JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989
Vol. 6, No. 4, Oktober 2018

Pengaruh Domain Teks Pada Korpus Terhadap Akurasi Mesin Penerjemah Statistik

Khamsah Akbar[#1], Herry Sujaini[#2], Rudy Dwi Nyoto[#3]

[#]*Program Studi Informatika Universitas Tangjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Kota Pontianak, 78115

[1]khamsahakbar@gmail.com
[2]hs@untan.ac.id
[3]rudydn@informatika.untan.ac.id

***Abstrak*— Salah satu faktor yang mempengaruhi tingkat akurasi suatu mesin penerjemah statistik adalah adanya suatu korpus yang baik sebagai sumber data yang digunakan untuk pembuatan mesin penerjemah statistik sehingga korpus yang dihasilkan dapat akurat dan memiliki persentase yang tinggi pada saat melakukan penerjemahan. Beberapa kriteria dari suatu korpus yang baik adalah orientasi ke bahasa atau variasi untuk dijadikan sampel, kriteria yang akan kita pilih yang meliputi mode teks, jenis teks, domain teks, bahasa, lokasi teks, tanggal teks serta sifat dan dimensi sampel [1]. Tujuan yang ingin dicapai dalam penelitian ini adalah untuk mengetahui seberapa besar pengaruh domain teks pada korpus terhadap nilai akurasi hasil terjemahan pada mesin penerjemah statistik Bahasa Inggris ke Bahasa Indonesia. Pengujian untuk mendapatkan nilai akurasi dilakukan dengan dua cara, yaitu pengujian otomatis menggunakan *Bilingual Evaluation Understudy* (BLEU) dan pengujian manual oleh ahli bahasa Inggris. Untuk pengujian otomatis dilakukan pada setiap mesin penerjemah yang sudah dibangun dengan pembagian fold pada korpus. Pengujian manual dilakukan oleh seorang ahli Bahasa Inggris dengan korpus uji sebanyak 100 kalimat. Berdasarkan hasil pengujian, domain teks pada korpus memiliki perbedaan nilai akurasi terjemahan dari mesin penerjemah statistik bahasa Inggris – bahasa Indonesia yaitu sebesar 7,6409% pada pengujian dengan BLEU dan 1,01% untuk pengujian oleh ahli bahasa.**

***Kata kunci*— korpus, mesin penerjemah statistik, domain teks, korpus spesifik, korpus campuran.**

I. PENDAHULUAN

Corpus (plural corpora) atau teks korpus adalah kumpulan teks yang besar dan terstruktur (sekarang biasanya tersimpan dan diproses secara elektronik). Korpus digunakan untuk melakukan analisis statistik dan pengujian hipotesis, memeriksa kejadian atau memvalidasi aturan linguistik dalam wilayah bahasa tertentu.

Sebuah korpus mungkin berisi teks dalam satu bahasa *(monolingual corpus)* atau data teks dalam berbagai bahasa *(multilingual corpus)*. Multilingual korpus yang telah diformat secara khusus untuk perbandingan *side by side* disebut korpus paralel sejajar. Ada dua tipe utama korpus paralel yang berisi teks dalam dua bahasa. Dalam terjemahan korpus, teks dalam satu bahasa adalah terjemahan teks dalam bahasa lain. Dalam korpus yang sebanding, teksnya sama dan mencakup konten yang sama, namun terjemahannya sama sekali tidak saling terkait. Untuk mengeksploitasi teks paralel, beberapa jenis alignment teks yang mengidentifikasi segmen teks setara (frase atau kalimat) adalah prasyarat untuk analisis. Algoritma terjemahan mesin untuk menerjemahkan antara dua bahasa sering dilatih menggunakan fragmen paralel yang terdiri dari korpus bahasa pertama dan korpus bahasa kedua yang merupakan elemen untuk terjemahan elemen korpus bahasa pertama.

Korpus adalah basis pengetahuan utama dalam linguistik korpus. Analisis dan pengolahan berbagai jenis korpus juga menjadi subyek banyak pekerjaan dalam bahasa linguistik komputasi, pengenalan ucapan dan terjemahan mesin. Korpus dan daftar frekuensi yang berasal darinya berguna untuk pengajaran bahasa. Korpus dapat dianggap sebagai jenis bantuan penulisan bahasa asing karena pengetahuan gramatikal kontekstual yang diperoleh oleh pengguna bahasa non-asli melalui pemaparan teks asli.

Salah satu faktor yang mempengaruhi tingkat akurasi suatu mesin penerjemah statistik adalah adanya suatu korpus yang baik sebagai sumber data yang digunakan untuk pembuatan mesin penerjemah statistik sehingga korpus yang dihasilkan dapat akurat dan memiliki persentase yang tinggi pada saat melakukan penerjemahan. Adapun beberapa kriteria dari suatu korpus yang baik adalah orientasi ke bahasa atau variasi untuk dijadikan sampel, kriteria yang akan kita pilih yang meliputi mode teks, jenis teks, domain teks, bahasa, lokasi teks, tanggal teks serta sifat dan dimensi sampel [1].

