JUSTIN
Jurnal Sistem dan Teknologi Informasi
ISSN : 2460-3562

Vol. 6, No. 2, April 2018

Algortima Pembagian Frasa dalam Kalimat untuk Menigkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Bugis Wajo

Mulyana[#1], Herry Sujaini[#2], Helen Sasty Pratiwi[#3]
[#]*Program Studi Informatika Universitas Tanjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak 78124
[1]ummufaizyana@gmail.com
[2]herry.sujaini@ee.untan.ac.id
[3]helensastypratiwi@gmail.com

Abstrak*—Salah satu cara yang digunakan untuk meningkatkan nilai akurasi hasil terjemahan adalah dengan melakukan pembagian frasa dalam kalimat korpus. Tujuan yang ingin dicapai dalam penelitian ini adalah untuk menigkatkan akurasi mesin penerjemah statistik Bahasa Indonesia – Bahasa Bugis Wajo. Penelitian ini mengimplementasikan algoritma pembagian frasa pada mesin penerjemah statistik, serta melakukan pengujian akurasi berdasarkan pembagian frasa dalam kalimat pada korpus paralel, dengan melakukan pemenggalan kalimat. Pemenggalan kalimat korpus dapat dilakukan dengan kondisi kalimat memiliki kata kunci seperti tanda koma, kata penghubung, kata negatif, kata keterangan penguat, kata tingkat perbandingan, kata yang menyatakan keadaan, kata depan dan adverbia. Pengujian dilakukan dengan membandingkan nilai akurasi hasil terjemahan tanpa dan dengan pembagian frasa dalam kalimat korpus. Pengujian dilakukan dengan cara otomatis menggunakan *bilingual evaluation understudy* (BLEU). Hasil dari pengujian algoritma pembagian frasa yang diimplementasikan pada mesin penerjemah statistik bahasa Indonesia – bahasa Bugis Wajo mengalami peningkatan. Nilai akurasi sebesar 59,15% meningkat dari mesin tanpa pembagian frasa dan sebesar 0,07% meningkat dari mesin dengan pembagian frasa tujuh algoritma.

***Kata Kunci*— bleu *score*, mesin penerjemah, mesin penerjemah statistik, algoritma pembagian frasa, bugis wajo.**

I. PENDAHULUAN

Bahasa Bugis Wajo fungsi-fungsi ideal, yaitu sebagai lambang identitas dan kebanggaan etnik, sebagai sarana komunikasi intraetnik, dan sebagai pemerkaya bahasa Indonesia. Fungsi-fungsi ini secara perlahan-lahan mengalami pengurangan, terutama pada generasi sekarang. Berbagai upaya telah dilakukan untuk mempertahankannya, misalnya bahasa daerah dijadikan salah satu mata pelajaran muatan lokal pada tingkat sekolah dasar, diadakan penelitian dan seminar dari waktu ke waktu, dan dibuka program studi atau jurusan sastra daerah di perguruan tinggi. Akan tetapi, semuanya ini tidak dapat menjadi solusi memadai untuk mempertahankannya[1].

Perkembangan teknologi yang pesat dapat mempengaruhi semua aspek kehidupan. Saat ini salah satu teknologi yang sedang dikembangkan, yaitu mesin penerjemah untuk mengatasi masalah penerjemahan bahasa. Akan tetapi, kualitas dari hasil terjemahan yang dihasilkan masih mengandung keterbatasan, yakni belum memberikan hasil terjemahan yang akurat. Akurasi hasil terjemahan dapat mempengaruhi beberapa faktor, salah satunya adalah frasa.

Hewavitharana melakukan penelitian dengan frasa benda pada mesin penerjemah statistik dengan menandai frasa benda pada kalimat dan menggantinya dengan frasa benda yang baru. Penelitian tersebut mendapatkan hasil yang lebih baik [2]. Demikian pula hasil penelitian Wibowo, yaitu penelitian tentang algoritma pembagian frasa dalam kalimat untuk meningkatkan akurasi mesin penerjemah statistik bahasa Indonesia – bahasa Jawa Kromo. Penelitian dilakukan dengan menggunakan tujuh aturan algortima yang digunakan untuk memenggal kalimat menjadi dua atau lebih dengan syarat memenuhi kondisi yang sudah ditentukan. Dimana proses tersebut dapat meningkatkan nilai akurasi terjemahan sebesar 12.75% [3]. Chiang melakukan penelitian dengan menyajikan model terjemahan berbasis frasa statistik yang menggunakan frasa hierarki - frasa yang mengandung sub - frasa. Model ini secara formal merupakan tata bahasa tanpa konteks sinkron namun dipelajari dari *bitext* tanpa informasi sintaksis, dengan demikian dapat dilihat sebagai perubahan mesin formal dari sistem terjemahan berbasis sintaks tanpa ada komitmen (ketetapan) linguistik. Penelitian tersebut menggunakan BLEU sebagai metrik, hierarchical phrasebased model mencapai peningkatan yang relatif, yaitu sebesar 7,5% [4].

Berdasarkan masalah yang telah dipaparkan, maka pada penelitian ini akan mengimplementasikan algoritma pembagian frasa pada mesin penerjemah statistik, serta melakukan pengujian akurasi berdasarkan pembagian frasa dalam kalimat pada korpus paralel, dimana akan dilakukan

Submitted 13-02-2018; Revised 20-02-2018; Accepted 22-02-2018

pemenggalan kalimat berdasarkan frasa sesuai dengan kategori sintaksis. Percobaan yang akan dilakukan menggunakan korpus paralel berupa bahasa Bugis Wajo dan bahasa Indonesia dengan memenggal sebuah kalimat menjadi dua buah kalimat atau lebih. Korpus tersebut dilakukan selanjutnya uji akurasi penerjemahan untuk mengetahui pengaruh dari pembagian frasa dalam kalimat terdap hasil terjemahan.

II. TINJAUAN PUSTAKA

A. Algoritma

Algoritma merupakan sebuah runtutan/urutan langkah-langkah perhitungan yang mengubah masukan (*input*) menjadi keluaran (*output*). Algoritma bisa juga dimaknai sebagai alat dalam menyelesaikan masalah perhitungan yang spesifik [5].

