MESIN PENERJEMAH BAHASA INDONESIA- BAHASA JAWA

Johan Pranata[1], Muljono[2]
[1,2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Nakula I No. 5-11, Semarang, 50131, (024) 3517261
E-mail : johanpandawa@gmail.com, gagaksinaga@gmail.com

Abstrak

Bahasa Jawa yang dulu merupakan bahasa yang besar, dengan bertambahnya waktu, penggunaannya semakin berkurang. Untuk mencegah terjadinya kepunahan bahasa atau biasa disebut dengan istilah kematian bahasa, maka sudah seharusnya masyarakat menyadari pentingnya pemeliharaan bahasa daerah. Di antaranya meliputi pemekaran kosakata dan kodifikasi berupa penyusunan kamus. Dalam era ini dengan pendekatan teknologi, kamus dapat dikembangkan dalam bentuk mesin penerjemah. Penelitian ini mengembangkan mesin penerjemah statistik berbasis frasa, yaitu yang merupakan suatu paradigma dari mesin penerjemah dimana penerjemahan dilakukan berbasis model statistik dengan parameter-parameter yang diturunkan dari analisis paralel korpus. Penelitian ini menggunakan 4.500 pasang kalimat sebagai korpus paralel yang bersumber dari Alkitab. Hasil dari implementasi mesin penerjemah statistik berbasis frasa ini diuji dan dievaluasi menggunakan evaluasi otomatis, BLEU yang menghasilkan nilai evaluasi sebesar 42,78%.

Kata Kunci: *mesin penerjemah, statistik, bahasa Indonesia, bahasa Jawa, basis frasa*

Abstract

Java language, which used to be a great language, with time, its use decreases. To prevent the extinction of a language or commonly referred to as the death of a language, then it should be public aware of the importance of maintenance of regional languages. Among these include the expansion of vocabulary and codification in the form of a dictionary compilation. In this era of technological approach, the dictionary can be developed in the form of machine translation. This study developed a phrase-based statistical machine translation, which is a paradigm of machine translation where translation is done based on a statistical model with parameters derived from the analysis of the parallel corpus. This research using 4,500 pairs of sentences as a parallel corpus derived from the Bible. The results of the implementation of phrase-based statistical machine translation is tested and evaluated using automated evaluation, BLEU which generates the evaluation value of 42.78%.

Keywords: *machine translation, statistical, Indonesian, Javanese, phrase-based*

1. PENDAHULUAN

Bahasa Jawa adaah suatu bahasa komunikasi yang digunakan secara khusus di lingkungan etnis Jawa. Bahasa ini merupakan bahasa pergaulan, yanng digunakan untuk berinterksi antarindividu dan memungkinkan terjadinnya komunikasi dan perpindahan informasi. Bahasa Jawa yang dulu merupakan bahasa yang besar, dengan bertambahnya waktu, pennggunannya semakin berkurang.

Saat ini para kaum muda di Pulau Jawa, khususnnya yang masih di usia sekolah, sebagian besar tidak menguasai bahasa Jawa. Hal ini disebabkan oleh gencarnnya serbuan beragam budaya asing dan arus informasi yang masuk melalui bermacam sarana seperti televisi dan lain-lain.

Hal ini tidak dapat menghindarkan terjadinya kepunahan bahasa, yaitu hilangnya bahasa daerah akibat drai penuturnya menggunakan bahasa lain. Untuk mencegah terjadinya kepunahan

bahasa, maka sudah seharusnya masyarakat khususnya etnis Jawa menyadari pentingnya pemeliharaan bahasa daerah.

Mesin penerjemah merupakan alat yang dapat digunakan dalam pemeliharaan bahasa sebagai wujud kesadaran akan terjadinya kepunahan bahasa daerah.

Dalam pembangunan mesin penerjemah dalam penelitian ini akan digunakan metode statistic-based yang merupakan paradigma dari mesin penerjemah dimana penerjemahan dilakukan berbasis model statistik dengan parameter-parameter yang diturunkan dari analisis paralel korpus.

Dalam pembangunannya, mesin penerjemah statistik ini disusun oleh beberapa komponen utama yaitu Languange Model, Translation Model dan decoder.

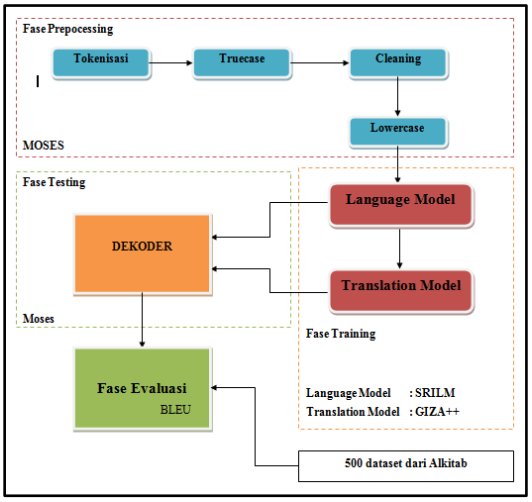
2. METODE

Pada penelitian ini penulis menggunakan metode statistik berbasis frasa. Ide utama dari pendekatan ini adalah terjemahan akan dibuat dari kata yang paling mungkin diterjemahkan. Mesin ini mengasmsikan bahwa setiap kalimat T pada bahasa target merupakan sebuah kemungkinan hasil terjemahan dari kalimat S pada bahasa sumber. Melalui pendekatan bahwa teks yang diterjemahkan berdasarkan distribusi probabilitas P(S|T) dapat dilakukan dengan teorema Bayes :