Terdapat penelitian yang dilakukan berkaitan dengan domain teks diantaranya penelitian tentang melihat kata-kata yang terjemahannya salah dan tidak terlihat pada saat berubah ke domain baru [2], penelitian tentang masalah domain

Submitted 31-08-2018; Revised 11-10-2018; Accepted 18-10-2018

menggunakan kombinasi bobot fitur dan adaptasi model bahasa untuk membedakan beberapa domain dengan model *log-linear* berbasis frasa[3], penelitian tentang adaptasi domain untuk mesin penerjemah statistik di mana domain di korpus bilingual tidak ada[4], penelitian tentang adaptasi domain yaitu menggabungkan teks komentar dan teks berita dengan domain kecil dari *bi-text out-of-domain* dan domain besar dari *corpus Europarl* [5], penelitian tentang pengekstrakan informasi klinis dari laporan radiologi dalam bahasa Portugis dengan menggunakan mesin terjemahan dan teknik pengambilan informasi lintas bahasa dengan domain teks [6].

Berdasarkan uraian tersebut, maka perlu dilakukan penelitian untuk melihat sejauh mana pengaruh domain teks terhadap akurasi mesin penerjemah statistik.

II. URAIAN PENELITIAN

A. 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 beberapa jenis tergantung tujuan penggunaannya. Ada delapan jenis yakni korpus khusus *(specialised corpus)*, korpus umum *(general corpus)*, korpus komparatif *(comparable corpus)*, korpus paralel *(parallel corpus)*, korpuspemelajar *(learner corpus)*, korpus pedagogis *(pedagogic corpus)*, korpushistoris atau diakronis *(historical or diachronic corpus)*, dan korpus monitor *(monitor corpus)* [8].

B. Definisi Penerjemahan

Dalam Kamus Besar Bahasa Indonesia (KBBI) kata "terjemah/menerjemahkan" merupakan menyalin (memindahakan) suatu bahasa ke bahasa lain atau mengalihbahasakan. Selain itu, penerjemahan adalah kegiatan mengalihkan secara tertulis pesan dari teks suatu bahasa (misalnya bahasa Inggris) ke dalam teks bahasa lain (misalnya bahasa Indonesia) [9]. Penerjemahan adalah pengalihan pikiran atau gagasan dari suatu bahasa sumber ke dalam bahasa yang lain. Penerjemahan adalah mengubah teks bahasa sumber ke dalam teks bahasa sasaran dengan mempertimbangkan makna kedua bahasa sehingga diusahakan semirip-miripnya, yang tak kalah pentingnya adalah terjemahan harus mengikuti kaidah-kaidah yang berlaku dalam bahasa sasaran [10].

C. Proses Penerjemahan

Proses penerjemahan terdiri dari 3 tahap yaitu *analysis, transfer* dan *restructuring*. Dalam proses *analysis*, penerjemah menganalisis isi pesan bahasa sumber berdasarkan gramatika dan makna. Pada tahap ini kalimat-kalimat bahasa sumber dipecah-pecah menjadi satuan-satuan gramatikal berstruktur kalimat-kalimat dasar, kata-kata dan frase-frase untuk menangkap makna yang ada dengan teknik analisis komponen. Tahap kedua, *transfer*, yaitu proses pengalihan materi-materi yang telah dianalisis dari bahasa sumber ke dalam bahasa sasaran. Tahap terakhir yaitu *restructuring*, bahwa penerjemah menyusun materi-materi yang telah dialihkan dan bertujuan untuk membuat pesan yang secara keseluruhan dapat diterima [11].

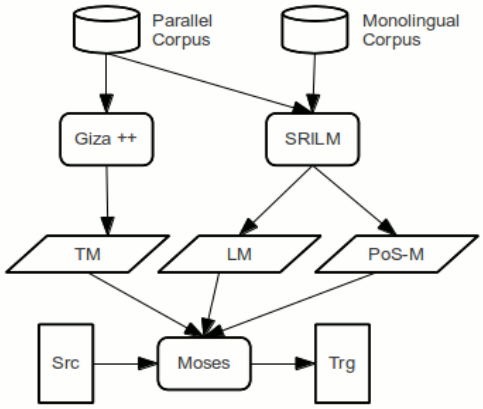
D. Mesin Penerjemah Statistik

Mesin penerjemah statistik merupakan salah satu jenis mesin penerjemah dengan menggunakan pendekatan statistik. Pendekatan statistik yang digunakan adalah konsep probabilitas. Setiap pasangan kalimat (S,T) akan diberikan sebuah P(T|S) yang diinterpretasikan sebagai distribusi probabilitas dimana sebuah penerjemah akan menghasilkan T dalam bahasa sasaran ketika diberikan S dalam bahasa sumber [12].

Mesin penerjemah statistik (MPS) *atau Statistical machine translation* (SMT) adalah suatu paradigma dari mesin penerjemah di mana penerjemahan dilakukan berbasis model statistik dengan parameter-parameter yang diturunkan dari analisis paralel korpus [13].

