Pembentukan *Sentence-Aligned* Korpus Paralel untuk Bahasa Sunda-Bahasa Indonesia Berbasis Wikipedia dengan *Bootstrapping* dan EM

Building Sentence-Aligned Parallel Corpus for Sundanese-Indonesian Language Based on Wikipedia using Bootstrapping and EM

Ignasius Indra Kusuma Wijaya[1], Arie Ardiyanti Suryani[2], Kurniawan Nur Ramadhani[3]

[1,2,3]Program Studi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom
[1]ignasiusindra@students.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id,
[3]kurniawannr@telkomuniversity.ac.id

Abstrak

Ketersediaan korpus paralel pada pasangan bahasa Sunda-Indonesia masih sangat sedikit. Korpus paralel tersebut penting dan bisa dimanfaatkan sebagai sumber data latih dalam sistem *machine translation* atau sistem *natural language processing*. Penelitian ini mencoba untuk mengumpulkan kalimat paralel yang didapatkan dari pasangan artikel Wikipedia berbahasa Sunda dan berbahasa Indonesia menggunakan fasilitas *interlanguage links*. Sebuah *bilingual lexicon* dan beberapa filter yang berdasarkan pada kemunculan kata, panjang kalimat dan *word overlap* antar kalimat digunakan untuk mendapatkan kalimat paralel. Metode *bootstrapping* kemudian digunakan untuk meningkatkan kualitas kalimat paralel dengan cara memperbarui *bilingual lexicon* memanfaatkan IBM Model 4 *expectation maximization* (EM) *learner* di dalam *tool* GIZA++. GIZA++ dijalankan pada kandidat kalimat paralel yang dihasilkan di setiap iterasi sistem sampai kondisi konvergensi tercapai. Hasil evaluasi manual menggunakan penilaian manusia menunjukkan bahwa 79,5% dari korpus paralel hasil bentukan sistem terbukti paralel.

Kata kunci: korpus paralel, Wikipedia, *bootstrapping, expectation maximization*

Abstract

The availability of Sundanese-Indonesian parallel corpus are few in number. Parallel corpus is important and could be used as a training data source for machine translation system or natural language processing system. This work is trying to collect parallel sentences extracted from pairs of Wikipedia articles using interlanguage links facility. A bilingual lexicon and a series of filters based on word occurrence, sentence length and word overlap between sentences were used to collect parallel sentence. The bootstrapping method was used to improve the quality of the parallel sentence by updating the bilingual lexicon using IBM Model 4 EM learner implemented in GIZA++ tool. GIZA++ was run on parallel sentence candidate which was generated in each iteration until the system reached convergence state. Manual evaluation result using human judgement shows that 79,5% of parallel corpus built by the system had proven to be parallel.

Keywords: parallel corpus, Wikipedia, bootstrapping, expectation maximization

1. Pendahuluan

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]. Korpus paralel pada bahasa-bahasa yang dominan digunakan pada penelitian seperti bahasa Inggris atau beberapa bahasa dari benua Eropa bisa didapatkan dengan mudah. Hal ini berbanding terbalik dengan bahasa lain yang kurang dominan seperti bahasa daerah. Korpus paralel dari bahasa-bahasa tersebut sulit didapatkan karena jumlahnya sangat sedikit atau bahkan tidak ada sama sekali. Penelitian ini mencoba untuk membentuk korpus paralel pasangan bahasa Sunda dan bahasa Indonesia untuk keperluan pengembangan sistem SMT. Pembentukan korpus paralel membutuhkan sumber korpus awal. Ketersediaan sumber korpus yang bisa dijadikan sebagai kandidat korpus paralel untuk bahasa Sunda dan bahasa Indonesia 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 menyediakan fasilitas *interlanguage links* yang bisa dimanfaatkan untuk menemukan pasangan artikel yang memiliki keterkaitan topik. Fasilitas *interlanguage links* memberikan tautan dari suatu artikel Wikipedia berbahasa tertentu ke artikel berbahasa lain yang memiliki artikel dengan topik yang sama.

Penelitian ini dilakukan dengan tujuan untuk membentuk korpus paralel pasangan bahasa Sunda dan bahasa Indonesia. Proses pembentukan korpus paralel diawali dengan melakukan *alignment* pada tingkat dokumen dengan memanfaatkan fasilitas *interlanguage links* pada Wikipedia. Tiap kalimat dari pasangan artikel Wikipedia berbahasa Sunda dan Indonesia kemudian dicari pasangan kalimat yang paralel dengan menggunakan pengukuran

cosine similarity dan filter berbasis panjang kalimat. Proses pengumpulan kalimat paralel dilakukan dengan memanfaatkan kamus kosakata bahasa Sunda-Indonesia (*bilingual lexicon*). Metode *bootstrapping* digunakan untuk memperbarui *bilingual lexicon* dengan kosakata baru yang didapatkan dari penggunaan IBM Model 4 EM *learner* pada *tool* GIZA++. Hasil korpus paralel yang dibentuk kemudian diuji melalui penilaian manusia dan juga dengan menggunakan sistem SMT Moses untuk melihat seberapa baik kualitas korpus paralel tersebut.

