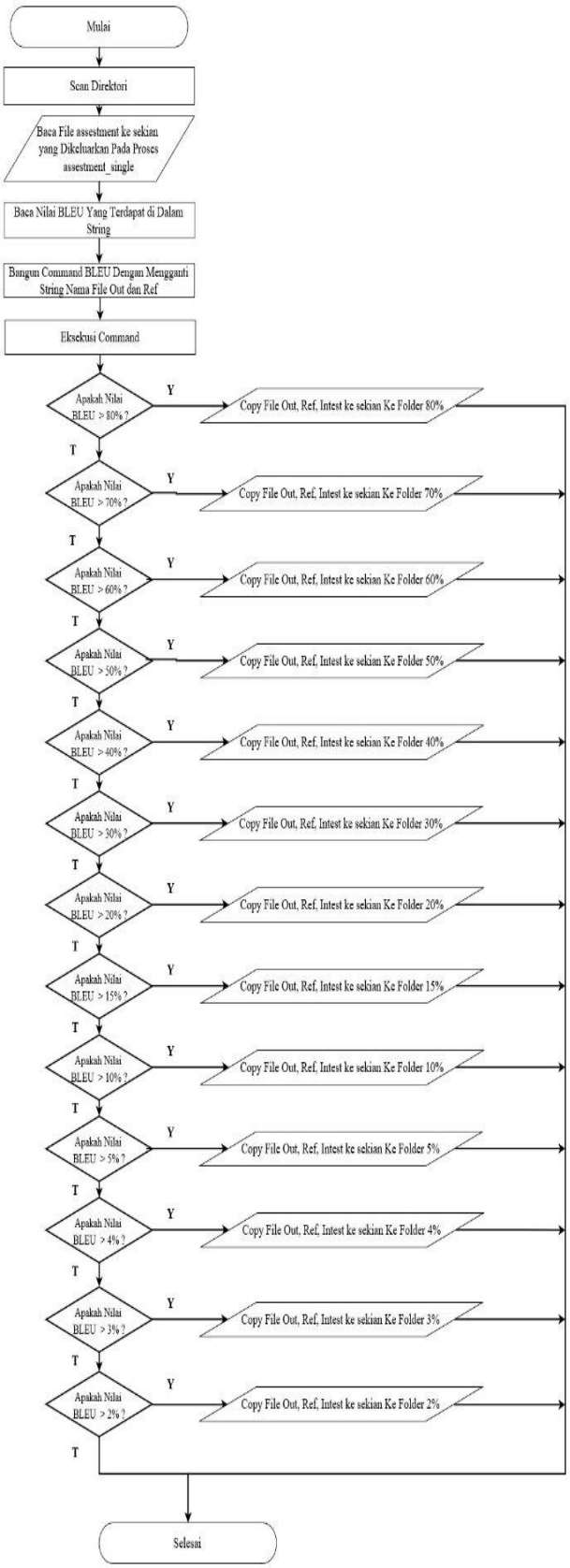
Gambar. 8 *Flowchart text_divider.php*

Gambar. 8 merupakan tampilan dari aplikasi bantu menggunakan bahasa pemrograman PHP yaitu *text_divider.php* yang berfungsi untuk membagi *file out* (hasil terjemahan otomatis dari mesin), *file ref* (kalimat kalimat terjemahan manual), dan *file intest* (kalimat - kalimat sumber) menjadi per kalimat.



Gambar. 9 *Flowchart assetsment_single.php*

Gambar. 9 merupakan aplikasi bantu dengan bahasa pemrograman PHP *yaitu assetsment_single.php* yang berfungsi untuk menghitung nilai BLEU masing-masing kalimat.



Gambar. 10 *Flowchart calculate.php*

Gambar. 10 merupakan aplikasi bantu menggunakan bahasa pemrograman PHP, yaitu *calculate.php* yang berfungsi untuk mengeleminasi kalimat – kalimat dengan nilai batas n (%) berdasarkan nilai BLEU yang telah dihitung per kalimat sebelumnya menggunakan aplikasi bantu *assestment_single.php*. Sebagai contoh kalimat yang dieliminasi dapat dilihat pada Tabel II, sedangkan jumlah kalimat yang dieliminasi dengan batas n% dapat dilihat pada Tabel III.

TABEL II

CONTOH KALIMAT YANG DI ELEMINASI

Bahasa Jawa Krama	Bahasa Indonesia	BLEU (%)
Surawana puniku, wusing tanya ki santri gya wangsul	Surawana puniku, tanya pulang ki santri	0.00
Wanci sang hyang arkamehtumamengwukir kadikaluwungkang Anjok	Pada saat sang hyang arka hamper datang ke pondok	0.00

TABEL III

JUMLAH KALIMAT YANG DI ELEMINASI DENGAN BATAS NILAI BLEU N (%)

n (%)	Jumlah Kalimat	Presentase Kalimat (%)
0	4999	100.00
2	1060	21.20
3	1060	21.20
4	1060	21.20
5	1060	21.20
10	1011	20.22
15	830	16.60
20	660	13.20
30	400	8.00
40	205	4.10
50	105	2.10
60	64	1.28
70	29	0.58
80	17	0.34

D. Pengujian Ulang Hasil Terjemahan Mesin Translasi

Langkah berikutnya adalah melakukan pengujian kembali hasil terjemahan mesin translasi bahasa Indonesia ke bahasa Jawa krama yang telah melewati proses optimasi korpus. Langkah pengujian yang dilakukan sama halnya dengan langkah pengujian sebelumnya, yakni dengan cara melakukan pengujian otomatis dengan mesin penerjemah yangakan memberikan *output* berupa korpus

dalam bahasa target yang telah diterjemahkan oleh mesin dan juga akan dilakukan pengujian oleh ahli bahasa.

