Perbaikan Probabilitas *Lexical Model* untuk Meningkatkan Akurasi Mesin Penerjemah Statistik

Soni Mandira¹, Herry Sujaini², Arif Bijaksana Putra Negara³.

Program Studi Teknik Informatika Fakultas Teknik Universitas Tanjungpura¹²³

e-mail: sonimandira@icloud.com¹, herry_sujaini@yahoo.com², arifbpn@gmail.com³

Abstrak—Bahasa merupakan alat komunikasi yang digunakan seseorang untuk menyampaikan ide, gagasan, konsep atau perasaan kepada orang lain. Ragam bahasa yang dimiliki setiap orang berbeda, hal ini dapat menghambat pertukaran informasi karena orang lain tidak memahami maksud dan tujuan yang ingin disampaikan. 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. Tujuan yang ingin dicapai dalam penelitian ini adalah melakukan perbaikan probabilitas *lexical model* pada mesin penerjemah statistik bahasa Jawa ke bahasa Indonesia untuk meningkatkan nilai akurasi hasil terjemahan. Pengujian dilakukan dengan membandingkan nilai akurasi hasil terjemahan sebelum dan setelah perbaikan probabilitas *lexical model*. Penelitian menggunakan korpus paralel sebanyak 5108 kalimat. Pengujian dilakukan dengan dua cara, yaitu pengujian otomatis menggunakan *Bilingual Evaluation Understudy* (BLEU) dan pengujian oleh ahli bahasa Jawa. Hasil dari pengujian adalah terdapat peningkatan persentase nilai BLEU sebesar 0.30% pada pengujian otomatis dan 10.69% pada pengujian oleh ahli bahasa.

Kata Kunci— BLEU score, korpus paralel, *lexical model*, mesin penerjemah statistik.

I. PENDAHULUAN

Bahasa merupakan alat komunikasi yang digunakan seseorang untuk menyampaikan ide, gagasan, konsep atau perasaan kepada orang lain. Adapun fungsi bahasa sebagai alat komunikasi yaitu sarana penyampaian informasi kepada orang lain baik secara lisan maupun tulisan mengenai apa yang ingin disampaikan sebagai media mengekspresikan diri, perasaan, pikiran, keinginan, serta kebutuhan agar orang dapat mengerti maksud dan tujuan yang diinginkan tanpa menghindari tata bahasa yang sudah ada. Berdasarkan Sensus Penduduk pada tahun 2010, dari berbagai bahasa yang terdapat di Indonesia, bahasa Jawa menempati urutan pertama penggunaan bahasa daerah terbanyak yang digunakan sehari-hari dengan persentase 31,79% (68.044.660 juta jiwa) dari total penduduk Indonesia [1].

Bahasa Jawa sebagai bahasa terbesar di Indonesia merupakan aset berharga yang harus dipertahankan agar bahasa Jawa tidak terlupakan dalam kehidupan sehari-hari, seperti yang dilakukan Pemerintah Provinsi Jawa Tengah dengan menjadikan bahasa Jawa sebagai pelajaran muatan lokal [2]. Upaya pelestarian bahasa Jawa juga telah dilakukan dengan memasukkan penggunaan bahasa Jawa

dalam acara pada media massa dan media elektronik. Beberapa upaya tersebut dinilai belum menjadi solusi optimal dikarenakan faktor media massa yang meracuni bahasa daerah dan bahasa Indonesia dengan maraknya penggunaan dialek Jakarta.

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].

Salah satu model yang digunakan untuk menghasilkan hasil terjemahan adalah *lexical model* yang dihasilkan melalui proses *training* secara otomatis dari korpus paralel. *Lexical model* berisi kosakata dari bahasa sumber yang memiliki arti pada bahasa sasaran ataupun sebaliknya (leksikal). Setiap kosakata yang dihasilkan memiliki jumlah probabilitas sebesar 1.0 yang terbagi dalam tiap arti. *Lexical model* yang diperoleh dari proses *training* kurang maksimal karena terdapat banyak kosakata yang tidak sesuai dengan artinya. Oleh karena itu diperlukan perbaikan *lexical model* secara manual dengan bantuan oleh ahli bahasa.

Berdasarkan permasalahan diatas, maka pada penelitian ini dilakukan perbaikan probabilitas *lexical model* untuk meningkatkan akurasi mesin penerjemah statistik bahasa Jawa - bahasa Indonesia.

II. URAIAN PENELITIAN

A. Mesin Penerjemah Statistik

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

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, menerjemahkan 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 [5]

Gambar 2 merupakan arsitektur sistem dari mesin penerjemah statistik Moses. Menurut Herry Sujaini dan Arif B.P.N, 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 [5].

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 sasaran. Saat melakukan penerjemahan 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 [6].

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 [7]. 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*) [8]. Berdasarkan jenis korpus tersebut, untuk penelitian ini penulis akan fokus pada korpus paralel.