Ada beberapa pendekatan untuk *machine translation* seperti pendekatan dengan menggunakan aturan *(rule-based machine translation)*, pendekatan dengan menggunakan contoh (*example-based machine translation*), dan pendekatan dengan menggunakan model statistik *(statistical machine translation).* Dalam mesin penerjemah statistik, terdapat 3 komponen yang terlibat dalam proses penerjemahan dari satu bahasa ke bahasa lain yaitu : *language model, translation model,* dan *decoder* [14].

Secara umum, arsitektur mesin penerjemah statistik moses ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur Mesin Penerjemah Statistik Moses [15]

Sumber data utama yang dipergunakan adalah *parallel corpus* (korpus paralel) dan *monolingual corpus* (monolingual korpus). Proses *training* terhadap korpus paralel

menggunakan GIZA++ menghasilkan *translation model* (TM). Proses *training* terhadap bahasa target pada korpus paralel ditambah dengan monolingual korpus bahasa target menggunakan SRILM menghasilkan *language model* (LM), sedangkan *PoS model* (PoS-M) dihasilkan dari bahasa target pada korpus paralel 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 [16].

E. Moses

Moses meupakan *software open source* yang merupakan implementasi dari mesin penerjemah statistik. Moses digunakan untuk melatih model statistik teks terjemahan dari bahasa sumber ke bahasa target. Saat melakukan penerjemahkan bahasa, moses membutuhkan korpus dalam dua bahasa, bahasa sumber dan bahasa target [17]. 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 Euro Matrix, dengan pendanaan oleh *European Commission* [18].

F. Automatic Evaluation

Sistem evaluasi otomatis yang populer saat ini adalah BLEU *(Bilingual Evaluation Uderstudy)*. BLEU adalah sebuah algoritma yang berfungsi untuk mengevaluasi kualitas dari sebuah mesin terjemahan yang telah diterjemahkan oleh mesin dari satu bahasa alami ke bahasa lain. Ide utama dibalik ini adalah "semakin dekat terjemahan sebuah mesin dengan terjemahan manusia, maka akan semakin baik" [19].

G. Tokenizing

Tahap *tokenizing* adalah tahap pemotongan string *input* berdasarkan tiap kata yang menyusunnya [20].

III. METODOLOGI PENELITIAN

A. Data Penelitian

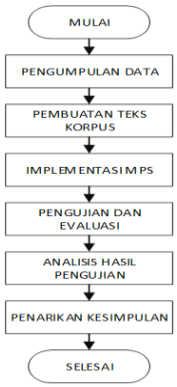
Data penelitian berupa berita yang diperoleh dari Antara *News*, *National Geography* Indonesia berita IPTEK dan situs-situs berita lainnya. Berita tersebut selanjutnya diolah menjadi korpus teks paralel bahasa Indonesia dan bahasa Inggris. Adapun jumlahnya yaitu 5500 pasangan kalimat korpus paralel bahasa Inggris dan bahasa Indonesia masing-masing untuk domain teks spesifik dan domain teks campuran. Pada domain teks spesifik korpus yang digunakan yaitu korpus ekonomi, sedangkan untuk domain teks campuran terdiri dari gabungan korpus ekonomi, olahraga, internasional dan sains. Pada korpus uji menggunakan korpus spesifik ekonomi berjumlah 500 kalimat.

B. Perangkat Penelitian

Perangkat penelitian yang digunakan dalam penelitian ini terdiri dari perangkat keras dan perangkat lunak. Perangkat keras menggunakan laptop ASUS X450CC dengan spesifikasi *processor* Intel Core i3 (2.0 GHz), *Graphic Processor* NVIDIA GeForce GT 720M, *Memory* 2048, HDD 500 GB. Perangkat lunak terdiri dari Sistem operasi Linux Ubuntu 16.04 LTS, SRILM untuk pemodelan bahasa, Giza++ untuk pemodelan translasi, Moses untuk *decoding*, BLEU untuk pengujian akurasi, dan Sublime Text 2 untuk teks editor.

C. Metodologi Penelitian

Metodologi penelitian yang dilakukan dijelaskan pada Gambar 2.



Gambar 2. Diagram Alir Penelitian

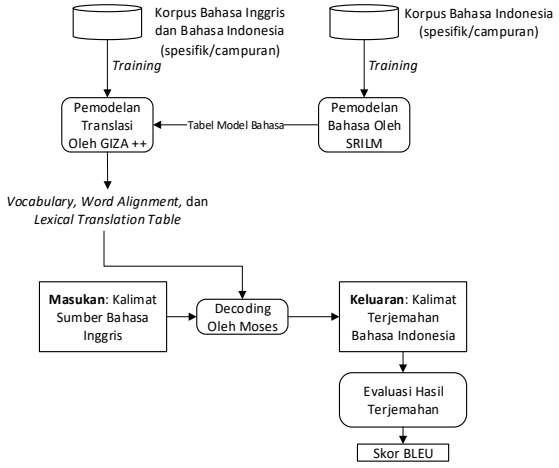
Penelitian ini akan menggunakan beberapa metode. Metode yang akan digunakan dalam penelitian ini antara lain.