B. Frasa

Frasa adalah gabungan dua kata atau lebih yang membentuk suatu kesatuan makna. Frasa bukan merupakan kalimat, karena tidak memiliki unsur-unsur kalimat. Akan tetapi, frasa dapat menjadi penyusun kalimat yang menduduki unsur tertentu.

Istilah frasa dalam *Statical Machine Translation* (SMT) adalah frasa yang mana ia dapat berupa *substring* apa saja, dan tidak mesti dimaknai sebagaimanamakna frasa yang umumnya dipahami dalam teori sintaksis memungkinkan model ini untuk mempelajari *local reorderings*, penerjemahan idiom-idiom singkat, atau penyisipan dan penghapusan yang peka terhadap muatan/konteks lokal. Frasa-frasa tersebut adalah mekanisme yang demikian sederhana dan kuat/berpengaruh bagi mesin penerjemah [4].

C. Mesin Penerjemah Statistik

Statical machine translation (SMT) adalah suatu paradigma dari mesin penerjemah dimana penerjemahan dilakukan berbasis model statistik dengan parameter-parameter yang diturunkan dari analisis paralel korpus [6].

D. Language Model

Language model merupakan sumber pengetahuan yang penting dalam mesin penerjemah statistik. *Language model* digunakan pada aplikasi *Natural Language processing* seperti *speech recognition*, *part-of-speech*, *tagging* dan *symantic parsing* [7]. Dalam *language model* statistik, bagian-bagian yang merupakan elemen kunci adalah probabilitas dari rangkaian-rangkaian kata yang dituliskan sebagai P($w1$, $w2$,…,wn) atau P($w1$, wn). *Language model* menetapkan probabilitas P($w1$, wn) ke serangkaian n kata dengan *means* sebuah distribusi probabilitas. Rangkaian-rangkaian tersebut bisa berupa frase-frase atau kalimat-kalimat dan probabilitasnya dapat diperkirakan dari korpus dokumen-dokumen yang besar. Salah satu contoh pendekatan language model adalah n-gram model. Model bahasa n-gram merupakan jenis probalilistik *language model* untuk memprediksi item berikutnya dalam urutan tersebut dalam bentuk (n-1).

E. Translation Model

Translation model digunakan untuk memasang teks input dalam bahasa sumber dengan teks *output* dalam bahasa sasaran. Dalam mesin penerjemah statistik terdapat dua model penerjemahan, yaitu *word-base translation model* (model translasi berbasis kata) dan *pharase-base translation model* (model traslasi berbasis frase) [8].

F. Decoder

Decoder bertugas menemukan teks dalam bahasa target yang memiliki probabilitas paling besar dengan pertimbangan faktor *translation model* dan *language model*.

G. SRILM

SRILM merupakan sebuah software yang digunakan untuk membangun dan menerapkan *languange model* statistik (LMs). SRILM biasanya digunakan untuk *keperluan speech recognition, statistical tagging*, dan mesin penerjemah [9].

H. Autommatic Evalution

Sistem evaluasi 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" [10].

Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score.* Semakin tinggi nilai BLEU, maka semakin akurat dengan rujukan. Nilai dari BLEU berada pada rentang 0 sampai 1. Suatu terjemahan akan mencapai nilai 1 jika terjemahan tersebut identik dengan terjemahan rujukan. Oleh karena itu, meskipun dengan penerjemahan oleh manusia tidak mungkin akan menghasilkan nilai 1. Sangat penting untuk diketahui bahwa semakin banyak terjemahan rujukan per kalimatnya, maka akan semakin tinggi nilainya. Untuk menghasilkan nilai BLEU yang tinggi, panjang kalimat hasil terjemahan harus mendekati panjang dari kalimat referensi dan kalimat hasil terjemahan harus memiliki kata dan urutan yang sama dengan kalimat referensi. Rumus BLEU berikut [8].

$$BP_{BLEU} = f(x) = \begin{cases} 1, & if\ c > r \\ e^{(1-r/c)}, & if\ c \leq r \end{cases} \quad (1)$$

$$P_n = \frac{\sum_{C \epsilon corpus n-gram \epsilon C} \sum count_{clip}(n-gram)}{\sum_{C \epsilon corpus n-gram \epsilon C} \sum count(n-gram)} \quad (2)$$

$$BLEU = BP_{BLEU} . e^{\sum_{n-1}^{N} w_n log P_n} \quad (3)$$

Keterangan:
BP = *brevity penalty*
c = jumlah kata dari hasil terjemahan otomatis
r = jumlah kata rujukan
Pn = modified precission score
wn = 1/N (standar nilai N untuk BLEU adalah 4)
pn = jumlah *n-gram* hasil terjemahan yang sesuai dengan rujukan dibagi jumlah n-gram hasil terjemahan

Korespondensi : Mulyana

I. Korpus Paralel

Korpus paralel adalah pasangan korpus yang berisi kalimat-kalimat dalam suatu bahasa dan terjemahannya. Korpus paralel merupakan bahan penting untuk melakukan eksperimen-eksperimen dalam bidang pemrosesan bahasa alami.

J. Flowchart

Flowchart adalah bagan-bagan yang mempunyai arus yang menggambarkan langkah-langkah penyelesaian suatu masalah. *Flowchart* merupakan cara penyajian dari suatu algoritma [11].

III. METOLOGI PENELITIAN

A. Data Penelitian

Data penelitian yang digunakan berupa bahasa Indonesia dari korpus penelitian sebelumnya dalam Tugas Akhir (TA) Mahasiswa. Data-data yang diperoleh tersebut selanjutnya diolah menjadi korpus teks pararlel bahasa Indonesia dan bahasa Bugis Wajo.

B. Perangkat Penelitian

Perangkat penelitian yang digunakan dalam penelitian ini terdiri dari perangkat keras dan perangkat lunak. Adapun perangkat yang digunakan diantaranya, yaitu perangkat keras Laptop Toshiba Satellite Pro C640 dengan spesifikasi Posesor Intel Core i3 CPU M 380, 2.53GHz, Ram 3072MB, VGA Intel HD Graphics, HDD 320GB. Perangkat lunak yang digunakan dalam penelitian ini adalah, Sistem Operasi Linux Ubuntu 14.04 LTS 64 Bit, SRILM untuk pemodelan bahasa, Giza++ untuk pemodelan translasi, Moses untuk *decoding*, BLEU untuk pengujian akurasi, Sublime Text 2 untuk teks *editor*.

