BAB III

METODOLOGI PENELITIAN

1.1 Instrumen penelitian

Pada instrumen penelitian akan dijelaskan kebutuhan perangkat lunak, kebutuhan perangkat keras, serta kebutuhan bahan penelitian. Yaitu sebagai berikut :

1.1.1 Kebutuhan perangkat lunak

Penelitian ini membutuhkan sekumpulan perangkat lunak, karena perangkat lunak merupakan faktor yang penting dan harus terpenuhi dalam penelitian ini, sehingga maksut dan tujuan penelitian dapat dicapai. perangkat lunak yang digunakan adalah :

1. Sistem Operasi

   Sistem operasi yang digunakan dalam penelitian ini adalah Ubuntu 14.04 LTS – 64 bit. Digunakan untuk menjalankan tools pembuatan mesin penerjemah statistik.

2. SRILM

   Perangkat lunak SRILM (The SRI Language Modeling) yang dikembangkan oleh SRI Speech Technology and Research Laboratory adalah toolkit untuk pemodelan n-gram pada language model. SRILM yang digunakan adalah SRILM versi 1.7.1 .

3. GIZA++

   Perangkat lunak GIZA++ dikembangkan oleh franz josef och. GIZA++ digunakan pada translation model, khususnya digunakan untuk proses penjajaran kata. GIZA++ yang digunakan adalah versi 1.0.7 .

4. Moses

   Perangkat lunak Moses dikembangkan oleh Hieu Hoang dan Philipp Koehn ,perangkat ini merupakan perangkat yang bersifat terbuka. Moses digunakan untuk prepocessing dan decoding. Moses yang digunakan adalah versi 2.1.1 .

1.1.2 Kebutuhan perangkat keras

Selain kebutuhan perangkat lunak, dibutuhkan juga perangkat keras guna mendukung penelitian ini. Adapun kebutuhan hardware yang akan digunakan untuk membangun sistem tersebut memiliki spesifikasi sebagai berikut :

1. Processor : Intel(R) Core(TM) i5-3210M CPU @ 2.50Ghz (4 CPUs)
2. Memory (RAM) : 4096MB

1.1.3 Kebutuhan bahan penelitian

Mesin penerjemah statistik memerlukan sekumpulan kalimat untuk pembuatan korpus paralel. Data atau kumpulan kalimat paralel yang digunakan sebagai pelatihan pada penelitian ini diambil dari Alkitab terjemahan Jawa dan Indonesia. Korpus Indonesia dan Korpus Jawa masing-masing memiliki 4.500 kalimat atau dataset. Pada penelitian ini juga membutuhkan 500 kalimat atau dataset digunakan untuk proses evaluasi.

1.2 Prosedur pengumpulan data

1. Pembelajaran literatur

   Pembelajaran literatur adalah proses mempelajari literer-literer yang berkaitan dengan penelitian yang dilakukan. Pembelajaran literasi juga