2. Penelitian Terkait

Penelitian ini menggunakan konsep serupa yang sudah pernah dilakukan sebelumnya oleh Fung dan Cheung [4]. Mereka menggunakan sumber korpus yang berupa transkrip berita berbahasa Mandarin dan Inggris dengan tujuan untuk mendapatkan korpus paralel dari korpus awal yang sangat tidak paralel dengan berfokus pada prinsip "*find-one-get-more*". Proses ekstraksi kalimat paralel yang dilakukan oleh Fung dan Cheung pertama-tama menemukan pasangan dokumen *comparable* dengan *cosine similarity* dan *bilingual lexicon*, lalu mengumpulkan seluruh kalimat yang memiliki kemungkinan paralel dari pasangan dokumen *comparable* yang ditemukan. Proses kemudian dilanjutkan dengan menghitung *cosine similarity* antar kandidat kalimat paralel lalu menyeleksi kalimat-kalimat tersebut untuk mengambil kalimat yang paralel. *Bilingual lexicon* yang digunakan kemudian diperbarui dengan mempelajarinya dari kalimat paralel yang dihasilkan memanfaatkan *tool* GIZA++. Dokumen *comparable* yang ditemukan kemudian dikembangkan dengan cara mencari dokumen yang berisi kalimat paralel hasil proes sebelumnya. Proses kemudian diulangi sampai kondisi konvergensi tercapai. Fung dan Cheung berhasil mendapatkan akurasi yang cukup bagus dari korpus paralel yang dihasilkan yaitu pada nilai 65,7% menggunakan metode tersebut.

Perbedaan antara proses yang dilakukan oleh Fung dan Cheung [4] dan penelitian ini adalah pada proses *alignment* dokumen, filter untuk mengklasifikasikan kalimat paralel, serta penelitian ini tidak mencoba untuk mengembangkan dokumen *comparable* yang telah digunakan. Proses *alignment* dokumen pada penelitan ini dilakukan dengan memanfaatkan fasilitas *interlanguage links* yang terdapat pada Wikipedia untuk menemukan pasangan dokumen yang memiliki topik serupa dalam versi bahasa yang berbeda.

Pemanfaatan *interlanguage links* pada Wikipedia untuk menemukan kemiripan informasi antar artikel yang memiliki versi bahasa lain ini sudah digunakan dalam beberapa penelitian. Adafre dan Rijke [5] menguji potensi Wikipedia sebagai sumber korpus multibahasa yang menyebutkan bahwa jarang terdapat ketidaksamaan informasi antar artikel Wikipedia dalam versi bahasa berbeda yang didapatkan dengan menggunakan struktur tautan antar artikel Wikipedia[1]. Tyers dan Pienaar [6] juga menggunakan fasilitas *interlanguage links* untuk proses pengumpulan pasangan kata terjemahan atau kamus multibahasa memanfaatkan Wikipedia dengan akurasi sebesar 69-92% dari empat pasangan bahasa. Serupa dengan penelitian ini, Mohammadi dan GasemAghaee [3] memanfaatkan *interlanguage links* untuk proses *alignment* dokumen dengan mengumpulkan setiap artikel Wikipedia berbahasa Persia yang memiliki versi lain dalam bahasa Inggris.

3. Wikipedia Sebagai Sumber Korpus Paralel

Korpus paralel merupakan kumpulan teks yang disusun sejajar dengan teks terjemahannya pada satu atau lebih bahasa lain dan dikenal juga dengan beberapa istilah lain seperti teks paralel, *bitext*, atau *multitext* [2]. Korpus paralel merupakan salah satu sumber data penting yang digunakan sebagai data latih untuk untuk keperluan pengembangan sistem MT dan juga untuk keperluan sistem yang melakukan pemprosesan *natural language* [1]. Sistem SMT memanfaatkan korpus paralel untuk melakukan proses *learning* yang dibutuhkan untuk mendapatkan keterkaitan antara teks bahasa asal (*source*) yang diterjemahkan ke dalam teks bahasa tertentu (*target*). Korpus paralel akan digunakan untuk mendapatkan *translation model* untuk menentukan kata dari teks bahasa *target* yang akan menghasilkan kata tertentu dalam bahasa *source*. Kegunaan lainnya yaitu untuk mendapatkan *language model* guna menentukan kata apa dalam bahasa *target* yang paling tepat digunakan bersama dengan kata yang lainnya. Korpus paralel disediakan dalam bentuk pasangan kalimat yang berkorespondensi satu-satu dengan masing-masing kalimat terjemahannya.