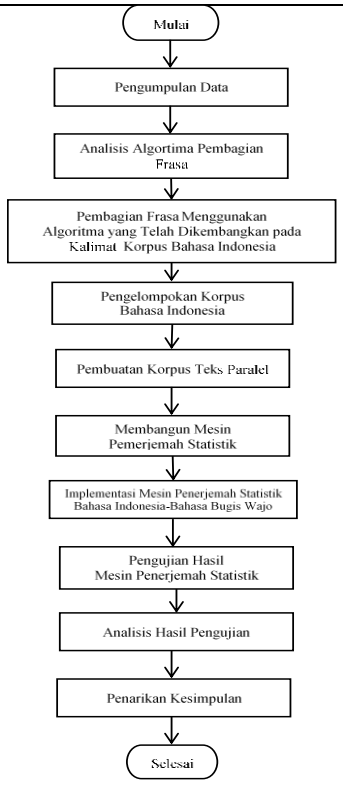
C. Metodologi Penelitian

Metode penelitian yang akan dilakukan dijelaskan pada diagram alir penelitian pada Gambar 1.

Tahapan pada penelitian ini adalah sebagai berikut:

1. Pengumpulan Data
   Proses pengumpulan data merupakan proses mengumpulkan data-data yang akan digunakan untuk penelitian. Data adalah berupa korpus bahasa Indonesia yang akan dibuat menjadi korpus teks paralel.
2. Analisis Algoritma Pembagian Frasa
   Proses analisis algoritma pembagian frasa adalah degan cara menelaah penelitian sebelumnya, yaitu Algoritma Pembagian Frasa dalam Kalimat untuk Meningkatkan Akurasi Mesin Perjemah Statistik Bahasa Indonesia – Bahasa Jawa Kromo [3]. Setelah itu mecari kekurangan/kelemahan yang ada pada penelitian tersebut. Kemudian merencanakan perbaikan dengan cara mengembangkan algoritma pembagian frasa dalam kalimat. Serta mencari referensi yang bisa digunakan dalam proses pengembagan algortima.
3. Pembagian Frasa dalam Kalimat Korpus Bahasa Indonesia
   Pembagian frasa dalam kalimat dilakukan pada korpus bahasa Indonesia yang sudah diperoleh sebelumnya menggunakan pengembagan algoritma pembagian frasa. Pembagian frasa yang dimaksud adalah membagi kalimat menjadi dua atau lebih dengan syarat memenuhi kondisi yang ada. Pembagian dilakukan dengan kondisi kalimat mempunyai kata kunci. Adapun kata kuncinya, yaitu tanda koma, kata penghubung, kata negatif, kata keterangan penguat, kata tinggkat perbandingan, kata yang menyatakan keadaan, kata depan dan adverbia. Kata kuci tersebut termasuk dalam jenis frasa golongan adjektiva, verba dan proposiona yang berdasarkan kategori sintaksis. Algortima pembagian frasa dalam kalimat korpus dapat dilihat pada Gambar 2, Gambar 3 dan Gambar 4.
4. Pengelompokkan Korpus Bahasa Indonesia
   Korpus dikelompokkan menjadi tiga kelompok, yaitu kelompok korpus tanpa pembagian frasa, kelompok korpus dengan pembagian frasa menggunakan tujuh Algoritma dan kelompok korpus dengan pembagian frasa menggunakan delapan Algoritma. Masing - masing kelompok korpus tersebut akan dibuat korpus teks paralel.
5. Pembuatan Korpus Teks Paralel
   Korpus teks paralel dibuat dari terjemahan kalimat-kalimat dari korpus bahasa indonesia tanpa pembagian frasa dan korpus pembagian frasa atau pemotongan kalimat. Kemudian korpus tersebut diterjemahkan ke dalam bahasa Bugis Wajo. Penerjemahan korpus dilakukan oleh ahli bahasa.
6. Membangun Mesin Penerjemah Statistik
   Membangun mesin penerjemah statistik dilakukan dengan melakukan instalasi perangkat lunak yang dibutuhkan diantaranya Linux Ubuntu 14.04 LTS, GIZA++ untuk pemodelan translasi, MOSES untuk *decoding*, SRILM untuk pemodelan bahasa dan BLEU untuk pengujian akurasi secara otomatis.
7. Implementasi Mesin Penerjemah Statistik Bahasa Indonesia Ke Bahasa Bugis Wajo. Implementasi dilakukan dengan cara melakukan pemodelan bahasa, pemodelan translasi dan *decoding* pada mesin penerjemah statistik bahasa Indonesia ke bahasa Bugis Wajo.
2. Pengujian Hasil Penerjemahan Mesin
   Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin penerjemah statistik dari kelompok korpus yang sudah dibuat sebelumnya. Pengujian dilakukan secara otomatis dengan menggunakan BLEU.
3. Analisis Hasil Pengujian
   Analisis Hasil pengujian dilakukan untuk mengetahui karakteristik mesin penerjemah statistik bahasa Indonesia ke bahasa Bugis Wajo dan mengidentifikasi apakah sudah sesuai dengan kebutuhan serta membandingkan nilai akurasi mesin penerjemah statistik sebelum melewati proses pembagian frasa dengan tujuh algortima dan delapan algortima.
4. Penarikan Kesimpulan
   Kesimpulan dirumuskan berdasarkan pengujian yang telah dilakukan apakah sistem yang diracang mampu memberikan solusi berdasarkan permasalahan yang ada.

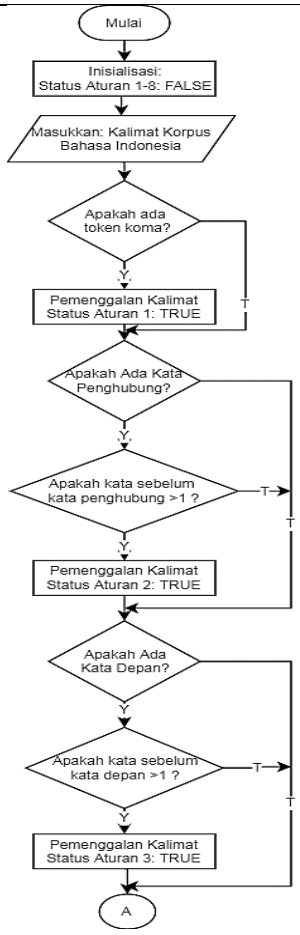
Korespondensi : Mulyana



Gambar. 1. Diagram Alir Penelitian

Gambar 2, Gambar 3 dan Gambar 4 merupakan algoritma pembagian frasa dalam kalimat korpus. Algoritma pembagian frasa terdiri dari 13 proses, kemudian pada Gambar 5 dilanjutkan penghapusan kalimat yang sama pada kalimat hasil pemenggalan. Proses-proses tersebut adalah sebagai berikut:

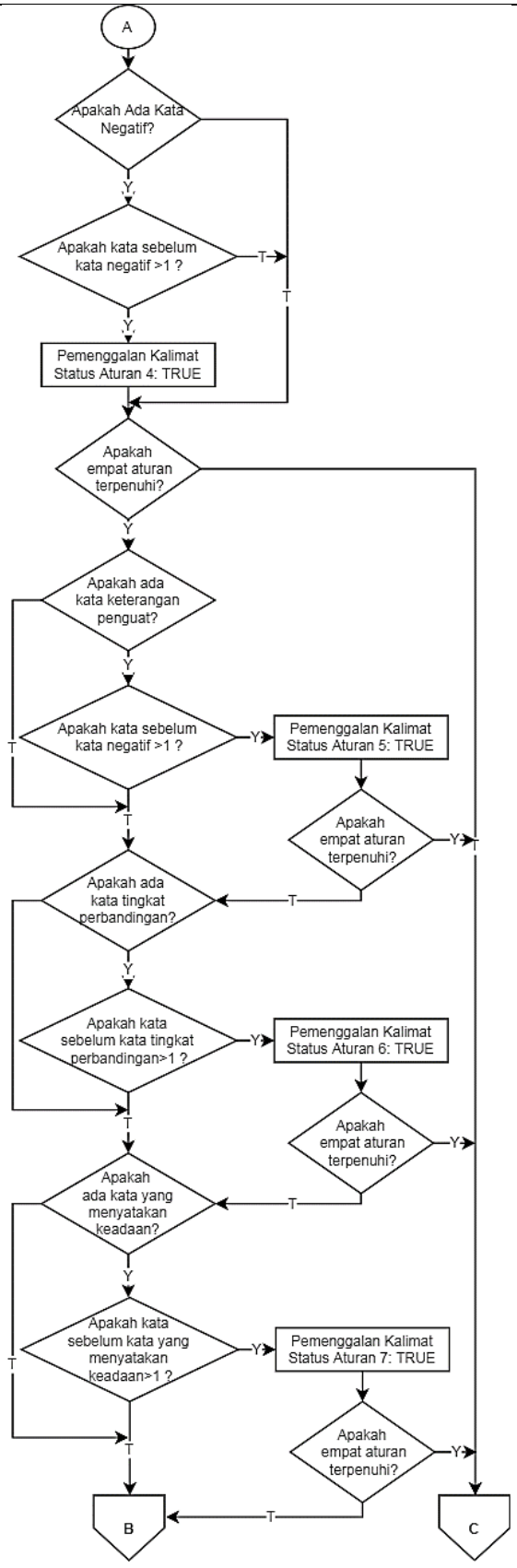
1. Inisialisasi status aturan pertama sampai aturan kedelapan. Kalimat masukan berupa kalimat korpus bahasa Indonesia. Kalimat korpus bahasa Indonesia ini akan melalui maksimal empat proses pemenggalan kalimat.
2. Mengecek apakah kalimat korpus terdapat token koma. Jika kalimat terdapat token koma maka dilakukan proses pemenggalan. Proses pemenggalan dilakukan pada tanda koma tersebut. Kemudian token koma dihapus dan status aturan satu menjadi *true*. Jika tidak terdapat tanda koma maka status aturan pertama tetap *false* dan proses pemenggalan lanjut ke aturan kedua. Contoh: permisi, saya ingin pergi ke dermaga, di penggal menjadi:
   - permisi
   - saya ingin pergi ke dermaga



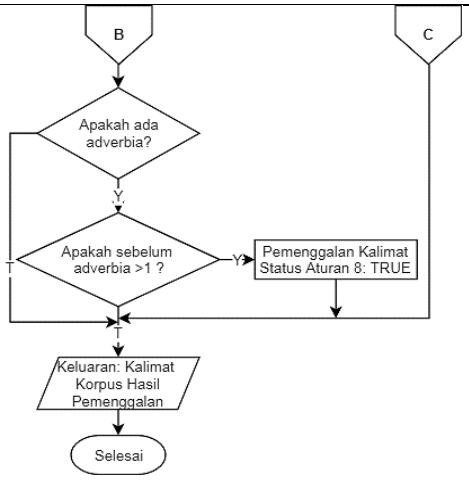
Gambar 2. Algoritma Pembagian Frasa dalam Kalimat Korpus

3. Mengecek apakah kalimat korpus terdapat kata penghubung (dan, atau, serta, tetapi, melainkan, padahal, sedangkan, yang, agar, supaya, biar, jika, kalau, jikalau, asal, asalkan, bila, manakala, sejak, semenjak, sedari, sewaktu, tatkala, ketika, sementara, seraya, sambil, demi, sesudah, sehabis, hingga, andaikan, seandainya, sekiranya, walau, walaupun,sekalipun, sungguhpun, kendatipun, seperti, sebagai, sebagaimana, ibarat, daripada, sebab, karena, maka, dengan, bahwa). Jika tidak maka status aturan kedua tetap *false* dan lanjut aturan berikutnya. Jika ada maka dilanjutkan mengecek apakah sebelum kata penghubung terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum kata penghubung pada kalimat korpus tersebut dan status aturan dua menjadi *true*. Jika hanya satu kata sebelum kata penghubung maka status aturan kedua tetap *false* dan proses pemenggalan lanjut ke aturan tiga. Contoh: beritahu saya ketika kita sampai di museum, dipenggal menjadi:
   - beritahu saya
   - ketika kita sampai di museum

Korespondensi : Mulyana



Gambar 3. Algoritma Pembagian Frasa dalam Kalimat Korpus (b)



Gambar 4. Algoritma Pembagian Frasa dalam Kalimat Korpus (c)