1. Pengumpulan Data
   Proses pengumpulan data merupakan proses mengumpulkan data-data yang akan digunakan untuk penelitian. Data berupa korpus bahasa Indonesia dan bahasa Inggris yang nantinya akan dibuat menjadi korpus teks paralel. Domain teks yang dipakai adalah domain teks spesifik dan domain teks campuran. Proses pengumpulan korpus diambil dari Antara *News*, *National Geography* Indonesia berita IPTEK dan situs-situs berita lainnya yang akan diolah menjadi teks korpus.
2. Analisis Kebutuhan
   Korpus teks paralel dibuat dari terjemahan kalimat – kalimat dari korpus bahasa Indonesia dan bahasa Inggris yang sudah dikelompokkan berdasarkan domain teks pada korpus. Kesetaraan korpus pada korpus spesifik dan campuran, dilihat dari jumlah kuantitas pada setiap korpus yang memiliki kuantitas (jumlah) kalimat yang sama dan memiliki token yang kurang lebih sama. Adapun jumlahnya yaitu 5500 pasangan kalimat korpus paralel bahasa Inggris dan bahasa Indonesia masing-masing untuk domain teks spesifik dan domain teks campuran. Pada domain teks spesifik korpus yang digunakan yaitu korpus ekonomi, sedangkan untuk domain teks campuran terdiri dari gabungan korpus ekonomi, olahraga, internasional dan sains. Pada korpus uji menggunakan korpus spesifik ekonomi berjumlah 500 kalimat.

3. Implementasi Mesin Penerjemah Statistik
   Proses awal yang dilakukan adalah melakukan *training* terhadap korpus bahasa sumber dan bahasa target yang terdiri dari *cleaning*, *tokenizing*, dan *lowercase.* Implementasi dilakukan dengan cara melakukan pemodelan bahasa, pemodelan translasi dan *decoding* pada mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia.
   Pada pemodelan bahasa diproses oleh SRILM yang akan menghasilkan table model bahasa. Proses selanjutnya akan dilakukan pemodelan translasi oleh GIZA++ yang akan menghasilkan 3 *folder yaitu vocabulary corpus*, *word alignment*, dan *lexical translation table*. Proses decoding telah dapat dilakukan dengan melakukan input bahasa sumber yaitu bahasa Inggris dan menghasilkan *output* bahasa target yaitu bahasa Indonesia
4. Pengujian dan Evaluasi Hasil Terjemahan Mesin Translasi
   Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi dari domain teks pada korpus yang sudah dibuat sebelumnya. Pengujian dilakukan secara otomatis menggunakan BLEU dan oleh ahli bahasa pada setiap mesin penerjemah yang telah dibuat sehingga didapatkan persentase akurasi hasil terjemahan dan didapat hasil tabel dari uji akurasi yang telah dilakukan. Dari persentase tersebut juga dapat diperoleh grafik perubahan nilai akurasi dari setiap penambahan kalimat pada domain teks spesifik dan domain teks campuran. Sehingga dapat dilihat perubahan persentase dari setiap mesin yang telah di uji baik pengujian menggunakan BLEU maupun oleh ahli bahasa.
5. Analisis Hasil Pengujian
   Analisis hasil pengujian dilakukan untuk mengetahui karakteristik mesin penerjemah statistik bahasa Inggris dan bahasa Indonesia dan mengidentifikasi apakah sudah sesuai dengan kebutuhan atau tidak berdasarkan dengan uji akurasi mesin penerjemah statistik yang telah dilakukan pada setiap domain teks yaitu domain teks spesifik dan domain teks campuran.
6. Penarikan Kesimpulan
   Penarikan kesimpulan dirumuskan berdasarkan analisis hasil pengujian dan sesuai dengan tujuan penelitian, sehingga mampu memberikan solusi berdasarkan rumusan masalah dilakukannya penelitian.

D. Perancangan Implementasi Mesin Penerjemah Statistik

Perancangan untuk arsitektur sistem pada mesin penerjemah statistik terdiri dari beberapa proses yaitu pemodelan bahasa, *decoding* dan evaluasi hasil terjemahan. Arsitektur mesin penerjemah statistik dari Bahasa Inggris ke bahasa Indonesia diperlihatkan pada Gambar 3.

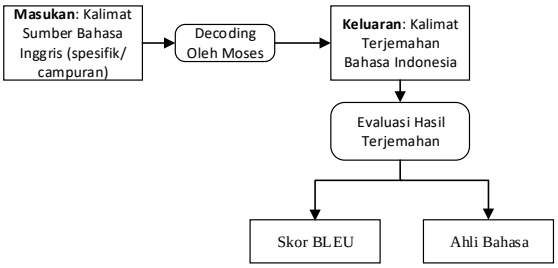


Gambar 3. Arsitektur Sistem Mesin Penerjemah Statistik

E. Perancangan Penelitian dan Evaluasi Hasil Terjemahan

pengujian secara otomatis menggunakan BLEU *(Bilingual Evaluation Understudy)* dan pengujian manual yang dilakukan oleh ahli bahasa Inggris. Proses pengujian otomatis dengan BLEU dilakukan untuk mendapatkan nilai akurasi terjemahan mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia terhadap kuantitas korpus dengan jumlah kalimat uji berdasarkan korpus maksimal.