Penelitian ini menggunakan Wikipedia sebagai sumber teks bahasa Sunda dan teks bahasa Indonesia. Wikipedia adalah ensiklopedi multibahasa berbasis *web* yang mengizinkan pengguna untuk mengubah isi konten dan dapat diakses secara bebas[2]. Wikipedia, pada saat penulisan penelitian ini, memiliki artikel dalam 282 bahasa yang masih aktif dikembangkan dengan jumlah artikel terbanyak yaitu lebih dari lima juta artikel dalam bahasa Inggris. Terkait dengan penelitian ini, jumlah artikel dalam bahasa Sunda yaitu lebih dari 19.000 artikel dan lebih dari 380.000 artikel dalam bahasa Indonesia[3]. Wikipedia memiliki fasilitas berupa tautan yang berfungsi untuk mengarahkan pengguna pada artikel Wikipedia berbahasa lain yang mengandung subjek atau topik yang sama.

---

[1] Wikipedia menyebutkan struktur link ini sebagai *interlanguage links*:
https://en.wikipedia.org/wiki/Help:Interlanguage_links
[2] https://en.wikipedia.org/wiki/Wikipedia:About
[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias

Tautan ini pada wikipedia dikenal dengan sebutan *interlanguage links*. Contohnya, pada artikel Wikipedia berbahasa Sunda mengenai *getih* (dalam bahasa Indonesia berarti darah), pada halaman tersebut akan tersedia *interlanguage links* yang akan mengarah pada artikel Wikipedia berbahasa Indonesia berjudul *darah* dan begitu pula sebaliknya.

Artikel yang memiliki pasangan artikel dalam versi bahasa lain yang ditemukan pada *interlanguage links* bisa dimanfaatkan untuk mendapatkan informasi kemiripan isi konten antar artikel yang memiliki versi bahasa lain. Misalnya pada pasangan artikel berjudul *getih* dan *darah* yang dijelaskan sebelumnya, jika dilihat dari kalimat penyusun artikelnya mungkin berbeda akan tetapi topik dari isi konten masing-masing artikel serupa (sama-sama membicarakan tentang darah) sehingga fasilitas ini dapat digunakan juga untuk proses *document alignment* [3, 5].

4. Bilingual Lexicon

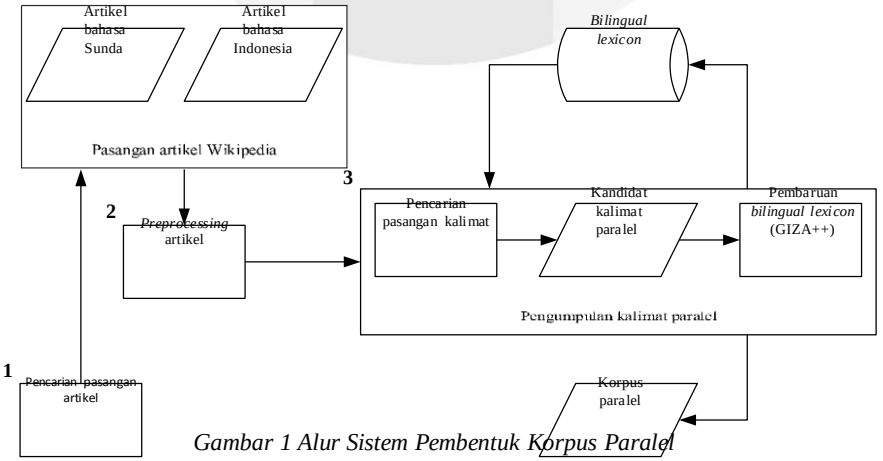
Tahap penerjemahan artikel berbahasa sunda ke dalam bahasa Indonesia untuk membantu proses perhitungan jumlah kata terjemahan antar kalimat sebagai bagian dari proses perhitungan *cosine similarity*, dilakukan dengan menggunakan *bilingual lexicon*. *Bilingual lexicon* yang dimaksudkan dalam penelitian ini adalah sebuah kamus dwibahasa bahasa Sunda-Indonesia berisi kumpulan kata dalam bahasa sunda dan dipasangkan dengan kata terjemahannya dalam bahasa Indonesia. Kualitas *bilingual lexicon* secara tidak langsung berpengaruh pada hasil pengukuran *cosine similarity*. Penelitian ini tidak menggunakan *bilingual lexicon* awal yang lengkap (tidak semua kata dalam bahasa Sunda tersedia), maka dari itu digunakanlah metode *bootstrapping* untuk meningkatkan kualitas *bilingual lexicon.*