4. Mengecek apakah kalimat korpus terdapat kata depan (di, ke). Jika tidak terdapat kata depan maka status aturan ketiga tetap *false* dan lanjut aturan berikutnya. Jika kalimat terdapat kata depan maka dilanjutkan mengecek apakah sebelum kata depan terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum kata depan pada kalimat korpus tersebut dan status aturan tiga menjadi *true*. Jika hanya satu kata sebelum kata depan maka status aturan ketiga tetap *false* dan proses pemenggalan lanjut ke aturan empat. Contoh: beritahu saya ketika kita sampai di museum, dipenggal mejadi:
   - beritahu saya ketika kita sampai
   - di museum
5. Mengecek apakah kalimat korpus terdapat kata negatif (tidak, bukan). Jika tidak terdapat kata negatif maka status aturan keempat tetap *false* dan lanjut aturan berikutnya. Jika kalimat terdapat kata negatif maka dilanjutkan mengecek apakah sebelum kata negatif terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum kata negatif pada kalimat korpus tersebut dan status aturan keempat menjadi *true*. Jika hanya satu kata sebelum kata negatif maka status aturan keempat tetap *false* dan proses pemenggalan lanjut ke proses selanjutnya. Contoh: mengapa anda tidak pergi ke sana ? dipenggal menjadi:
   - mengapa anda
   - tidak pergi ke sana ?
6. Mengecek apakah empat kondisi utama sudah terpenuhi. Pengecekan dilakukan pada status masing-masing keempat aturan utama tersebut. Jika status keempat aturan *true* maka aturan terpenuhi dan proses pemenggalan selesai. Jika status keempat aturan utama ada terdapat *false* maka dilakukan proses selanjutnya dengan pengecekan dan pemenggalan dengan empat aturan lainnya.
7. Mengecek apakah kalimat korpus terdapat kata

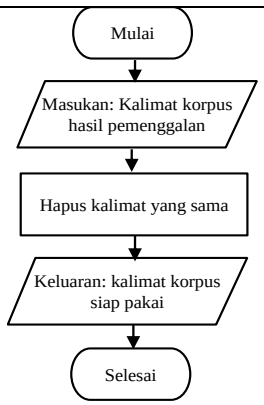
keterangan penguat (sangat, amat, terlalu). Jika tidak terdapat kata keterangan penguat maka status aturan kelima tetap *false* dan lanjut ke aturan keenam. Jika kalimat korpus terdapat kata keterangan penguat maka dilanjutkan mengecek apakah sebelum kata keterangan penguat terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum kata keterangan penguat pada kalimat korpus tersebut dan status aturan kelima menjadi *true*. Jika hanya satu kata sebelum kata keterangan penguat maka status aturan kelima tetap *false* dan proses pemenggalan lanjut ke proses selanjutnya. Contoh: peralatan listrik sangat murah, dipenggal menjadi:

- peralatan listrik
- sangat murah

8. Mengecek apakah empat kondisi sudah terpenuhi. Pengecekan dilakukan pada status masing-masing aturan tersebut. Jika sudah ada empat konsisi yang status aturannya *true* maka aturan terpenuhi dan proses pemenggalan selesai. Jika status belum terpenuhi empat kondisi *true* maka dilakukan proses selanjutnya dengan pengecekan dan pemenggalan dengan tiga aturan lainnya.
9. Mengecek apakah kalimat korpus terdapat kata tingkat perbandingan (lebih). Jika tidak terdapat kata tingkat perbandingan maka status aturan keenam tetap *false* kemudian dilanjutkan pada proses aturan ketujuh. Jika kalimat korpus terdapat kata tingakat perbandingan maka dilanjutkan mengecek apakah sebelum kata tingkat perbandingan terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum kata tingkat perbandingan pada kalimat korpus tersebut dan status aturan keenam menjadi *true*. Jika hanya satu kata sebelum kata keterangan penguat maka status aturan keenam tetap *false* dan proses pemenggalan lanjut ke proses selanjutnya. Contoh: lima puluh sen per kata ini lebih mahal daripada yang saya perkirakan, dipenggal menjadi:
    - lima puluh sen per kata ini
    - lebih mahal daripada yang saya perkirakan
10. Mengecek apakah empat kondisi sudah terpenuhi. Pengecekan dilakukan pada status masing-masing aturan tersebut. Jika sudah ada empat konsisi yang status aturannya *true* maka aturan terpenuhi dan proses pemenggalan selesai. Jika status belum terpenuhi empat kondisi *true* maka dilakukan proses selanjutnya dengan pengecekan dan pemenggalan dengan dua aturan lainnya.
11. Mengencek apakah kalimat korpus terdapat kata yang menyatakan keadaan (sudah, harus, dapat). Jika tidak terdapat kata yang menyatakan keadaan maka status aturan ketujuh tetap *false* kemudian dilanjutkan pada aturan kedelapan. Jika kalimat korpus terdapat kata yang menyatakan keadaan maka dilanjutkan mengecek apakah sebelum kata yang menyatakan keadaan terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum kata yang menyatakan keadaan pada kalimat korpus tersebut dan status aturan ketujuh menjadi *true*. Jika hanya satu kata sebelum kata keterangan penguat maka status aturan ketujuh tetap *false* dan proses pemenggalan lanjut ke proses selanjutnya. Contoh: kapankah saya harus mengembalikannya? dipenggal menjadi:
    - kapankah saya
    - harus mengembalikannya ?
12. Mengecek apakah empat kondisi sudah terpenuhi. Pengecekan dilakukan pada status masing-masing aturan tersebut. Jika sudah ada empat konsisi yang status aturannya *true* maka aturan terpenuhi dan proses pemenggalan selesai. Jika status belum terpenuhi empat kondisi *true* maka dilakukan proses selanjutnya dengan pengecekan dan pemenggalan dengan dengan aturan terakhir.
13. Mengecek apakah kalimat korpus terdapat adverbia (akan, boleh, suka, ingin, mau, sedang, pernah, selalu, masih, sering). Jika tidak terdapat adverbia maka status aturan kedelapan tetap *false*. Jika kalimat korpus terdapat adverbia maka dilanjutkan mengecek apakah sebelum adverbia terdapat lebih dari satu kata. Jika benar maka akan maka proses pemenggalan dilakukan sebelum adverbia pada kalimat korpus tersebut dan status aturan kedelapan menjadi *true*. Jika hanya satu kata sebelum kata keterangan penguat maka status aturan kedelapan tetap *false*. Pada tahap ini berapapun jumlah aturan dengan status *true* maupun *false* pada proses terakhir maka keluaran kalimat langsung menjadi kalimat korpus hasil dari pemenggalan. Contoh: penglihatan anda akan kembali pada waktunya, dipenggal menjadi:
    - penglihatan anda
    - akan kembali pada waktunya

Selanjutnya, penghapusan kalimat yang sama pada kalimat hasil dari pemenggalan. Gambar 5 adalah proses penghapusan kalimat-kalimat yang sama pada kalimat hasil pemenggalan. Korpus yang sudah melalui proses pemenggalan ada beberapa yang sama jadi hanya diambil satu kalimat di antara kalimat-kalimat yang sama. Hasil dari proses ini adalah kalimat – kalimat korpus yang siap digunakan untuk pembuatan korpus teks paralel.