BLEU *(Bilingual Evaluation Understudy)* mengukur *modified n-gram precision score* antara hasil terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang disebut *brevity penalty*. Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score.* Adapun perancangan mesin penerjemah statistik dapat dilihat pada gambar 4.



Gambar 4. Perancangan Mesin Penerjemah Statistik

Pada penelitian ini digunakan 5500 kalimat yang dibagi atas 11 *fold* yang masing-masing berjumlah 500 kalimat seperti yang diperlihatkan pada Tabel 1. Untuk sebuah percobaan diambil 1 buah *fold* untuk testing dan 10 buah *fold* sisanya untuk *training*. Dengan kata lain, dari 5500 kalimat korpus paralel dibagi menjadi 500 kalimat untuk *testing* dan 5000 kalimat untuk *training* seperti yang terlihat pada Tabel 1. Sehingga untuk sebuah data set korpus paralel dilakukan 11 kali percobaan dan dari hasil 1-11 percobaan tersebut diambil

nilai rata-ratanya. Proses pengujian otomatis dengan BLEU dilakukan untuk mendapatkan nilai akurasi terjemahan mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia dan menggunakan domain teks spesifik sebagai korpus uji. Pembagian *fold* dapat dilihat pada Tabel 1 dan jumlah pembagian *fold* pada Tabel 2.

TABEL 1
PEMBAGIAN FOLD PADA KORPUS

Fold	Jumlah
F1	1-500
F2	501-1000
F3	1001-1500
F4	1501-2000
F5	2001-2500
F6	2501-3000
F7	3001-3500
F8	3501-4000
F9	4001-4500
F10	4501-5000
F11	5001-5500

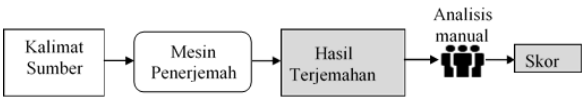
TABEL 2
JUMLAH MESIN DENGAN PEMBAGIAN *FOLD*

Mesin	*Testing*	*Training*
1	F1	F2 F3 F4 F5 F6 F7 F8 F9 F10 F11
2	F2	F1 F3 F4 F5 F6 F7 F8 F9 F10 F11
3	F3	F1 F2 F4 F5 F6 F7 F8 F9 F10 F11
4	F4	F1 F2 F3 F5 F6 F7 F8 F9 F10 F11
5	F5	F1 F2 F3 F4 F6 F7 F8 F9 F10 F11
6	F6	F1 F2 F3 F4 F5 F7 F8 F9 F10 F11
7	F7	F1 F2 F3 F4 F5 F6 F8 F9 F10 F11
8	F8	F1 F2 F3 F4 F5 F6 F7 F9 F10 F11
9	F9	F1 F2 F3 F4 F5 F6 F7 F8 F10 F11
10	F10	F1 F2 F3 F4 F5 F6 F7 F8 F9 F11
11	F11	F1 F2 F3 F4 F5 F6 F7 F8 F9 F10

Setiap percobaan akan dilakukan dua tahap. Tahap pertama akan dibuat mesin translasi bahasa Inggris ke bahasa Indonesia untuk domain teks spesifik. Tahap kedua akan dibuat mesin translasi bahasa Inggris ke bahasa Indonesia untuk domain teks campuran.

Langkah pada pengujian otomatis, korpus yang akan diuji terlebih dahulu melalui langkah translasi otomatis. Translasi otomatis akan memberikan *output* berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin, Setelah membuat *output* berupa hasil translasi otomatis dari mesin penerjemah, langkah selanjutnya adalah mendapatkan nilai BLEU dari output dengan cara membandingkan *output* tersebut dengan korpus bahasa target yang telah dibuat sebelumnya.

Pengujian secara manual akan dilakukan oleh ahli bahasa Inggris yaitu oleh Antonius Yonathan seorang mahasiswa Program Studi Informatika, Universitas Tanjungpura. Data yang digunakan dalam pengujian yaitu 100 kalimat bahasa Inggris sebagai kalimat bahasa sumber dan bahasa Indonesia sebagai kalimat bahasa target, masing-masing untuk domain teks spesifik dan domain teks campuran yang kemudian akan dinilai ketepatannya berdasarkan pemahaman dan pengetahuan ahli bahasa. Adapun perancangan proses penilaian secara manual baik untuk domain teks spesifik maupun domain teks campuran dapat di lihat pada Gambar 5.