Metode *bootstrapping* pada penelitian ini adalah suatu proses mandiri yang terdiri dari beberapa tahapan untuk meningkatkan kualitas dari *bilingual lexicon* melalui iterasi dengan bantuan kandidat kalimat paralel yang dihasilkan di setiap iterasi. Pada setiap akhir iterasi sebelum mencapai kondisi konvergensi, sejumlah kandidat kalimat paralel akan dihasilkan dan kemudian digunakan untuk memperbarui (menambahkan) entri baru yang belum pernah ada pada *bilingual lexicon* yang digunakan pada iterasi terakhir. Algoritma *expectation maximization* (EM) pada IBM *alignment model* berperan dalam pengumpulan pasangan kata dwibahasa Sunda-Indonesia baru yang didapatkan dari kumpulan kandidat kalimat paralel. Sebuah *tool* bernama GIZA++ yang mengimplementasikan IBM *alignment model* digunakan untuk keperluan penelitian ini.

IBM *alignment model* merupakan serangkaian model *alignment* berbasis kata yang membutuhkan kalimat paralel untuk proses *learning* dengan memanfaatkan algoritma EM [8]. Algoritma EM digunakan karena data yang didapatkan dari kalimat paralel tidak lengkap. Kalimat paralel hanya menyediakan data berupa kata bahasa *source* dan bahasa *target*, akan tetapi tidak tersedia data *alignment* antar kata. EM akan diterapkan pada tiap IBM Model untuk mengestimasi *alignment* antar kata sampai ditemukan *alignment* yang memiliki kemungkinan paling tepat (konvergensi). Penelitian ini menggunakan IBM Model sampai dengan Model 4.

5. Pembentukan Korpus Paralel

Sebuah sistem digunakan pada penelitian ini untuk melakukan proses pembentukan korpus paralel dari artikel Wikipedia berbahasa Sunda dan Indonesia. Gambar 1 menunjukkan alur sistem yang digunakan pada penelitian ini. Sistem ini terdiri dari tiga bagian utama yaitu pencarian pasangan artikel Wikipedia yang memiliki versi dalam bahasa Sunda dan bahasa Indonesia, *preprocessing* artikel Wikipedia dari pasangan artikel berbahasa Sunda-Indonesia, kemudian proses pencarian pasangan kalimat yang kemungkinan akan menjadi kalimat paralel termasuk proses pembaruan *bilingual lexicon* dengan menggunakan bantuan program GIZA++. Bagian 3.1 sampai dengan 3.3 akan menjelaskan tahapan dari awal sampai terbentuk korpus paralel.



Gambar 1 Alur Sistem Pembentuk Korpus Paralel

5.1 Pencarian Pasangan Artikel

Pada tahap ini dilakukan proses *document matching* atau dalam konteks penelitian ini pencarian pasangan artikel Wikipedia yang bertopik sama. Pencarian pasangan artikel dilakukan dengan memanfaatkan fasilitas *interlanguage links* yang terdapat pada setiap artikel Wikipedia. Pertama-tama sistem akan mengambil seluruh daftar judul artikel Wikipedia berbahasa Sunda. Sistem kemudian akan mengecek satu per satu dari daftar judul artikel tersebut untuk mencari judul artikel bahasa Sunda yang memiliki pasangan artikel Wikipedia dalam bahasa Indonesia.

5.2 *Preprocessing* Artikel

Pada tahapan sebelumnya, sistem hanya akan mengambil pasangan judul artikel tanpa isi artikel itu sendiri. Pada tahap ini sistem akan mengambil isi artikel dari pasangan artikel Wikipedia berbahasa Sunda dan Indonesia. Setelah teks dari suatu artikel Wikipedia didapatkan, tahapan *preprocess* akan dilakukan pada data teks yang didapatkan. Hal ini dilakukan karena teks yang didapatkan masih berupa sebuah *string* panjang tanpa batas kalimat dan juga masih mengandung penanda lainnya seperti *tag* XML ataupun penanda yang digunakan Wikipedia. *Preprocess* yang dilakukan yaitu:

- Merapikan dan menghapus duplikat karakter *whitespace* agar mudah dipisahkan kata dan kalimat pada tahap selanjutnya termasuk menandai bagian mana yang merupakan akhir kalimat dan pemisah antar kata.
- Menghapus bagian lampiran/*appendix* dan penanda subbab dari artikel Wikipedia dan seluruh teks setelahnya karena tidak cocok untuk dijadikan kalimat paralel.
- Menyingkirkan kalimat yang jumlah katanya kurang dari dua.
- Membuat teks menjadi dua versi, dengan cara membuat duplikatnya versi pertama merupakan versi orisinal tanpa penghapusan tanda baca yang hanya akan dipisah per kalimat dan versi kedua adalah teks yang akan dipisah per kata serta dihapus tanda bacanya untuk keperluan perhitungan *similarity* dan GIZA++.