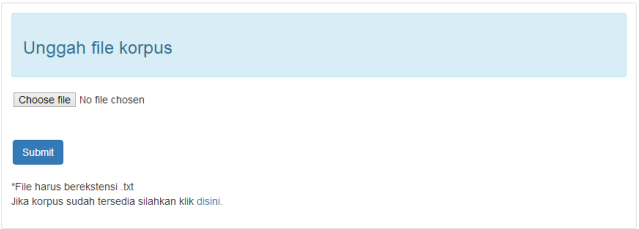
Korespondensi : Mulyana



Gambar 5. Proses Penghapusan Kalimat yang Sama Pada Korpus Hasil dari Pemenggalan

D. Pembagian Frasa dalam Kalimat Bahasa Indonesia

Pembagian frasa dalam kalimat dilakukan pada korpus bahasa Indonesia yang sudah diperoleh sebelumnya. Pembagian frasa yang dimaksud adalah membagi sebuah kalimat menjadi dua atau lebih dengan dengan syarat memenuhi kondisi yang sudah ditentukan. Pada proses ini dibantu apilikasi pembagian frasa dalam kalimat. Dimana aplikasi ini merupakan penerepan dari pengembagan algoritma pembagian frasa dalam kalimat. Gambar 2, Gambar 3, dan Gambar 4 merupakan diagram alir pengembangan algoritma pada aplikasi yang akan dibagun.



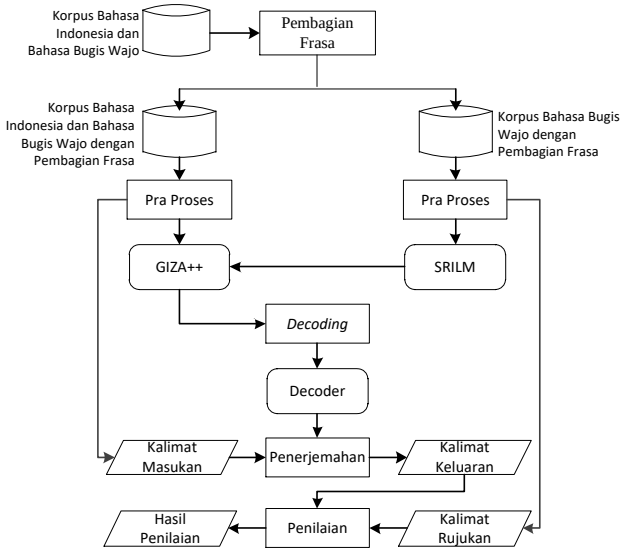
Gambar 6. Tampilan Antarmuka Aplikasi Pembagian Frasa dalam Kalimat Korpus (Halaman Unggah *file* korpus)

Gambar 7. Antarmuka Aplikasi Pembagian Frasa dalam Kalimat Korpus (Halaman Utama)

Tampilan antarmuka pada Gambar 6 merupakan sebuah halaman untuk mengunggah *file* korpus yang akan dilakukan proses pembagian frasa. Tampilam antarmuka pada Gambar 7 terdapat tiga bagian. Bagian pertama berisi keterangan aturan yang digunakan dalam pembagian frasa setiap kalimat, keterangan jumlah korpus sebelum, sesudah proses pembagian frasa setiap kalimat dan jumlah kalimat yang siap digunakan. Bagian kedua terdapat *textarea* yang menampilkan kalimat hasil pembagian frasa dan sebuah tombol untuk memproses korpus hasil pembagian frasa menjadi kalimat korpus yang siap digunakan. Bagian ketiga terdapat *textarea* yang menampilkan kalimat-kalimat korpus yang siap digunakan dan sebuah tombol untuk menyimpan kalimat-kalimat korpus yang siap digunakan. Kalimat-kalimat tersimpan tersebut menjadi sebuah *file text.*

E. Implementasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Bugis Wajo

Arsitektur sistem pada penelitian ini terdiri dari beberapa proses, yaitu pemodelan bahasa, pemodelan translasi, *decoding* dan evaluasi hasil terjemahan. Arsitektur sistem mesin penerjemah statistik dapat dilihat pada Gambar 8.



Gambar 8. Arsitektur Sistem Mesin Penerjemah Statistik Bahasa Indoneisa – Bahasa Bugis Wajo dengan Pembagian Frasa dalam Kalimat

Gambar 8 merupakan arsitektur sistem mesin penerjemah statistik bahasa Indonesia ke bahasa Bugis Wajo. Korpus paralel dalam *file* teks terdiri dari dua buah korpus, yaitu korpus bahasa Indonesia dan korpus bahasa Bugis Wajo. Kemudian korpus teks paralel dilakukan proses pembagian frasa. Setelah itu korpus dengan pembagian frasa dilakukan pra proses sebelum dilakukan pemodelan bahasa oleh SRILM dan pemodelan translasi oleh Giza. Kemudian pra proses akan digunakan sebagai kalimat masukan untuk proses terjemahan dan kalimat rujukan untuk proses penilaian. Kemudia masuk pada proses selanjutnya, yaitu pemodelan bahasa, pemodelan translasi, decoding, dan proses evaluasi hasil terjemahan.

Korespondensi : Mulyana

F. Pengujian Hasil Terjemahan Translasi

Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi dan sebagai awal untuk dibandingkan dengan nilai akurasi setelah dilakukan proses pembagian frasa dalam kalimat. Pengujian dilakukan cara pengujian hasil secara otomatis menggunakan BLEU.