1. Pengujian Otomatis

Pengujian dilakukan dengan cara membandingkan nilai BLEU hasil terjemahan otomatis dari mesin penerjemah bahasa Indonesia-bahasa Jawa krama sebelum dan setelah melewati tahap optimasi korpus. Korpus yang akan diuji menggunakan 1011 kalimat hasil optimasi dengan nilai batas n (10%) yang terlebih dahulu akan dibagi dalam sepuluh *fold*, yaitu: *fold* 1: kalimat nomor 1-100, *fold* 2: kalimat nomor 101-200, *fold* 3: kalimat nomor 201-300, *fold* 4: kalimat nomor 301-400, *fold* 5: kalimat nomor 401-500, *fold* 6: kalimat nomor 501-600, *fold* 7: kalimat nomor 601-700, *fold* 8: kalimat nomor 701-800, *fold* 9: kalimat nomor 801-900, *fold* 10: kalimat nomor 901-1011 [12], lalu akan melalui langkah translasi otomatis yang akan memberikan output berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin. Kalimat uji menggunakan 15 kalimat yang berasal dari luar korpus yang sama seperti pengujian sebelumnya. Nilai masing-masing Grup Uji setelah optimasi dapat dilihat pada Tabel IV. Adapun perbandingan nilai BLEU sebelum dan sesudah optimasi dapat dilihat pada Tabel V dan VI.

Berdasarkan Tabel V dapat dilihat bahwa terjadi peningkatan nilai BLEU setelah dilakukan optimasi korpus dengan korpus uji berasal dari luar korpus dengan jumlah kalimat uji sebesar 15 kalimat.Terdapat peningkatan rata-rata nilai BLEU sebesar 10.53%.

TABEL IV

NILAI BLEU SETELAH OPTIMASI KORPUS

Grup Uji	Korpus (fold)	Skor BLEU (%)
A	2,3,4,5,6,7,8,9,10	43,16
B	1,3,4,5,6,7,8,9,10	39,01
C	1,2,4,5,6,7,8,9,10	38,84
D	1,2,3,5,6,7,8,9,10	37,56
E	1,2,3,4,6,7,8,9,10	53,24
F	1,2,3,4,5,7,8,9,10	46,59
G	1,2,3,4,5,6,8,9,10	50,03
H	1,2,3,4,5,6,7,9,10	50,23
I	1,2,3,4,5,6,7,8,10	49,73
J	1,2,3,4,5,6,7,8,9	46,03

TABEL V

TABEL PERBANDINGAN NILAI BLEU I

No	Korpus (fold)	Sebelum Optimasi	Sesudah Optimasi	Peningkatan
A	2,3,4,5,6,7,8,9,10	31,96	43,16	11,2
B	1,3,4,5,6,7,8,9,10	35,04	39,01	3,97
C	1,2,4,5,6,7,8,9,10	38,38	38,84	0,46
D	1,2,3,5,6,7,8,9,10	36,82	37,56	0,74
E	1,2,3,4,6,7,8,9,10	18,07	53,24	35,17
F	1,2,3,4,5,7,8,9,10	41,95	46,59	4,64
G	1,2,3,4,5,6,8,9,10	30,43	50,03	19,6
H	1,2,3,4,5,6,7,9,10	38,21	50,23	12,02
I	1,2,3,4,5,6,7,8,10	41,52	49,73	8,21
J	1,2,3,4,5,6,7,8,9	36,74	46,03	9,29

Kemudian dilakukan pengujian lanjutan dengan menggunakan kalimat uji yang berasal dari korpus hasil optimasi sebesar 100 kalimat untuk mendapatkan perbandingan peningkatan nilai BLEU dari mesin penerjemah sebelum dioptimasi dan sesudah dilakukan optimasi.