Gambar 5. Proses Evaluasi Secara Manual

IV. HASIL DAN ANALISIS

A. *Perancangan Pengujian dan Evaluasi Hasil Terjemahan*

Tahapan implementasi mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia terlebih dahulu korpus teks paralel yang telah dibuat dilakukan proses *cleaning*, *tokenizing*, dan *lowercase* baik untuk domain teks spesifik maupun untuk domain teks campuran. Berikut merupakan perintah untuk melakukan *cleaning*, *tokenizing* dan *lowercase* yang dapat dilihat pada Gambar 6.

```
#cleaning
perl clean-corpus-n.perl corpus en id korpus.clean 1 50
#tokenizing
perl tokenizing.plx corpus.clean.en corpus.tokenized.en
perl tokenizing.plx corpus.clean.id corpus.tokenized.id
#lowercase
perl lowercase.perl corpus.clean.en corpus.lowercased.en
perl lowercase.perl corpus.clean.id corpus.lowercased.id
```

Gambar 6. Perintah Proses *Cleaning*, *Tokenizing*, dan *Lowercase*

B. *Implementasi SRILM untuk Pemodelan Bahasa*

Pemodelan bahasa dilakukan untuk mendapatkan model bahasa dari bahasa target yaitu bahasa Indonesia. Model bahasa yang digunakan dalam penelitian ini yaitu *n-gram* data. Model bahasa dibangun dengan *tools* SRILM. Berikut merupakan perintah untuk membangun model bahasa pada gambar 7.

```
/home/khamsah/NLP/srilm/bin/i686-ubuntu/ngram-count -order 3 -interpolate -unk -text corpus.lowercased.id -lm corpus.id.lm
```

Gambar 7. Perintah Membangun Model Bahasa

C. Implementasi GIZA++ untuk Pemodelan Translasi

Pemodelan translasi digunakan untuk memasangkan teks *input* dalam bahasa sumber dengan teks *output* dalam bahasa target. Model translasi dibangun dengan *tools* Giza++. Berikut merupakan perintah untuk membangun model translasi yang dapat dilihat pada Gambar 8.

```
/home/khamsah/NLP/mosesdecoder-master/scripts/training/train-model.perl -external-bin-dir /home/khamsah/NLP/giza/GIZA++-v2 -root-dir . --corpus corpus.lowercased --f en --e id --lm 0:3:$(pwd)/corpus.id.lm:0
```

Gambar 8. Perintah Membangun Model Translasi

Berdasarkan Gambar 8 proses pemodelan translasi oleh Giza++ menghasilkan dokumen *vocabulary corpus, word alignment* dan tabel model translasi. Dokumen-dokumen tersebut terdapat dalam folder *"train"* yang didalamnya terdapat 4 *file* yaitu *"giza.en-id, giza.id-en, trained.corpus.en-id, trained.model.en-id".*

D. Decoding

Proses *decoding* digunakan untuk menemukan teks dalam bahasa target yang memiliki probabilitas paling besar dengan pertimbangan faktor *translation model* dan *language model*. Tools yang digunakan untuk proses decoding adalah Moses. Berikut merupakan perintah untuk melakukan *decoding* dengan moses pada Gambar 9.

```
/home/khamsah/NLP/mosesdecoder-master/moses-cmd/bin/gcc-7/release/link-static/threading-multi/moses -f model/moses.ini
```

Gambar 9. Perintah *Decoding* oleh Moses

E. Pengujian Hasil terjemahan Secara Otomatis

Evaluasi secara otomatis dilakukan dengan BLEU *(Bilingual Evaluation Understudy).* Pengujian hasil terjemahan mesin penerjemah bahasa Inggris ke bahasa Indonesia terhadap kuantitas korpus menggunakan kalimat uji dengan jumlah korpus maksimal. Gambar 10 merupakan perintah membuat output dalam bahasa target.

```
/home/khamsah/NLP/mosesdecoder-master/moses-cmd/bin/gcc-7/release/link-static/threading-multi/moses -f model/moses.ini < corpusuji.en > translated.corpusuji.en
```

Gambar 10. Perintah Membuat *Output*

Setelah membuat *output* berupa hasil terjemahan 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. Perintah untuk menghitung skor terdapat pada Gambar 11.

```
/home/khamsah/NLP/mosesdecoder-master/scripts/generic/multi-bleu.perl corpusuji.id < translated.corpusuji.en
```

Gambar 11. Perintah Menghitung Skor *Output*

Untuk pengujian otomatis dilakukan pada setiap mesin penerjemah yang sudah dibangun dengan pembagian *fold* pada korpus. Pengujian yang dilakukan terhadap kuantitas korpus terdiri dari 5500 korpus diantaranya 500 untuk korpus uji dan 5000 untuk korpus mesin spesifik dan mesin campuran. Mesin yang dibangun berjumlah 11 mesin yang terdiri dari 500 korpus uji Inggris-Indonesia berdasarkan korpus spesifik. Setiap mesin menghasilkan nilai BLEU yang berbeda. Setelah membuat *output* berupa hasil terjemahan otomatis dari mesin penerjemah statistik, langkah selanjutnya adalah mendapatkan nilai BLEU dari *output* dengan cara membandingkan *output* tersebut dengan korpus bahasa target yang telah dibuat sebelumnya. Secara umum hasil pengujian terjemahan terhadap kuantitas korpus diperlihatkan pada Tabel 3.