Mesin 1 adalah mesin yang menggunakan kalimat korpus tanpa pembagian frasa, mesin 2 adalah mesin yang menggunaka kalimat korpus dengan pembagian frasa menggunakan tujuh algoritma, dan mesin 3 adalah mesin yang menggunakan kalimat korpus dengan pembagian frasa menggunakan delapan algoritma. Setiap mesin terbagi menjadi lima mesin. Kalimat korpus yang digunakan untuk kalimat *training* adalah 75% dari jumlah kalimat korpus setiap mesin, sedangkan korpus uji yang digunakan berjumlah 650 untuk masing-masing diujikan pada lima mesin, sehingga korpus uji berjumlah 3150.

Berikut merupakan cara perhitungan kenaikan atau peningkatan nilai BLEU.

$$Peningkatan = \frac{Nilai\ BLEU\ mesin\ x - Nilai\ BLEU\ mesin\ y}{Nilai\ BLEU\ mesin\ y}\ x\ 100\% \quad (4)$$

Keterangan :
mesin x: mesin yang ingin dicari nilai kenaikannya.
mesin y: mesin yang akan dibandingkan dengan mesin yang dicari nilai kenaikannya.

G. Analisis Hasil Pengujian

Analisis hasil pengujian dilakukan untuk mengetahui karakteristik mesin penerjemah statistik dan mengidentifikasi apakah sudah sesuai dengan kebutuhan serta membandingkan nilai akurasi mesin penerjemah statistik.
Nilai akurasi yang dibandingkan dibagi menjadi tiga bagian.

1. Nilai akurasi yang didapatkan dari mesin penerjemah tanpa pembagian frasa.
2. Nilai akurasi yang didapat dari mesin penerjemah dengan pembagian frasa menggunakan tujuh algotritma.
3. Nilai akurasi yang didapat dari mesin penerjemah dengan pembagian frasa menggunakan delapan algotritma.

IV. HASIL DAN ANALISIS

A. Implementasi Mesin Penerjemah Statistik Bahasa Bahasa Indonesia – Bahasa Bugis Wajo

Tahapan implementasi mesin penerjemah statistik bahasa Bugis Wajo ke bahasa Indonesia terlebih dahulu korpus teks paralel yang telah dibuat dilakukan proses *cleaning* dan tokenisasi. Berikut merupakan perintah untuk melakukan cleaning dan tokenisasi yang dapat dilihat pada Gambar 9.

```
1  cd ~/moses/mesin3A
2  ~/moses/mosesdecoder/scripts/training/clean-corpus-n.perl txt
   indo8 wajo8 txt.clean 1 40
3  perl ~/moses/clean.plx txt.clean.indo8 txt.clean1.indo8
4  perl ~/moses/clean.plx txt.clean.wajo8 txt.clean1.wajo8
5  perl ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   txt.clean1.wajo8 > txt.lowercased.wajo8
6  perl ~/moses/mosesdecoder/scripts/tokenizer/lowercase.perl <
   txt.clean1.indo8 > txt.lowercased.indo8
```

Gambar 9. *Cleaning*, Tokenisasi, dan *Case Folding* Pada Mesin 3

Berdasarkan Gambar 9 terdapat perintah untuk proses *cleaning*, tokenisasi dan *lowercase*. Perintah tersebut adalah pra proses sebelum dilakukan proses implementasi mesin penerjemah. Pada baris pertama untuk membuka folder mesin penerjemah. Baris kedua digunakan untuk memenggal kalimat yang memiliki kata lebih dari 40, baris ketiga dan keempat merupakan perintah untuk menghapus tanda baca titik di akhir kalimat dan menyisipkan spasi atara kata dan tanda baca. Baris kelima dan keenam digunakan untuk mengubah huruf kapital yang terdapat dalam korpus menjadi huruf kecil.

B. Implementasi Giza++ untuk Pemodelan Translasi

Model translasi digunakan untuk memasangkan teks *input* dalam bahasa sumber dengan teks *output* dalam bahasa target. Model translasi dibangun dengan *tool* Giza++.

```
~/moses/mosesdecoder/scripts/training/train-
model.perl -root-dir .
--corpus txt.lowercased --f indo81 --e
wajo81 --lm 0:3:
/home/ummufaiz/moses/mesin3A/wajo8.lm:0
```

Gambar 10. Perintah Membangun Model Translasi Mesin 3A

Gambar 10 merupakan proses pemodelan translasi oleh Giza++ menghasilkan dokumen *vocabulary corpus*, *word alignment* dan *lexical model table.*

C. *Decoding* oleh Moses

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.

```
1  cd ~/moses/mesin1A
2  ~/moses/mosesdecoder/moses-cmd/src/moses
   -f model/moses.ini < txt.indoa > out
```

Gambar 11. Perintah membuat output mesin 3A

Gambar 11 merupakan perintah untuk membuat translasi otomatis dari mesin penerjemah dari bahasa sumber ke dalam bahasa target. *Decoder* moses akan menerjemahkan kalimat masukan berupa kalimat sumber (Bahasa Indonesia). Selanjutnya kalimat masukan tersebut akan diproses oleh *decoder* moses dan akan menghasilkan kalimat keluaran berupa kalimat hasil terjemahan ke dalam bahasa target (Bahasa Bugis Wajo).

D. Pengujian Hasil Terjemahan Mesin Translasi Oleh BLEU

Pengujian hasil translasi dilakukan dengan cara pengujian otomatis dari mesin penerjemah. Pengujian otomatis dari mesin penerjemah menghasilkan keluaran berupa nilai akurasi yang dihasilkan oleh BLEU *(Billingual Evaluation Understudy)*. Hasil pengujian ini nantinya akan menjadi parameter untuk membandingkannya dengan hasil pengujian setelah dilakukan perbaikan *lexical model.*

Korespondensi : Mulyana

```
1  cd ~/moses/mesin3A
2  ~/moses/mosesdecoder/scripts/generic/multi-
   bleu.perl txt.wajoa < out
BLEU = 24.48, 46.8/27.8/19.9/15.3 (BP=0.975,
ratio=0.976, hyp_len=9116, ref_len=9343)
```

Gambar 12. Perintah menghitung dan hasil skor BLEU mesin 3A

Gambar 12 merupakan perintah untuk penilai otomatis oleh BLEU dari mesin 3A. Adapun hasil keseluruhan dapat dilihat pada tabel 1.

