Bab 1 Pendahuluan

1.1 Latar Belakang

Korpus paralel merupakan kumpulan teks yang disusun sejajar dengan teks terjemahannya pada satu atau lebih bahasa lain. Korpus paralel penting dan mempunyai pengaruh yang cukup besar dalam pengembangan *machine translation* (MT) dan *natural language processing* (NLP) [1]. *Statistical machine translation* (SMT) yang merupakan salah satu bagian dari MT, memanfaatkan korpus paralel untuk melakukan proses *learning* yang dibutuhkan untuk mendapatkan keterkaitan antara teks bahasa *source* (misalnya bahasa Sunda) dan teks bahasa *target* (misalnya bahasa Indonesia) dengan penerapan *translation model* dan *language model*. Pemanfaatan korpus paralel tersebut membuat sistem SMT sangat bergantung pada kuantitas, kualitas, dan domain dari data latih yang digunakan [2].

Korpus paralel untuk bahasa-bahasa yang umum digunakan atau bahasa yang masih dikembangkan secara aktif dan sering menjadi objek penelitian seperti bahasa Inggris atau beberapa bahasa lain yang umum digunakan di Eropa bisa dengan mudah didapatkan dengan memanfaatkan korpus paralel yang sudah disusun seperti Europarl[1], korpus yang dikembangkan oleh Joint Research Centre[2], OPUS[3] atau dari LDC[4]. Hal ini berbeda jauh dengan ketersediaan korpus paralel untuk bahasa-bahasa yang kurang dominan digunakan karena korpus paralel yang tersedia sangat sedikit atau bahkan tidak ada sama sekali. Pada penelitian ini, akan dilakukan pembentukan korpus paralel pasangan bahasa Sunda dan bahasa Indonesia untuk keperluan pengembangan sistem SMT.

Ketersediaan korpus yang bisa dijadikan sebagai kandidat korpus paralel untuk bahasa Sunda masih sangat sedikit, oleh karena itu proses pengumpulan akan memanfaatkan Wikipedia yang merupakan salah satu sumber korpus multibahasa yang besar dan dapat diakses secara bebas [3]. Wikipedia juga dipilih karena menyediakan *interlanguage links* yang menyediakan tautan artikel sebanding di dalam Wikipedia dalam bahasa yang berbeda. Pemanfaatan fasilitas ini akan memberikan tambahan informasi terkait pasangan artikel Wikipedia yang ditulis dalam bahasa Sunda dan juga bahasa Indonesia dan memiliki topik yang serupa atau masih berada pada domain yang sama. Fasilitas tersebut akan digunakan dalam proses *alignment* pada tingkat dokumen.

Kalimat-kalimat pada tiap artikel yang akan di bandingkan kemiripannya direpresentasikan ke dalam bentuk vektor kata agar proses perhitungan menjadi lebih sederhana [19, 20]. Perhitungan *similarity* antar vektor tidak bisa dilakukan hanya dengan melihat panjang vektor kata karena kalimat yang dipasangkan tidak memiliki panjang yang sama. Penelitian ini menggunakan *cosine similarity* untuk menghitung *similarity* antar kalimat yang direpresentasikan dalam bentuk vektor kata dengan cara mengukur besar sudut dua vektor yang dibandingkan dan bukan

[1] http://www.statmt.org/europarl/
[2] https://ec.europa.eu/jrc/en/language-technologies
[3] http://opus.lingfil.uu.se/
[4] https://www.ldc.upenn.edu/

panjang dari vektor yang dibandingkan. Hal ini dilakukan karena panjang kalimat pada korpus nonparalel seperti Wikipedia tidak seimbang [5].

Pengumpulan kalimat paralel dilakukan dengan memanfaatkan kamus kosakata bahasa Sunda-Indonesia (*bilingual lexicon*) untuk membantu perhitungan *cosine similarity* antar kandidat kalimat paralel. Kosakata baru yang belum ada pada *bilingual lexicon* awal akan diperbarui melalui iterasi dengan menerapkan metode *bootstrapping* yang dibantu dengan IBM Model 4 EM *learner* pada *tool* GIZA++ [4]. *Tool* GIZA++ akan digunakan dalam proses *learning* pada kandidat kalimat paralel yang dihasilkan di akhir setiap iterasi untuk mendapatkan pasangan kata terjemahan yang baru. Iterasi akan berhenti ketika kondisi konvergensi tercapai. Metode *bootstrapping* ini terbukti mampu membantu proses ektraksi kalimat paralel dari korpus nonparalel dengan akurasi yang lebih tinggi jika dibandingkan dengan hanya menggunakan IBM Model 4 EM [5].