TABEL 3
HASIL PENGUJIAN PENERJEMAHAN TERHADAP KUANTITAS KORPUS

Mesin	Hasil Mesin Spesifik	Hasil Mesin Campuran
1	16,63%	8,95%
2	18,06%	7,36%
3	15,03%	7,28%
4	16,01%	7,77%
5	16,99%	8,43%
6	15,19%	6,60%
7	16,60%	8,42%
8	17,31%	6,93%
9	19,33%	7,15%
10	18,38%	7,81%
11	18,84%	7,35%
Rata-Rata	17,1245%	7,6409%

Berdasarkan Tabel 3, diperoleh nilai BLEU pada mesin penerjemah statistik Bahasa Inggris-Indonesia pada korpus spesifik dengan nilai rata-rata 17,1245%. Nilai rata-rata BLEU pada mesin penerjemah statistik bahasa Inggris-Indonesia pada korpus campuran adalah 7,6409%.

Berikut merupakan perhitungan perbandingan nilai BLEU dari mesin penerjemah statistik Bahasa Inggris-Indonesia pada korpus spesifik dan campuran.

Perbandingan=Nilai BLEU Spesifik-Nilai BLEU Campuran

Perbandingan=17,1245-7,6409

Perbandingan=9,4836

Dengan demikian penilaian otomatis terhadap hasil terjemahan seluruh kalimat uji pada mesin penerjemah statistik bahasa Inggris-Indonesia memiliki perbandingan nilai BLEU antara domain teks spesifik dan domain teks campuran yaitu sebesar 9,4836%.

F. Pengujian Hasil Terjemahan Secara Manual

Pengujian secara manual dilakukan oleh ahli bahasa Inggris. Dalam penelitian ini ahli bahasa dilakukan oleh Antonius Yonathan seorang mahasiswa Program Studi Informatika, Universitas Tanjungpura. Pengujian secara manual menggunakan kalimat dari kalimat uji yang diambil dari setiap pembagian *fold*. Jumlah kalimat uji 200 kalimat diambil dari 100 kalimat dari korpus spesifik dan diambil dari 100 kalimat dari korpus campuran dimana bahasa Inggris sebagai kalimat bahasa sumber dan bahasa Indonesia sebagai kalimat bahasa target. Berdasarkan hasil penilaian dari ahli Bahasa, maka dapat dilakukan perhitungan akurasi manual sebagai berikut.

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$$

dengan:

$\bar{x}$ = nilai rata-rata *(mean)* akurasi terjemahan

$\sum x$ = total skor dari bobot penilaian

n = banyaknya data

Perhitungan akurasi manual menggunakan dilakukan dengan range nilai skala peringkat untuk setiap kalimat uji yaitu 1-5 dimana :

1 = Sangat Buruk
2 = Buruk
3 = Cukup
4 = Baik
5 = Sangat Baik

TABEL 4
HASIL PENGUJIAN TERHADAP KALIMAT UJI

Domain Teks	Jumlah Kalimat Pada Nilai Skala				
	1	2	3	4	5
Spesifik	0	2	72	21	5
Campuran	7	60	31	2	0

Pada Tabel 4 terlihat hasil pengujian ahli bahasa pertama pada mesin penerjemah statistik bahasa Inggris – Indonesia untuk domain teks spesifik diperoleh 0 kalimat pada nilai skala satu, 2 kalimat pada nilai skala dua, 72 kalimat pada nilai skala tiga, 21 kalimat pada nilai skala empat, dan 5 kalimat pada nilai skala lima. Sedangkan mesin penerjemah statistik bahasa Inggris – Indonesia pada domain teks campuran diperoleh nilai skala satu 7 kalimat, nilai skala dua 60 kalimat, nilai skala tiga 31 kalimat, nilai skala empat 2 kalimat, dan nilai skala lima 0 kalimat. Berdasarkan persamaan perhitungan sehingga total nilai skala untuk mesin penerjemah statistik bahasa Inggris – Indonesia pada domain teks spesifik adalah 329 dan mesin penerjemah statistik bahasa Inggris – Indonesia pada domain teks campuran adalah 228.

Secara umum hasil penilaian manual bahasa Inggris ke Inggris dapat dilihat pada Tabel 5.

TABEL 5
HASIL PENILAIAN MANUAL OLEH AHLI BAHASA

Pengujian Secara Manual	Korpus	$\sum x$, n	Nilai Akurasi
	Spesifik	329, 100	3,29
	Campuran	228, 100	2,28

Dari tabel di atas diperoleh nilai akurasi hasil pengujian manual mesin penerjemah statistik bahasa Inggris ke bahasa Indonesia, yaitu 3,29 untuk mesin penerjemah statistik pada korpus spesifik dan 2,28 untuk mesin penerjemah statistik pada korpus campuran.

Berikut merupakan perhitungan untuk perbandingan hasil pengujian manual oleh ahli bahasa dari mesin penerjemah statistik untuk domain teks korpus spesifik dan korpus campuran.

Perbandingan=3,29-2,28

Perbandingan=1,01

Penilaian ahli bahasa terhadap hasil terjemahan 100 kalimat uji pada mesin penerjemah statistik untuk korpus spesifik dan 100 kalimat uji pada mesin penerjemah statistik untuk korpus campuran memiliki perbandingan sebesar 1,01%.