$$P(T|S) = P(T).P(S|T) \quad (1)$$

Sistem ini terdiri dari tiga komponen penting yaitu, Languange Model (LM) untuk menghitung probabilitas bahasa target ‘T’ sebagai probabilitas P(T). Translation Model (TM) membantu menghitung probabillitas bersyarat kalimat target yang diberi aturan sumber, P (T|S). Decoder memaksimalkan hasil probabilitas dari LM dan TM, pada tahap ini proses penerjemahan dilakukan [1].

Berikut merupakan arsitektur penyususnan mesin penerjemah dengan metode statistik berbasis frasa menggunakan 4,500 dataset corpus paralel yang berumber dari Alkitab :



Gambar 1. Arsitektur Penyusunan Mesin Penerjemah.

2.1 Fase Preprocessing

Sebelum melakukan proses dengan Language Model (LM), Translation Model (TM), dan decoder, data mentah corpus paralel harus melalui tahap preprocessing meliputi :

1. Proses menata kedua bahasa dalam korpus paralel menjadi susunan baris-baris kalimat.
2. Tokenisasi, proses memberi jarak antarkata, termasuk juga memberi jarak kata dengan tanda baca yang ada.
3. Truecasing, merupakan proses setiap awal kata dari tiap kalimat dikonversi ke tempat yang paling mungkin.
4. Cleaning, proses yang memberi batas maksimal pada panjang kalimat.
5. Lowercase, yaitu proses untuk

menyeragamkan besar-kecilnya huruf.

2.2 Fase Training

Setelah melalui tahap preprocessing, pada tahap selanjutnnya dilakukan beberapa proses melalui komponen utama dalam mesin penerjemah ini, yaitu Languange Model (LM) dan Translation Model (TM).

2.2.1 Languange Model (LM)

Language Model (LM) merupakan proses untuk menghitung probabilitas kalimat. Untuk menghitung probabilitas kalimat, diperlukan untuk menghitung probabilitas kata, mengingat urutan kata yang mendahuluinya dengan aturan rantai seperti sebagai berikut:

$$\mathbf{P(K) = P\ (w_1 w_2 \dots w_n)} \qquad (2)$$

Dengan 'K' merupakan notasi 'sentence' dan 'w' merupakan notasi 'word' [2].

Rumusan tersebut dikenal dengan sebutan n-gram model. Probabillitas bersyarat dapat dihitung dari jumlah frekuensi n-gram [1] :

$$\mathbf{P\ (w_n | w_{n-1}) = \frac{count\ (w_{n-1}\ w_n)}{count\ (w_{n-1})}} \qquad (3)$$

Terdapat tiga model n-gram yang diperoleh dari proses language mmodel, yaitu unigram, bigram, dan trigram. Unigram adalah kemunculan kata yanng tidak dipengaruhi kata lain. Bigram merupakan kemunculan setiap kata yang dipengaruhi oleh kata lain. Model trigram adalah kemunculan sebuah kata yang dipengaruhi oleh kata sebelumnya.

2.2.2 Translation Model (TM)

Translation Model (TM) merupakan tahap untuk mencari ketepatan. P(S|T) digunakan sebagai notasi yang menunjukkan TM. Peluang untuk mendapatkan translation model, ditunjukkan dalam persamaan sebagai berikut :

$$P(S|T) = P(s_1|t_1)P(s_2|t_2) \dots P(s_n|t_n) \qquad (4)$$

Dengan, s_n adalah kata dari kalimat sumber pada posisi ke-n yang akan diketahui peluangnya, t_n adalah kata kalimat target pada posisi ke-n yang menerjemahkan kata s_n [3].

Ada beberapa macam metode TM yang dapat diterapkan yakni pendekatan berbasis kata (*word based*), berbasis frasa (*phrase based*), dan kombinasi keduanya. Pendekatan berbasis kata sering kali tidak mampu menerjemahkan dengan baik konteks lokal suatu bahasa. Dalam TM berbasis frase, prosesnya dapat dibagi kedalam tiga bagian. Pertama, membuat kalimat sumber menjadi sebuah tabel frasa. Kedua, menerjemahkan setiap frase kedalam bahasa target. Kemudian, dilakukan tahap *reordering* [2].

2.3 Fase Testing

Dalam fase testing ini komponen yang digunakan adalah decoder. Decoder bertugas menemukan teks dalam bahasa target yang memiliki probabilitas paling besar dengan pertimbangan faktor translation model dan language model [1]. Decoder disebut juga sebagai alogaritma pencarian. Bentuk matematikanya adalah sebagai berikut :

$$\hat{e} = arg_e max(S|T)P(T) \qquad (5)$$

Dengan $arg_e max$ adalah pencarian nilai probabilitas terbesar yang diperoleh dari language model dan translation model.