TABEL VI

TABEL PERBANDINGAN NILAI BLEU II

No	Korpus (fold)	Sebelum Optimasi	Sesudah Optimasi	Peningkatan
A	2,3,4,5,6,7,8,9,10	71,22	77,95	6,73
B	1,3,4,5,6,7,8,9,10	69,82	78,74	8,92
C	1,2,4,5,6,7,8,9,10	71,80	78,80	7,00
D	1,2,3,5,6,7,8,9,10	70,10	79,74	9,64
E	1,2,3,4,6,7,8,9,10	72,58	73,47	0,89
F	1,2,3,4,5,7,8,9,10	50,66	79,44	28,78
G	1,2,3,4,5,6,8,9,10	53,15	78,63	25,48
H	1,2,3,4,5,6,7,9,10	70,04	80,72	10,68
I	1,2,3,4,5,6,7,8,10	70,86	80,05	9,19
J	1,2,3,4,5,6,7,8,9	70,94	79,91	8,97

Pada Tabel VI dapat dilihat bahwa terjadi peningkatan nilai BLEU setelah dilakukan optimasi korpus dengan

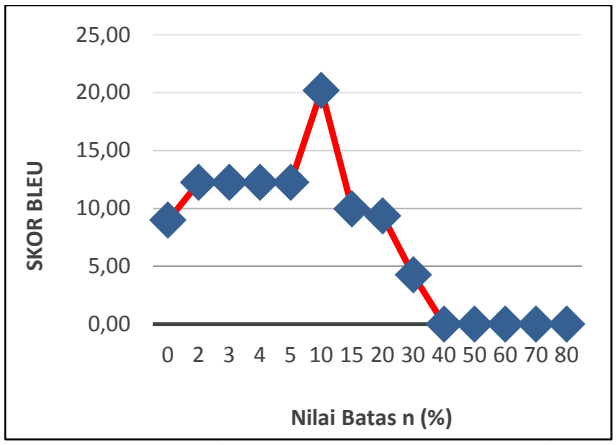
korpus uji berasal dari luar korpus dengan jumlah kalimat uji sebesar 100 kalimat yang berasal dari korpus hasil optimasi. Terdapat peningkatan rata-rata nilai BLEU sebesar 11.63%.

Untuk setiap nilai batas BLEU (n), juga dilakukan pengujian dengan menyertakan 15 kalimat uji yang berasal dari luar korpus yang digunakan.

TABEL VII

NILAI BLEU SISTEM MPS HASIL PEMFILTERAN

n (%)	Skor BLEU (%)
0	8,97
2	12,24
3	12,24
4	12,24
5	12,24
10	20,17
15	9,95
20	9,34
30	4,24
40	0,00
50	0,00
60	0,00
70	0,00
80	0,00



Gambar.11 Tampilan grafik nilai BLEU sistem MPS hasil pemfilteran

Dari data hasil ekpserimen pada Tabel VII dan gambar. 11, dapat dilihat bahwa nilai BLEU untuk sistem dengan n = 5 % dan 10% meningkat dari nilai dasar, akan tetapi menurun mulai dengan sistem korpus n = 15% dan seterusnya. Dari data tersebut terlihat bahwa keseimbangan yang optimal terdapat pada sistem dengan korpus n = 10%.

2. Pengujian Ahli Bahasa

Pengujian ahli bahasa dilakukan terhadap hasil terjemahan mesin penerjemah statistik bahasa Indonesia ke bahasa Jawa krama. 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 15 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 VIII

PENILAIAN AKURASI AHLI BAHASA

Kalimat Hasil Terjemahan	Ahli Bahasa	C,R	$P = \frac{C}{R}$ 100%
Sebelum Optimasi Korpus	Danuri	C = 102, R = 158	0.64%
Setelah Optimasi Korpus	Danuri	C = 108, R=162	0.66%

Tabel VIII merupakan tampilan tabel akurasi dari ahli bahasa sebelum optimasi korpus, nilai dari ahli bahasa sebesar 0.64% dan setelah dilakukan optimasi korpus didapat nilai dari ahli bahasa sebesar 0.66%. Sehingga peningkatan akurasi dapat dihitung C = $\frac{b-a}{a}$ 100% , sehingga C = $\frac{0.66-0.64}{0.64}$ 100% dan didapat nilai peningkatan sebesar 0.03%.

IV. KESIMPULAN

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

1. Optimasi korpus dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia ke bahasa Jawa krama.
2. Persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia ke bahasa Jawa krama dengan 15 kalimat uji dari luar korpus sebesar 10.53%
3. Persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia ke bahasa Jawa krama dengan 100 kalimat uji dari korpus hasil optimasi sebesar 11.63%
4. Penilaian yang dilakukan oleh ahli bahasa menghasilkan persentase peningkatan akurasi hasil terjemahan sebesar 0.03%.
5. Dibutuhkan penelitian lebih lanjut dengan merubah jumlah korpus uji, asal korpus uji, permodelan bahasa, permodelan translasi, pembagian *fold,* penambahan kuantitas korpus, dan lain sebagainya, untuk mengetahui sejauh mana pengaruh optimasi korpus dalam meningkatkan kualitas mesin penerjemah statistik.

REFERENSI

[1] Sujaini, H. & Bijaksana, A., Strategi Memperbaiki Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik.*Seminar Nasional Teknologi Informasi XI*, 2016

[2] 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

[3] 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

[4] 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.

[5] 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.

[6] 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.

[7] 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. 38-44, Oktober 2014.

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

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

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

[12] Sujaini, H., 2018. Peningkatan Akurasi Penerjemah Bahasa Daerah dengan Optimasi Korpus Paralel.*Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI).* Vol 7, No. 1