G. Analisis Hasil Pengujian

Dari dua pengujian yang sudah dilakukan, dapat disimpulkan beberapa poin hasil analisis yang didapatkan.

1. Pada pengujian otomatis menggunakan BLEU, akurasi terjemahan pada mesin penerjemah statistik bahasa Inggris-Indonesia pada domain teks korpus spesifik diperoleh nilai BLEU rata-rata sebesar 17,1245%. Sedangkan pada mesin penerjemah statistik bahasa Inggris-Indonesia pada domain teks korpus campuran, diperoleh akurasi terjemahan rata-rata sebesar 7,6409%. Dengan demikian, tampak bahwa terlihat perbandingan akurasi terjemahan sebesar 9,4836%.
2. Pada pengujian manual yang dilakukan oleh ahli bahasa, diperoleh nilai akurasi sebesar 3,29% untuk domain teks spesifk dan diperoleh nilai akurasi sebesar 2,28% untuk

domain teks campuran. Dengan demikian, tampak terlihat perbandingan akurasi terjemahan sebesar 1,01%.

3. Dari kedua pengujian, yaitu pengujian otomatis dan pengujian manual terlihat pada domain teks spesifik memiliki tingkat akurasi yang lebih tinggi daripada domain teks campuran, penulis menyimpulkan bahwa adanya pengaruh penggunaan domain teks pada korpus dikarenakan menggunakan domain teks spesifik sebagai korpus uji pada mesin penerjemah statistik bahasa Inggris-Indonesia, yang menyebabkan domain teks spesifik memiliki nilai BLEU yang lebih baik dari nilai BLEU pada domain teks campuran.

REFERENSI

[1] Sinclair, J. 2004. Intuition and annotation - the discussion continues. In Advances in corpus linguistics. *Papers from the 23rd International Conference on English Language Research on Computerized corproa (ICAME 23).Göteborg 22-26 May 2002.* , eds. Karin Aijmer and Bengt Altenberg, 39-59. Amsterdam/New York: Rodopi.

[2] Daume III, Hal and Jagarlamudi, Jagadeesh (2011): *Domain Adaptation for Machine Translation by Mining Unseen Words,* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies.

[3] Jia Xu and Yonggang Deng and Yuqing Gao and Hermann Ney (2007): *Domain Dependent Statistical Machine Translation*, Proceedings of the MT Summit X.

[4] Wu, Hua and Wang, Haifeng and Zong, Chengqing (2008): *Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora*, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008).

[5] Nakov, Preslav and Ng, Hwee Tou (2009): NUS at WMT09: *Domain Adaptation Experiments for English-Spanish Machine Translation of News Commentary Text*, Proceedings of the Fourth Workshop on Statistical Machine Translation.

[6] Andre Castilla and Alice Bacic and Sergio Furuie (2005): *Machine Translation on the Medical Domain: The Role of BLEU/NIST and METEOR in a Controlled Vocabulary Setting*, Proceedings of the Tenth Machine Translation Summit (MT Summit X).

[7] McEnery, T. & Gabrielatos, C. (2006). *English corpus linguistics. In Aarts, B. & McMahon, A. (Eds.), The Handbook of English Linguistics (pp. 33-71).* Oxford: Blackwell.

[8] Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

[9] Amalia, Farida. 2009. *"Ideologi dalam Penerjemahan"*. Makalah disajikan dalam Forum Ilmiah Pengajar Bahasa Prancis Prancis se Indonesia di Bandung.

[10] Sudarno, A.P. 2011. *Penerjemahan Buku Teori dan Aplikasi.* Surakarta :UNS Press.

[11] Sheddy, N. Tjandra. 2005. *Analisis Penerjemahan.* Jakarta, library UI Vol 8 No 1, hal 168-173, 2005.

[12] Tanuwijaya, Hansel. 2009. *Penerjemahan Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan Word Reordering dan Phrase Reordering*. Jakarta, Jurnal ilmu Komputer dan Informasi Vol 2 No 1, hal 17-24, 2009.

[13] 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, hal 127-135, 2014.

[14] Manning, Christopher D. dan Schutze, Hinrich. 2000. Foundations Of Statistical Natural Language Processing. London : The MIT Press Cambridge Massachusetts.

[15] Sujaini, Herry., Negara, Arif Bijaksana Putra. 2015. *Analysis of Extended Word Similarity Clustering based Algorithm on Cognate Language.* Gujarat: ESRSA Publications Pvt. Ltd.

[16] Y.Jarob, H. Sujaini dan N. Safriadi. 2016. *Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan*. JEPIN, Vol. 2 No. 2.

[17] Hasbiansyah, Muhammad. 2016. *Tuning For Quality Untuk Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia - Bahasa Dayak Kanayatn*. Pontianak, JUSTIN Vol 1 No 1, hal 1-6, 2016.

[18] 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.

[19] Papineni, Kishore., Roukos, Salim., Ward, Todd., and Zhu, Wei-Jing. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, Juli 2002. IBM T. J. Watson Research Center.

[20] Triawati, Candra. 2009. *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia*. Jakarta: IT TELKOM.