Tabel 1. Hasil Penilai BLEU Pada Mesin Penerjemah Statistik

Hasil Penilaian BLEU			
Fold	**Mesin 1 (%)**	**Mesin 2 (%)**	**Mesin 3 (%)**
A	14,04	24,27	24,48
B	18,36	25,62	25,13
C	13,25	19,35	18,85
D	5,69	8,9	9,61
E	3,11	3,51	3,64
Total Nilai (%)	**10,268**	**16,33**	**16,342**
Persentase Kenaikan (%)		**59,04**	**59,15**
			0,07

Tabel 1 berisi total penilai mesin 1 adalah 10,268%, mesin 2 adalah 16,330% sedangkan mesin 3 adalah 16,342%. Perhitungan hasil penilaian berdasarkan persamaan 4. Hasil Penerjemahan Mesin Translasi. Sebagai contoh perhitungan penilaian, yaitu peningkatan Mesin 3 terhadap Mesin 1 berdasarkan persamaan 4.

$$Peningkatan = \frac{Total\ Nilai\ BLEU\ mesin\ 3 - Total\ Nilai\ BLEU\ mesin\ 1}{Total\ Nilai\ BLEU\ mesin\ 1}\ x\ 100\%$$

$$Peningkatan = \frac{16{,}342 - 10{,}268}{10{,}268}\ x\ 100\%$$

$$Peningkatan = 59{,}15\%$$

Peningkatan hasil akurasi mesin 2 terhadap mesin 1 adalah sebesar 59,04% sedangkan peningkata mesin 3 terhadap mesin 1 adalah sebesar 59,15%. Adapun peningkatan mesin 3 terhadap mesin 2 adalah sebesar 0,07%. Total akurasi tertinggi yaitu pada mesin 3.

E. Analisis Hasil Pengujian

Berikut merupakan analisis terhadaphasil pengujian yang telah dilakukan.

1. Jumlah kalimat korpus bahasa Indonesia sebelum proses pembagian frasa berjumlah 3150 kalimat. Setelah dilakukan pembagian frasa dengan tujuh alogoritma kalimat korpus menjadI 7440 kalimat. Sedangkan setelah dilakukan pembagian frasa dengan delapan algoritma kalimat korpus menjadi 7580 kalimat.
2. Penilaian otomatis terhadap hasil terjemahan seluruh korpus uji pada mesin penerjemah statistik bahasa Indonesia – bahasa Bugis Wajo tanpa pembagian frasa menghasilkan nilai BLEU sebesar 10,268 % . Hasil terjemahan seluruh koprus uji dengan pembagian frasa tujuh algoritma menghasilkan nilai BLEU sebesar 16,330 % . Sedangkan hasil terjemahan seluruh koprus uji dengan pembagian frasa tujuh algoritma menghasilkan nilai BLEU sebesar 16,342 %
3. Penigkatan nilai BLEU mesin penerjemahan statistik dengan pembagian frasa tujuh algortima terhadap hasil terjemahan mesin penerjemah statistik tanpa pembagian frasa adalah 59,04%. Sedangkan penigkatan mesin penerjemah statistik dengan pembagian frasa delapan algoritma terhadap hasil terjemahan mesin penerjemah statistik tanpa pembagian frasa adalah 59,15%. Peningkatan nilai BLEU mesin penerjemah statistik pembagian frasa dengan delapan algoritma terhadap hasil terjemahan mesin penerjemahan statistik dengan pembagian frasa tujuh algortiman adalah 0,07%.

V. KESIMPULAN/RINGKASAN

Berdasarkan hasil analisis dan pengujian, maka kesimpulan yang dapat diambil sebagai berikut.

1. Algoritma pembagian frasa dalam kalimat korpus dapat diimplementasikan dalam mesin penerjemah statistik bahasa Indonesia – bahasa Bugis Wajo.
2. Algoritma pembagian frasa yang diimplementasikan pada mesin penerjemah statistik bahasa Indonesia – Bugis Wajo dapat meningkatkatkan nilai akurasi mesin penerjemahan statistik. Peningkatan sebesar 59,15% terhadap mesin tanpa pembagian frasa. Sedangkan peningkatan 0,07% peningkatan terhadap mesin pembagian frasa pada kalimat dengan tujuh algoritma

DAFTAR PUSTAKA

[1] Darwis, Muhammad. 2011. *Nasib Bahasa Daerah Di Era Globalisasi: Peluang dan Tantangan*. Seminar Pelestarian Bahasa Daerah Bugis Makasar. Parepare, tanggal 15 Oktober 2011. Balitbang Agama Makassar.

[2] Hewavitharana, Sanjika,. Lavie, Alon,. And Vogel, Stephan,. 2007. *Experiments with a Noun-Phrase driven Statistical Machine Translation System*. In Proceeding of MT Summit.

[3] Wibowo, Wasis. 2016. *Algoritma Pembagian Frasa dalam Kalimat Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Jawa Kromo*. Fakultas Teknik Prodi Teknik Informatika Universitas Tanjungpura: Pontianak.

[4] Chiang, David. *A Hierarchical Phrase-Based Model for Statistical Machine Translation*. Proceedings of the 43rd Annual Meeting of the ACL. Ann Arbor, Juni 2005. Association for Computational Linguistics.

[5] Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronald L., Stein, Clifford. 2009. *Introduction to Algorithms Third Edition*. The MIT Press: London.

[6] Hadi, Ibnu. 2014. *Uji Akurasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Melayu Sambas dan Mesin Penerjemahan Statistik Bahasa Melayu Sambas ke Bahasa Indonesia*. Jurnal Sistem dan Teknologi Informasi (JustIN) Vol 2, No 3.

[7] Mandira, Soni. 2016. *Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik.* Jurnal Edukasi dan Penelitian Informatika (JEPIN) Vol. 2, No. 1.

[8] Tanuwijaya, Hansel. 2009. *Penerjemahan Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan Word Reordering dan Phrase Reordering*. Universitas Indonesia: Jakarta.

[9] Stolcke, A., Zheng, J., Wang, W., dan Abrash, V. 2011. *SRILM at Sixteen: Update and Outlook. 13 Januari 2018. https://www.sri.com/sites/default/files/publications/srilm.pdf.*

Korespondensi : Mulyana

[10] 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.
[11] Ladjamuddin, Al-Bahra Bin. (2006). *Rekayasa Perangkat Lunak*. Graha Ilmu: Yogyakarta.

Korespondensi : Mulyana