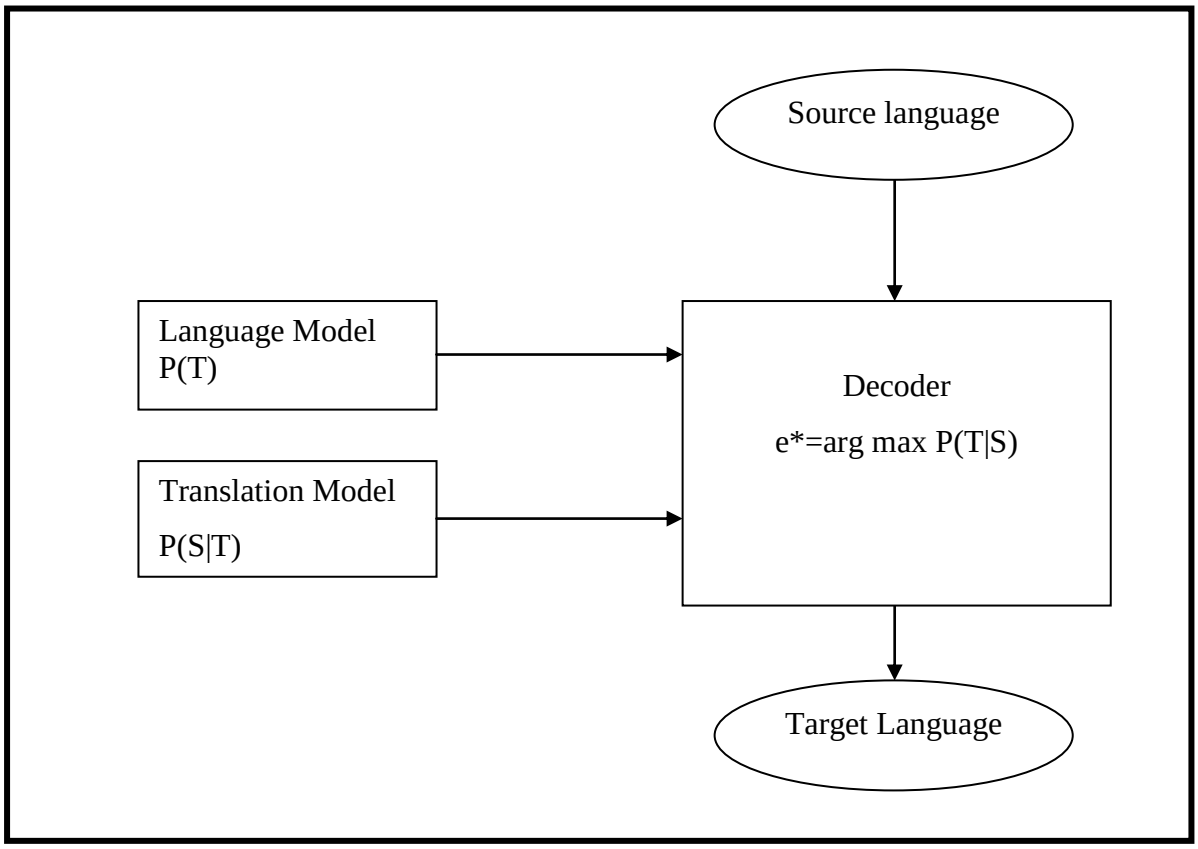
merupakan sebuah proses dalam mengevaluasi penelitian-penelitian yang pernah dilakukan dalam bidang yang sama. Dalam proses ini pula pembelajaran terhadap korpus Bahasa Indonesia-Jawa dilakukan kemudian dikumpulkan sebagai bahan pembuatan sistem.

2. Dokumentasi

   Data yang didapatkan dari pembelajaran literature kemudian dijadikan sebagai acuan dalam membangun sistem, kemudian setiap langkah yang dilakukan didokumentasikan sebagai bahan pembahasan.