Seluruh kalimat yang diambil dari dalam artikel Wikipedia akan direpresentasikan dalam bentuk vektor kata. hal ini dilakukan untuk mempermudah pengukuran kemiripan antar kalimat menggunakan *cosine similarity*.

5.3 Pengumpulan Kalimat Paralel

Pada tahap ini, *bilingual lexicon* akan mulai digunakan. Sistem akan melakukan pengumpulan kalimat paralel yang kemudian akan disusun menjadi sebuah korpus paralel. Gambar 2 menunjukkan poin penting yang dilakukan pada tahap ini dalam pembentukan korpus paralel sekaligus pembaruan *bilingual lexicon* dengan menggunakan bantuan IBM Model 4 EM *learner* yang diimplementasikan di dalam *tool* GIZA++.

1. Pencarian pasangan kalimat
1) Terjemahkan seluruh kata dari artikel Wikipedia Sunda dengan *bilingual lexicon*.
2) Pasangkan tiap kalimat dari artikel Wikipedia berbahasa Sunda dengan seluruh kalimat artikel Wikipedia berbahasa Indonesia.
3) Untuk setiap pasangan kalimat yang dibentuk, hitung *cosine similarity*.
4) Urutkan seluruh pasangan kalimat berdasarkan skor *cosine similarity* dari yang terbesar sampai yang terkecil.
5) Cek setiap pasangan kalimat pada seluruh filter, jika lolos → CP;

2. Pembaruan bilingual lexicon
1) Gunakan CP sebagai masukan GIZA++, dan dapatkan seluruh keluaran pasangan kata dwibahasa → W1;
2) Ambil pasangan kata dwibahasa yang belum ada pada *bilingual lexicon* jika *translation probability* pasangan kata dwibahasa W1 >= *threshold* → W2;
3) Tambahkan W2 pada *bilingual lexicon*;
4) Cek konvergensi, jika ya keluarkan hasil S2, jika tidak ulangi dari tahap 1.1.

Gambar 2 Proses Pembentukan Korpus Paralel

Proses pengumpulan kalimat paralel diawali dengan penerjemahan setiap kata dari kalimat bahasa Sunda ke dalam bahasa Indonesia menggunakan *bilingual lexicon*. Kata yang tidak memiliki entri di dalam *bilingual lexicon* akan dibiarkan tetap dalam bahasa Sunda. Penerjemahan merupakan awal proses pembobotan seluruh kata menggunakan *inverse document frequency*. Setelah dilakukan pembobotan, sistem akan melakukan perhitungan *cosine similarity* pada seluruh kemungkinan pasangan kalimat bahasa Sunda dan kalimat bahasa Indonesia. Kalimat yang berbahasa sama tidak akan dipasangkan. Setelah perhitungan selesai dilakukan, seluruh pasangan

kalimat tersebut kemudian proses seleksi kandidat kalimat paralel dilakukan. Proses seleksi dilakukan dengan menerapkan beberapa filter, yaitu:

1. Nilai hasil perhitungan *cosine similarity* harus lebih dari nol.
2. *Word overlap*. Setiap kalimat dalam suatu pasangan harus memiliki paling sedikit *n*-persen kata yang sama di dalam kalimat pasangannya.
3. Rasio panjang (jumlah kata) antara dua kalimat yang berpasangan tidak lebih dari dua.
4. Hapus duplikat. Salah satu kalimat harus belum pernah dimasukkan menjadi kandidat kalimat paralel, untuk mencegah duplikat. Filter ini memanfaatkan hasil pengurutan skor *cosine similarity* untuk mengukur derajat keterkaitan antar kalimat di dalam setiap pasangan kalimat.

Beberapa kalimat akan lolos filter yang diterapkan dan dimasukkan ke dalam kandidat kalimat paralel. Seluruh kandidat kalimat paralel ini kemudian digunakan sebagai data untuk proses *learning* algoritma EM pada IBM Model 4 menggunakan *tool* GIZA++. GIZA++ akan menghasilkan salah satu *output* berisi pasangan kata dwibahasa (Sunda-Indonesia) beserta nilai probabilitas kemungkinan terjemahan dari suatu kata ke kata pasangannya.

Sistem kemudian akan mengambil pasangan kata dwibahasa yang memiliki nilai probabilitas satu karena kualitas serta kuantitas *bilingual lexicon* awal kurang baik dan sebagian besar hasil *lexicon translation probability* yang dihasilkan GIZA++ kurang baik. Pasangan kata dwibahasa yang memiliki nilai probabilitas sama dengan satu kemudian dicek kembali apakah kata bahasa Sunda yang didapatkan sudah ada pada *bilingual lexicon* atau belum. Jika belum ada, maka akan kata tersebut termasuk pasangan kata bahasa Indonesianya akan digunakan sebagai entri baru *bilingual lexicon*. Sistem akan mengulangi proses dari tahap penerjemahan sampai kondisi konvergensi tercapai dan pada iterasi terakhir, akan dihasilkan sebuah korpus paralel.