D. 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 precision 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 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 sebagai berikut [9]:

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

$$P_n = \frac{\sum_{c \in \text{corpus } n\text{-grams}} \sum_{e \in \text{corpus } n\text{-grams}} \text{count}_{clip(n\text{-gram})}}{\sum_{c \in \text{corpus } n\text{-grams}} \sum_{e \in \text{corpus } n\text{-grams}} \text{count}_{(n\text{-gram})}}$$

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

Keterangan:

BP = *brevity penalty*

c = jumlah kata dari hasil terjemahan otomatis

r = jumlah kata rujukan

P_n = *modified precision 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 DISKUSI

A. Data Penelitian

Data penelitian berupa buku cerita rakyat yang berasal dari Jawa Tengah dan Yogyakarta. Dokumen beserta cerita tersebut selanjutnya diolah menjadi korpus teks paralel bahasa Indonesia dan bahasa Jawa. Adapun jumlahnya yaitu 5108 pasangan kalimat korpus paralel bahasa Jawa dan bahasa Indonesia. Dengan jumlah 6961 *token* untuk korpus bahasa Jawa dan 5536 *token* untuk bahasa Indonesia.

B. Implementasi Mesin Penerjemah Statistik Jawa ke Bahasa Indonesia

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 Jawa - Bahasa Indonesia.

```
\data\
ngram 1=5538
ngram 2=28565
ngram 3=4454

\1-grams:
-4.323414antar          -0.1050127
-3.925996antara         -0.1050127
-----
\2-grams
-0.6994465 hari ini     -0.06201818
-1.412109hari itu      -0.05569446
-----
\3-grams
-0.2124338 cokorda anak agung
-0.6026156 dan anak buahnya
```

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

2. Implementasi Giza++ Untuk Pemodelan Translasi

Model translasi digunakan untuk memasang 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.jw-id, giza.id-jw dan model”.

1	UNK	0
2	sing	1244
3	ora	811
4	iku	773
5	dheweke	767
6	ing	646
7	lan	592
8	wis	517
9	ana	475
10	karo	445

Gambar 4. Dokumen *vocabulary corpus* bahasa Jawa

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 Jawa – bahasa Indonesia terdiri dari 6961 token untuk korpus bahasa Jawa dan 5536 token untuk bahasa Indonesia.

```
# Sentence pair (2381) source length 5 target length
5 alignment score : 0.0729788

ora ana sing bisa ditindakake

NULL ({ }) tidak ({ 1 }) ada ({ 2 }) yang ({ 3 })
bisa ({ 4 }) dilakukan ({ 5 })
```

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

Gambar 5 merupakan dokumen *alignment* Bahasa Jawa – bahasa Indonesia yang memiliki tiga baris kalimat. Baris pertama berisi letak kalimat target (2381) dalam korpus, panjang kalimat sumber (5), panjang kalimat target (5) dan skor *alignment*. Baris kedua merupakan bahasa sumber dan baris ketiga merupakan *alignment* kalimat bahasa target terhadap kalimat bahasa sumber. Kata “dilakukan” ({ 5 }) memiliki makna bahwa kata “dilakukan” pada kalimat

bahasa target, di-align ke kata kelima pada kalimat bahasa sumber yaitu “ditindakake”.

senyap trintim	1.0000000
rumahnya ngomahe	1.0000000
lamaranmu lamaranmu	1.0000000
pojok pojok	1.0000000
perang perange	1.0000000

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

Gambar 6 merupakan tampilan dari tabel *lexical model* pada mesin penerjemah statistik bahasa Jawa - bahasa Indonesia. 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 (*Bilingual Evaluation Understudy*). Hasil pengujian ini nantinya akan menjadi parameter untuk membandingkannya dengan hasil pengujian setelah dilakukan perbaikan *lexical model*.

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. Korpus uji yang digunakan pada tahap ini berjumlah 5108.

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.

```
bangdira@bang-dira:~$
~/NLP/mosesdecoder/scripts/generic/multi-bleu.perl
~/asli/full.id < ~/asli/outputfullasli.id
BLEU = 95.32, 98.1/96.1/94.4/92.8 (BP=1.00, ratio
=1.001, hyp_len=58655, ref_len=58608)
```

Gambar 7. Tampilan nilai dari outputfullasli.id

Gambar 7 merupakan sintaks untuk memperoleh nilai BLEU dan hasilnya yang menampilkan nilai dari outputfullasli.id sebesar 95.32%.

D. Perbaikan Probabilitas Lexical Model

Setelah mendapatkan nilai awal dari korpus uji, maka langkah selanjutnya adalah melakukan perbaikan probabilitas *lexical model*. Perbaikan probabilitas *lexical model* dilakukan penulis dengan menggunakan aplikasi bantu yang telah dibuat sebelumnya.

Langkah awal perbaikan probabilitas *lexical model*, dilakukan perbaikan dengan kelipatan sekitar 50 perbaikan *lexical model* dengan menyimpannya pada tiap-tiap folder yang berbeda namun dengan nama file yang sama, yakni lex.f2e. Perbaikan *lexical model* yang dilakukan meliputi,

- menghapus kata yang tidak sesuai dengan arti seperti yang ditampilkan pada gambar 8,

Gambar 8. Tampilan menghapus

- mengubah probabilitas kata yang tidak sesuai yang ditunjukkan pada gambar 9.