2.4 Fase Evaluasi

Suatu permasalahan jika mencoba menerjemahkan satu kalimat dengan menggunakan beberapa mesin penerjemah, akan diperoleh berbagai jawaban yang berbeda. Mempertimbangkan hal-hal tersebut

maka setiap pembangunan mesin penerjemah dibutuhkan tahap evaluasi terhadap mesin penerjemah tersebut. Dalam penelitian ini penulis menggunakan BLEU (*Billingual Evaluation Understudy*). BLEU bekerja dengan cara mengukur skor presisi dari n-gram termodifikasi (*modified n-gram precision score*) antara terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty* (BP).

3. HASIL DAN PEMBAHASAN

Hasil dari preprocessing 4.500 corpus paralel menunjukkan bahwa hanya 1 kalimat yang terbuang atau sebesar 0.022%. Proses terbuangnya ini dikarenakan jumlah kata yang ada pada korpus Bahasa Indonesia dan Bahasa Jawa yang diambil dari alkitab memiliki panjang kalimat yang mendekati presisi atau presisi.

Kemudian dalam tahap selanjutnya, yaitu fase training dihasilkan sebuah berkas konfigurasi mesin penerjemah statistik dengan nama *moses.ini*, file ini digunakan untuk proses *decoding* atau penerjemah kalimat. Setelah didapatkan berkas configurasi, maka proses penerjemahan dapat dilakukan.

Kualitas terjemahan dari mesin penerjemah dalam penelitian ini diuji menggunakan alat ukur BLEU secara otomatis. Proses evaluasi dilakukan dengan membandingkan hasil terjemahan dari mesin penerjemah dengan pembanding korpus atau dataset lain yang terpisah dari data training berjumlah 500 baris kalimat yang telah diolah atau melalui tahap *tokenize, lowercase, cleaning*,dan *truecasing*.

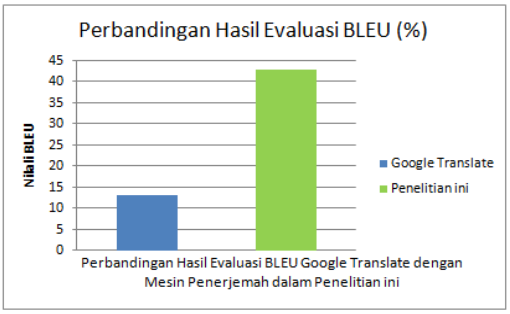
Tabel 1: Hasil Evaluasi BLEU

	Hasil Evaluasi BLEU		
Jumlah	Brevity	Ratio	Nilali
Dataset	Penalty		Akhir (%)
4500	1,000	1,006	42,78

Tabel 2: Perbandingan Hasil Evaluasi Google Translate dengan Penelitian ini

Peneliti	Jenis Penerjemahan	Hasil BLEU (%)
Google Translate	Bahasa Indonesia –Jawa	12,99 %
Penelitian ini	Bahasa Indonesia –Jawa	42,78%

Dari Tabel 1 dan Tabel 2 dapat dilihat bahwa mesin penerjemah dalam penelitian ini memiliki nilai evaluasi sebesar 42, 78%, sedangkan Google Translate memiliki nilai evaluasi 12,99% untuk penerjemahan Bahasa Indonesia-Bahasa Jawa.



Gambar 2: Perbandingan Hasil Evaluasi BLEU Google Translate dengan Mesin Penerjemah dalam Penelitian ini

Jika dibandingkan dengan mesin penerjemah lain, dalam hal ini adalah Google translate yang merupakan mesin penerjemah yang biasa digunakan oleh kalangan umum, dapat dilihat bahwa mesin penerjemah dalam penelitian ini memiliki kualitas yang lebih baik dibandingkan dengan Google Translate, dengan rentang nilai yang cukup tinggi.

4. KESIMPULAN DAN SARAN

Mesin penerjemah statistik dapat diimplementasikan untuk melakukan penerjemahan bahasa Indonesia-bahasa Jawa. Berdasarkan hasil pengujian, skor

BLEU pada korpus uji 4.500 sebesar 42,78%.

Penambahan korpus paralel yang lebih banyak diperlukan untuk menghasilkan mesin penerjemah yang lebih baik dan memiliki nilai evaluasi yang tinggi. Diperlukan pula pengecekan ulang terhadap korpus paralel karena kesejajaran korpus paralel harus dijaga.

DAFTAR PUSTAKA

[1] Andri Hidayat, "Aplikasi Penerjemah Dua Arah Bahasa Indonesia – Bahasa Melayu Sambas Berbasis Web dengan Menggunakan Decoder Moses," Teknik Informatika Universitas Tanjung Pura, Tnjung Pura, Skripsi 2011.

[2] P.Koehn, "MOSES," in *Statistical Machine Translation System, User Manual and Code Guide*. Cambridge, London: Cambridge University, 2010, p. 245.

[3] Rizky Aditya Nugroho, "Penerjemahan Bahasa Indonesia dan Bahasa Jawa dengan Pendekatan Statistik," *Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015)*, Maret 2015.