6. Evaluasi

Korpus paralel hasil bentukan sistem yang menerapkan metode *bootstrapping* dengan bantuan IBM Model 4 EM diuji dengan menggunakan penilaian manusia dan juga dibandingkan dengan penggunaan korpus paralel tersebut pada sistem SMT Moses[4]. Pengujian dengan penilaian manusia dibagi menjadi lima skala Likert yang melambangkan kualitas tiap kalimat paralel sesuai dengan dengan nomor urutnya:

Skala **1**: Kalimat Sunda dan Indonesia memiliki arti yang seluruhnya sama.
Skala **2**: Kalimat Sunda dan Indonesia memiliki arti yang sebagian besar sama.
Skala **3**: Kalimat Sunda dan Indonesia memiliki arti yang cukup sama.
Skala **4**: Kalimat Sunda dan Indonesia memiliki arti yang sebagian kecil sama.
Skala **5**: Kalimat Sunda dan Indonesia memiliki arti yang tidak sama.

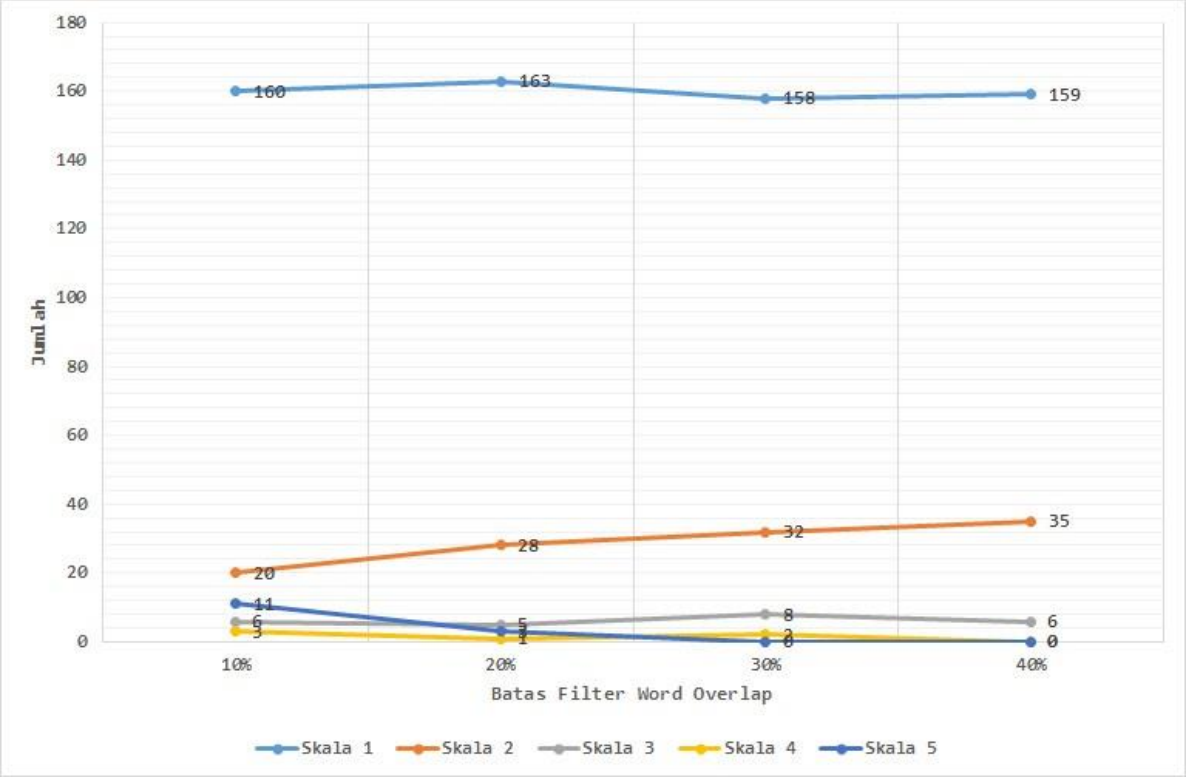
Sistem pada penelitian ini menggunakan empat buah filter yang digunakan untuk menentukan pasangan kalimat mana saja yang cocok menjadi kandidat kalimat paralel dan pada akhirnya menjadi kalimat paralel. Salah satu dari filter tersebut yaitu filter untuk membatasi jumlah minimal kata yang sama atau *word overlap* dalam satu pasang kalimat bergantung pada *bilingual lexicon* yang digunakan. Penelitian ini tidak menggunakan *bilingual lexicon* yang lengkap sehingga batas optimal akan dicari pada rentang 10% sampai 40% dengan perbedaan 10 poin menjadi empat kategori yaitu batas nilai filter *word overlap* 10%, 20%, 30%, dan 40%. Pengujian dilakukan untuk melihat apa efek dari penggunaan nilai batas filter *word overlap* yang berbeda sekaligus melihat bagaimana kualitas korpus paralel yang dihasilkan. Pengujian dengan menggunakan penilaian manusia dilakukan dengan menggunakan data uji yang terdiri dari 200 kalimat paralel. Penggunaan jumlah data uji yang lebih banyak sulit dilakukan akibat keterbatasan sumber daya. Hasil pengujian dari penilaian manusia dapat dilihat pada Gambar 3.