1.3 Metode yang diusulkan

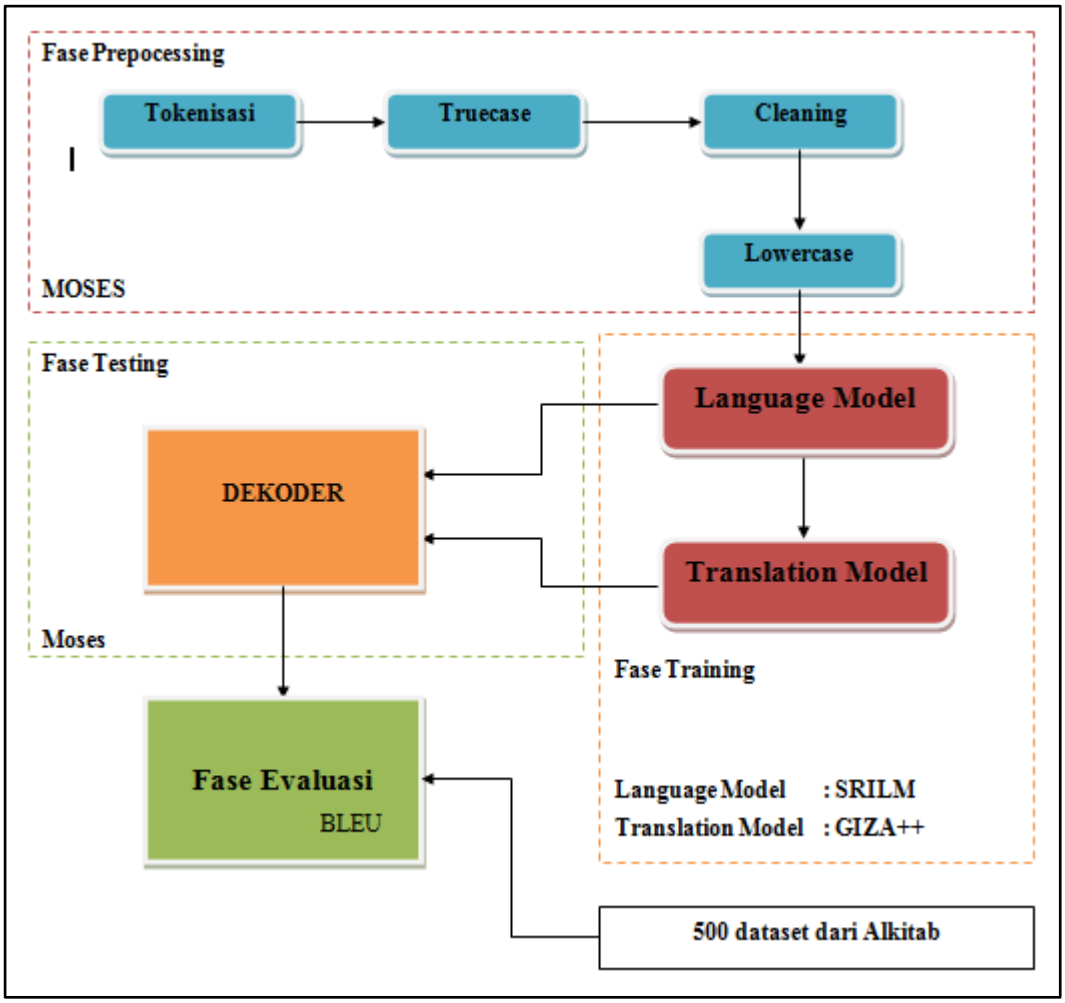
Untuk pembuatan mesin penerjemah pada penelitian ini penulis menggunakan metode mesin penerjemah statistik berbasis frasa. Karena metode ini dianggap sebagai metode era baru untuk mesin penerjemah dibandingkan dengan metode lainnya seperti metode Direct dan Rule-based. Ide utama dari pendeketan ini adalah terjemahan akan dibuat dari kata yang paling mungkin di terjemahkan. Sistem ini terdiri dari tiga komponen yang berbeda. Language Model (LM) menghitung probabilitas bahasa target ‘T’ sebagai probabilitas P (T), pada language model (LM) dilakukan proses penghalusan untuk mendapatkan model n-gram. Translation Model (TM), membantu untuk menghitung probabilitas bersyarat kalimat target diberi aturan sumber, P (T | S). pada translation model terdapat beberapa elemen seperti penjajaran (alignment) dan extraction frasa (phrase extraction). Decoder memaksimalkan hasil probabilitas dari Language Model (LM) dan Translation Model (TM) pada model ini proses penerjemahan dilakukan. Untuk penjelasan dasar Mesin Penerjemah Statistik Berbasis Korpus ditujukan pada gambar berikut :



Gambar 1 : Dasar mesin penerjemah statistik berbasis korpus

1.4 Teknik analisis data

Teknik analisis data yang digunakan dalam pembangunan mesin penerjemah terdiri dari beberapa tahapan. Tahapan pertama prepocessing kemudian beralih ke tahapan training. Tahapan prepocessing adalah tahapan mempersiapkan korpus paralel. Pada tahapan training dilakukan proses mengolah korpus untuk memperoleh model bahasa dan model penerjemahan. Setelah tahapan training memperoleh hasil lalu melakukan proses testing atau disebut juga dengan tahapan decoding. Dan yang terakhir adalah evaluasi . Berikut gambaran arsitektur pembangunan mesin penerjemah statistik :