Gambar 9. Tampilan mengubah

E. Pengujian Ulang Hasil Terjemahan Mesin Translasi

Langkah berikutnya adalah menguji kembali hasil terjemahan mesin translasi bahasa Jawa - bahasa Indonesia yang telah melewati perbaikan *lexical model*. Langkah pengujian dilakukan dengan cara melakukan pengujian otomatis 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 Jawa - bahasa Indonesia sebelum dan setelah melewati tahap perbaikan probabilitas *lexical model*.

Gambar 10. Tampilan grafik nilai BLEU sebelum dan setelah perbaikan *lexical model*

Gambar 10 merupakan tampilan grafik sebelum mengalami perbaikan probabilitas *lexical model*, nilai BLEU pada korpus uji 5108 sebesar 95.32% dan setelah dilakukan perbaikan probabilitas *lexical model* didapat nilai BLEU sebesar 95.61%. Terdapat peningkatan nilai BLEU sebesar 0.30% dilihat dari perbandingan sebelum dan sesudah mengalami perbaikan probabilitas *lexical model*.

2. Pengujian Ahli Bahasa

Pengujian ahli bahasa dilakukan terhadap hasil terjemahan mesin penerjemah statistik bahasa Jawa - bahasa Indonesia dengan mengambil kalimat yang mengalami perubahan pada hasil terjemahan otomatis yang terdapat pada korpus uji 5108 sebelum dan sesudah dilakukan perbaikan probabilitas *lexical model* 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 Perubahan <i>Lexical Model</i>	Sudarni, BA	C = 206, R = 234	88.03%
Setelah Perubahan <i>Lexical Model</i>	Sudarni, BA	C = 228, R = 234	97.44%

Tabel 1 merupakan tampilan tabel akurasi dari ahli bahasa sebelum mengalami perbaikan probabilitas *lexical model*, nilai dari ahli bahasa pada korpus uji 5108 sebesar 88.03% dan setelah dilakukan perbaikan probabilitas *lexical model* didapat nilai dari ahli bahasa sebesar 97.44%. Terdapat peningkatan nilai BLEU sebesar 10.69% dilihat dari perbandingan sebelum dan sesudah mengalami perbaikan probabilitas *lexical model*.

IV. KESIMPULAN

A. Kesimpulan

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

1. Mesin penerjemah statistik dapat diimplementasikan untuk menerjemahkan bahasa Jawa ke bahasa Indonesia.
2. Berdasarkan hasil penelitian, perbaikan probabilitas *lexical model* dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Jawa – bahasa Indonesia.
3. Persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Jawa – bahasa Indonesia yang dicapai sebesar 0.30% pada pengujian otomatis oleh BLEU.
4. Penilaian yang dilakukan oleh ahli bahasa menghasilkan persentase peningkatan akurasi hasil terjemahan sebesar 10.69%.

B. Saran

Beberapa saran yang dapat diberikan sebagai pengembangan dari penelitian ini adalah sebagai berikut.

1. Perlu penambahan jumlah korpus untuk meningkatkan kualitas terjemahan mesin penerjemah statistik.
2. Perlu adanya teknik dalam mencari *lexical model* yang diperbaiki mengingat banyaknya baris pada *lexical model*.
3. Perlu dilakukan penelitian lanjutan untuk melakukan analisis dalam menghasilkan terjemahan bahasa jawa ke bahasa Indonesia dengan mempertimbangkan hubungan antar frase dalam kalimat.
4. Perlu dilakukan pengujian terhadap korpus yang bukan merupakan bagian dari korpus paralel untuk mengetahui tingkat akurasi.
5. Perlu dilakukan pengecekan ulang terhadap korpus teks paralel untuk mencegah kesalahan penulisan (*typo*).

DAFTAR PUSTAKA

- [1] Na'im, Akhsan., Syaputra, Hendri. 2010. Hasil Sensus Penduduk 2010: Kewarganegaraan, Suku Bangsa, Agama, dan Bahasa Sehari-Hari Penduduk Indonesia. Jakarta: Badan Pusat Statistik.
- [2] Republik Indonesia. 2010. Keputusan Gubernur Jawa Tengah Nomor 423.5/5/2010 tentang Kurikulum Mata Pelajaran Muatan Lokal (Bahasa Jawa) untuk Jenjang Pendidikan SD/ SDLB/ MI, SMP/ SMPLB/ MTs Negeri dan Swasta. Semarang: Pemerintah Provinsi Jawa Tengah.
- [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] Sujaini, Herry., Negara, Arif Bijaksana Putra. 2015. Analysis of Extended Word Similarity Clustering based Algorithm on Cognate Language. Gujarat: ESRSA Publications Pvt. Ltd.
- [6] 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.
- [7] McEnery, T., et al. 2006. Corpus-Based Language Studies: An Advanced Resource Book. Oxon: Routledge.
- [8] Hunston, S. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.
- [9] 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).