Sebagai informasi, jumlah kalimat paralel yang ditampilkan di tiap kategori nilai filter pada Gambar 3 merupakan kumpulan nilai modus skala tiap kalimat paralel dari sejumlah hasil data uji dan bukan merupakan nilai rata-rata karena menggunakan skala Likert. Berdasarkan hasil pengujian, semakin besar batas nilai filter *word overlap* memberikan efek tidak langsung yaitu peningkatan kualitas korpus paralel. Efek tersebut dicapai melalui pengurangan jumlah kalimat paralel yang hanya mendapatkan skala 4 dan 5 – melambangkan pasangan kalimat sebagian kecil sama dan tidak sama secara berurutan – atau dengan kata lain pasangan kalimat yang tidak atau sedikit sekali memiliki kemiripan kata dalam kalimat antar pasangan, semakin kecil kemungkinannya untuk masuk ke dalam korpus paralel. Efek langsung dari semakin besarnya nilai filter *word overlap* yaitu pengurangan jumlah kandidat kalimat paralel di setiap iterasi *bootstrapping* dan adanya kecenderungan peningkatan jumlah kata yang saling menerjemahkan dalam tiap pasangan kalimat (*word overlap*). Peningkatan nilai *word overlap* bisa berubah menjadi penurunan jika batas filter *word overlap* yang digunakan terlalu tinggi dan *bilingual lexicon* tidak mampu menangani penerjemahan dengan baik sebagai akibat penggunaan *bilingual lexicon* yang tidak lengkap. Peningkatan nilai *word overlap* memang tidak serta-merta meningkatkan kualitas korpus paralel, akan tetapi meningkatkan kemungkinan suatu pasangan kalimat memiliki subjek yang sama. Kombinasi *bootstrapping* dan

[4] http://www.statmt.org/moses/

filter lainnya membuat korpus paralel yang dihasilkan semakin kecil kemungkinannya tercemar oleh pasangan kalimat yang cenderung sangat tidak paralel dan berujung pada korpus yang lebih berkualitas.

Bootstrapping berperan penting dalam sistem pembentuk korpus paralel ini. Tanpa *bootstrapping*, sistem hanya akan berjalan satu kali tanpa iterasi dan menjadi sangat bergantung pada *bilingual lexicon* yang digunakan untuk menerjemahkan kata berbahasa Sunda ke bahasa Indonesia. Penelitian ini tidak menggunakan *bilingual lexicon* yang lengkap dan dengan menggunakan teknik *bootstrapping* efek dari penggunaan *bilingual lexicon* yang tidak lengkap dapat diminimalkan. Pada setiap iterasi, sistem akan memperbarui *bilingual lexicon* yang digunakan dan juga mengganti kandidat kalimat paralel dengan kandidat yang sudah menggunakan *bilingual lexicon* yang baru.



Gambar 3 Grafik Perbandingan Batas Filter Word Overlap Terhadap Kualitas Korpus Paralel

Pengujian dengan memanfaatkan bantuan penilaian manusia menunjukkan bahwa peningkatan batas nilai filter *word overlap* membantu meningkatkan kualitas korpus paralel secara tidak langsung dengan cara mencegah pasangan kalimat yang tidak paralel masuk ke dalam korpus paralel. Pengujian kemudian dilanjutkan dengan menggunakan korpus paralel hasil sistem sebagai data latih pada sistem SMT Moses. Hasil dari pengujian ini sedikit berbeda dengan hasil penilaian manusia. Skor BLEU yang dihasilkan lebih menggambarkan kualitas korpus paralel secara keseluruhan dibandingkan dengan pengujian manual menggunakan penilaian manusia yang bisa memberikan gambaran mengenai kualitas korpus sampai pada tingkat kalimat. Tabel 1 menunjukkan hasil pengujian menggunakan korpus paralel uji dari tiap kategori yang masing-masing berisi 200 pasang kalimat.