1.2 Rumusan Masalah

Mengacu pada latar belakang yang sudah dijelaskan sebelumnya, rumusan masalah dari penelitian ini adalah:

1. Bagaimana membangun korpus paralel bahasa Sunda-bahasa Indonesia dengan memanfaatkan Wikipedia dan menggunakan metode *bootstrapping* dengan bantuan IBM Model 4 EM.
2. Bagaimana akurasi dari korpus paralel bahasa Sunda-bahasa Indonesia yang dihasilkan menggunakan metode *bootstrapping* dan IBM Model 4 EM.

1.3 Tujuan

Mengacu pada rumusan masalah yang dijelaskan sebelumnya, tujuan dari penelitian ini adalah:

1. Membangun korpus paralel bahasa Sunda-bahasa Indonesia dengan memanfaatkan artikel Wikipedia dan menggunakan metode *bootstrapping* dengan bantuan IBM Model 4 EM.
2. Melakukan perhitungan untuk mengukur akurasi dari pasangan kalimat pada korpus paralel bahasa Sunda-bahasa Indonesia yang dihasilkan menggunakan metode *bootstrapping* dan IBM Model 4 EM.

1.4 Batasan Masalah

Batasan masalah yang ada selama pelaksanaan penelitian ini adalah:

1. Artikel Wikipedia yang digunakan hanyalah artikel dalam bahasa Sunda yang memiliki artikel yang sama dalam bahasa Indonesia.
2. *Tool* GIZA++ akan digunakan untuk implementasi *word alignment* IBM Model 4 EM.
3. Hasil akhir dari penelitian ini berupa korpus paralel bahasa Sunda-bahasa Indonesia dengan topik bebas.
4. Pengukuran *similarity* antar kandidat kalimat paralel menggunakan metode *cosine similarity* dengan metode pembobotan *inverse document frequency* yang sudah dinormalisasi.

5. Tidak dilakukan analisis morfologi pada kata bahasa Sunda maupun kata bahasa Indonesia sehingga kata "makan" dan "memakan" akan dianggap sebagai dua *token* yang berbeda. Analisis *part-of-speech* pada kalimat pun juga tidak dilakukan.
6. Pengujian hasil kalimat paralel yang dihasilkan akan menggunakan bantuan penutur asli bahasa Sunda yang juga mengerti bahasa Indonesia.
7. Proses *alignment* yang dilakukan pada kalimat paralel yang dihasilkan, akan dilakukan sampai tingkat kalimat saja.

1.5 Metode Penelitian

Metodologi yang dilakukan untuk menyelesaikan permasalahan dalam penelitian ini adalah:

1. Studi Literatur
   Melakukan pencarian dan pengkajian teori dari teknik atau metode yang digunakan pada penelitian ini beserta informasi pendukung dari berbagai sumber seperti buku teks, jurnal ilmiah, dan Internet yang terkait pemanfaatan Wikipedia sebagai sumber korpus, pembuatan korpus paralel, proses *alignment* kata atau kalimat, dan pengukuran akurasi antar kalimat paralel.
2. Pengumpulan Data dan *Preprocessing* Data
   Melakukan pengumpulan data memanfaatkan *application programming interface* (API) yang disediakan MediaWiki untuk mendapatkan teks atau artikel berbahasa Sunda yang memiliki atikel setopik dalam bahasa Indonesia pada Wikipedia. *Preprocessing* dilakukan agar data yang didapatkan bisa diproses pada sistem.
3. Analisis dan Perancangan Sistem
   Melakukan proses analisis dan perancangan terhadap sistem yang akan digunakan untuk membuat korpus paralel dengan mengacu pada hasil studi literatur dan data yang ada.
4. Implementasi
   Mengimplementasikan sistem yang digunakan untuk proses pembangunan korpus paralel untuk bahasa Sunda dan bahasa Indonesia menggunakan metode *bootstrapping* bersama dengan IBM Model 4 EM.
5. Pengujian dan Analisis
   Pengujian dilakukan terhadap korpus paralel yang sudah dibuat untuk melihat akurasi dari korpus paralel tersebut. Analisis dilakukan dengan melihat hasil dari pengujian.
6. Pembuatan Laporan
   Melakukan pembuatan laporan akhir penelitian yang menjelaskan proses penyelesaian masalah beserta hasil akhir dan analisisnya.