Gambar 2 : Teknik analisis pembangunan mesin penerjemah statistik

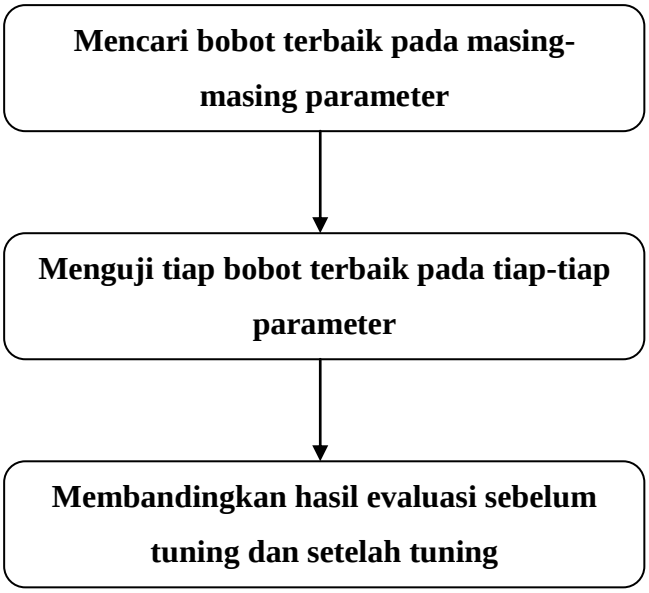
1. Penjajaran kalimat adalah proses menata kedua bahasa tersebut menjadi susunan baris-baris kalimat.
2. *Tokenisasi* adalah proses memberi jarak antar kata, termasuk juga memberi jarak kata dengan tanda baca yang ada.
3. *Cleaning* adalah proses yang memberi batas maksimal pada panjang kalimat , *lowercase* merupakan proses untuk menyeragamkan besar-kecilnya huruf.
4. *Truecasing* adalah proses setiap kata awal dari tiap kalimat dikonversi ke tempat yang paling mungkin.

5. *Language model* adalah proses pembentukan n-gram, penghalusan (*smoothing*), dan penambahan simbol sebagai batas (*sentences boundary symbols*) pada tiap kalimat.
6. *Translation Model* adalah proses menjajarkan kata pada korpus paralel, membuat tabel frasa (*phrase table*), pembuatan *tabel lexicalized reordering* atau pemanggilan kembali kata maupun frasa yang telah dibuat , pemberian score pada frasa, dan terakhir membuat berkas configuration file.
7. Decoder adalah proses menerjemahkan bahasa sumber ke bahasa target sesuai dengan masukan kalimat yang ditulis.

1.5 Teknik *tuning*

Pada pembuatan mesin penerjemah statistik terdapat dua macam proses *tuning* yaitu tuning otomatis dan *tuning* manual. Proses yang dilakukan dalam penelitian ini adalah proses *tuning* yang dilakukan secara manual. Yang dilakukan dalam proses *tuning* manual adalah mecari bobot terbaik dari masing-masing parameter yang ada pada mesin penerjemah. Parameter tersebut antara lain, *phrase translation model, language model, reordering model,*dan *word pinalty*. Proses *tuning* dilakukan setelah proses evaluasi terhadap hasil terjemahan dan hasil evaluasi dari mesin penerjemah statistik bahasa Indonesia-bahasa Jawa selesai dikerjakan.

Pencarian bobot terbaik dilakukan dengan uji-coba tiap tiap nilai bobot dari tiap-tiap parameter. Parameter phrase translation,language model, reordering model mempunyai rentang nilai bobot yaitu 0,1 hingga 1. Untuk parameter word penalty mempunyai rentang -3 hingga 3. Setelah didapatkan bobot terbaik dari masing-masing parameter maka dilakukan uji-coba dengan menggunakan bobot terbaik masing-masing parameter. Dan proses terakhir adalah membandingkan hasil evaluasi setelah tuning dan sebelum tuning pada mesin penerjemah statistik bahasa Indonesia-bahasa Jawa.


Mencari bobot terbaik pada masing-masing parameter
Menguji tiap bobot terbaik pada tiap-tiap parameter
Membandingkan hasil evaluasi sebelum tuning dan setelah tuning