Tabel 1 Skor BLEU Korpus Uji

Jumlah data : 200 kalimat paralel tiap kategori filter					
***Word Overlap* Filter**	**Total BLEU (%)**	**1-gram (%)**	**2-gram (%)**	**3-gram (%)**	**4-gram (%)**
10%	51,12	73,6	56,7	44,9	36,5
20%	54,14	75,3	59,5	48,2	39,8
30%	54,14	75,5	59,4	48,1	39,8
40%	54,44	75,7	59,7	48,5	40,1

Tabel 1 menggunakan korpus paralel yang sama dengan pengujian manual menggunakan penilaian manusia. Korpus paralel tersebut akan digunakan sebagai data latih pada sistem Moses untuk mendapatkan *translation model*. Data uji yang digunakan pada Moses terdiri dari 120 kalimat bahasa Sunda yang diambil dari Wikipedia dan untuk perhitungan BLEU digunakan data referensi hasil terjemahan manual dari data uji. Hasil

yang ditampilkan pada Tabel 1 mendukung bukti hasil pengujian manual bahwa perubahan batas nilai *word overlap* membantu meningkatkan kemungkinan suatu pasangan kalimat memiliki subjek yang sama dan secara bersamaan menyingkirkan pasangan kalimat yang buruk. Terlihat dari skor total BLEU yang semakin meningkat mengikuti batas nilai filter *word overlap* menunjukkan bahwa *translation model* yang dilatih dari korpus paralel uji menjadi lebih baik ketika batas nilai *word overlap* meningkat. Pengukuran dengan menggunakan skor BLEU ini juga menunjukkan bahwa korpus paralel yang dihasilkan sistem pada penelitian ini mampu menghasilkan korpus dengan tingkat *adequacy* cukup tinggi akan tetapi tingkat *fluency*-nya rendah. Hal tersebut bisa dilihat dari skor BLEU 1-gram yang cukup tingi tetapi kemudian terjadi penurunan pada skor *n*-gram lebih dari satu. BLEU 1-gram cenderung menunjukkan *adequacy* dan *n*-gram lebih dari 1 menunjukkan *fluency* dari korpus paralel yang terkait [7]. *Adequacy* menunjukkan bahwa hasil terjemahan dari Moses memiliki arti yang sama dalam tingkatan tertentu sedangkan *fluency* menunjukkan ketepatan hasil terjemahan Moses (meliputi tata bahasa dan pemilihan kata).

7. Kesimpulan

Sistem yang digunakan mampu membentuk korpus paralel dengan syarat sumber korpus yang digunakan masih memiliki keterkaitan subjek. Persyaratan tersebut disebabkan oleh penggunaan *bilingual lexicon* yang tidak lengkap. IBM Model 4 EM *learner* yang diterapkan pada GIZA++ dimanfaatkan untuk memperbarui *bilingual lexicon* sekaligus meminimalkan dampak dari *bilingual lexicon* yang tidak lengkap. Dampak tersebut diminimalkan dengan penggunaan metode *bootstrapping* yang kemudian akan meningkatkan nilai *word overlap* di setiap iterasi. Peningkatan nilai *word overlap* secara tidak langsung meningkatkan kemungkinan ditemukannya kalimat paralel melalui perhitungan *similarity*.

Hasil analisis memperlihatkan 79,5% dari 200 kalimat paralel yang diujikan memang bersifat paralel. Adapun akurasi penerjemahan dengan menggunakan korpus paralel tersebut pada mesin penerjemah Moses menghasilkan skor BLEU sebesar 54,44. Akurasi tersebut masih bisa ditingkatkan dengan memperbaiki faktor lainnya yang mempengaruhi penerjemahan dengan mesin.

Daftar Pustaka:

[1] P. Resnik and N. A. Smith, "The Web as a Parallel Corpus," *Computational Linguistics*, vol. 29, no. 3, pp. 349-380, 2003.

[2] A. Lopez, "Statistical Machine Translation," *ACM Computing Surveys*, vol. 40, no. 3, pp. 8:1-8:49, 2008.

[3] M. Mohammadi and N. GhasemAghaee, "Building Bilingual Parallel Corpora Based on Wikipedia," in *Computer Engineering and Applications (ICCEA), 2010 Second International Conference*, Bali, Indonesia, 2010.

[4] P. Fung and P. Cheung, "Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[5] S. F. Adafre and M. d. Rijke, "Finding Similar Sentences across Multiple Languages in Wikipedia," in *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, Trento, Italy, 2006.

[6] F. M. Tyers and J. A. Pienaar, "Extracting Bilingual Word Pairs from Wikipedia," in *Collaboration: interoperability between people in the creation of language resources for less-resourced languages (A SALTMIL workshop)*, Marrakech, Morocco, 2008.

[7] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, US, 2002.

[8] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263-311, 1993.