Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan

Yosep Jarob R[#1], Herry Sujaini[#2], Nofi Safriadi[#3]

[#]*Program Studi Teknik Informatika Fakultas Teknik Universitas Tanjungpura*

[1]yosepjarob11@gmail.com

[2]herry_sujaini@yahoo.com

[3]bangnops@gmail.com

***Abstrak*— Bahasa merupakan alat komunikasi dan kunci pokok yang penting bagi kehidupan manusia, karena dengan menggunakan bahasa kita dapat berinteraksi dan mengetahui informasi yang dibutuhkan, bahasa juga digunakan seseorang untuk menyampaikan ide, gagasan, konsep atau perasaan kepada orang lain. Bahasa yang dimiliki setiap orang berbeda-beda, keragaman bahasa ini dapat menghambat pertukaran informasi karena orang lain tidak memahami maksud dan tujuan yang ingin disampaikan. Oleh karena itu diperlukan penerjemah untuk menjembatani bahasa yang berbeda. Mesin Penerjemah Statistik *(Statistical Machine Translation)* merupakan sebuah pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel. Korpus paralel adalah pasangan korpus yang berisi kalimat-kalimat dalam suatu bahasa dan terjemahannya. Salah satu model yang digunakan untuk menentukan akurasi hasil terjemahan adalah dengan melakukan proses *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya. Tujuan yang ingin di capai dalam penelitian ini adalah menguji akurasi penerjemahan bahasa Indonesia-Dayak Taman dengan membandingkan nilai akurasi sebelum dan setelah dilakukan proses *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya. Penelitian menggunakan korpus paralel sebanyak 3110 korpus. Pengujian dilakukan dengan dua cara, yaitu pengujian otomatis menggunakan *Bilingual Evaluation Understudy* (BLEU) dan pengujian oleh ahli bahasa Dayak Taman. Hasil dari pengujian adalah terdapat peningkatan nilai BLEU sebesar 0.36% pada pengujian otomatis dan 20.57% pada pengujian oleh ahli bahasa.**

***Kata Kunci*— BLEU *score*, Dayak Taman, Indonesia, korpus paralel, mesin penerjemah statistik, *tagging*.**

I. PENDAHULUAN

Bahasa merupakan alat komunikasi dan kunci pokok yang penting bagi kehidupan manusia, karena dengan menggunakan bahasa kita dapat berinteraksi dan mengetahui informasi yang dibutuhkan, dengan bahasa juga manusia dapat mengekspresikan diri, menyampaikan kritik dan pendapat, pikiran serta keinginannya. Bahasa juga merupakan sumber daya bagi kehidupan bermasyarakat, bahasa merupakan alat yang digunakan untuk sosialisasi dan adaptasi sosial antara manusia satu dengan manusia lainnya agar kehidupannya semakin berkembang. berdasarkan catatan Badan Pengembangan dan Pembinaan Bahasa serta Kementerian Pendidikan dan Kebudayaan (Kemendikbud) sedikitnya ada 442 bahasa yang dimiliki Indonesia [1].

Bahasa daerah merupakan simbol paling sempurna sebagai cara pengekspresian tata cara, adat komunikasi sosial, dan paranata sosial. Sama seperti bahasa daerah lain nya, bahasa Dayak Taman juga memiliki fungsi yang sama, yaitu sebagai lambang kebanggaan daerah, lambang identitas daerah, alat penghubung antar warga masyarakat daerah. Namun fungsi dari bahasa daerah tersebut saat ini mengalami banyak penurunan. Salah satu upaya yang dilakukan untuk mempertahankan penggunaan bahasa daerah, yaitu dengan menjadikan salah satu mata pelajaran muatan lokal pada tingkat sekolah dasar serta peran pemerintah baik di pusat maupun daerah. Namun belum memadai untuk menjadi solusi dari permasalahan. Ada beberapa faktor yang menjadi penyebab. Faktor kedwibahasaan atau kemultibahasaan yang berkembang di masyarakat, perbedaan kelompok umur, pendidikan, wilayah permukiman, jenis kelamin. tidak menutup kemungkinan penyebab pergeseran bahasa tidak hanya dipengaruhi oleh faktor diatas, akan tetapi juga dapat disebabkan oleh adanya perbedaan etnis atau adat istiadat penutur bahasa dari mana mereka berasal [2].

Berdasarkan faktor-faktor yang telah dijabarkan, untuk menghindari terjadinya kepunahan bahasa daerah, salah satu caranya adalah dengan mesin penerjemah statistik. Mesin penerjemah statistik (*Statistical Machine Translation*) merupakan sebuah pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel [3]. Salah satu model yang digunakan untuk menghasilkan hasil terjemahan adalah *tagging* kata per kata dengan mengambil kata dasar dan imbuhan.

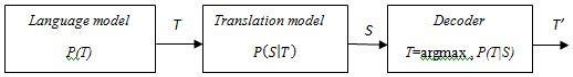
ISSN 2460-0741

Berdasarkan permasalahan diatas, maka pada penelitian ini dilakukan *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya untuk menguji akurasi mesin penerjemah statistik bahasa Indonesia – bahasa Dayak Taman.

II. LANDASAN TEORI

A. Mesin Penerjemah Statistik

Mesin penerjemah statistik merupakan salah satu jenis mesin penerjemah dengan menggunakan pendekatan statistik. Terdapat 3 komponen yang terlibat dalam proses penerjemahan dari satu bahasa ke bahasa lain pada mesin penerjemah statistik, yaitu *language model*, *translation model*, dan *decoder*, seperti terlihat pada Gambar 1.



Gambar 1. Komponen mesin penerjemah *statistik [4]*

Sumber data utama yang dipergunakan adalah parallel corpus dan *monolingual corpus*. Proses *training* terhadap *parallel corpus* menggunakan GIZA++ menghasilkan translation model (TM). Proses *training* terhadap bahasa target pada *parallel corpus* ditambah dengan *monolingual corpus* bahasa target menggunakan SRILM menghasilkan *language model* (LM), sedangkan *PoS model* (PoS-M) dihasilkan dari bahasa target pada *parallel corpus* 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[5].

B. Moses

Moses adalah salah satu Mesin Penerjemah Sattistik yang memungkinkan untuk menerjemahkan secara otomatis setiap pasangan bahasa. Moses digunakan untuk melatih model statistik teks terjemahan dari bahasa sumber ke bahasa sasaran. Saat melakukan penerjemahkan bahasa, Moses membutuhkan korpus dalam dua bahasa, bahasa sumber dan bahasa sasaran. 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 EuroMatrix, dengan pendanaan oleh European Commission[6].

C. 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 delapan jenis, yaitu korpus khusus (*specialised corpus*), korpus umum (*general corpus*), korpus komparatif (*comparable corpus*), korpus paralel (*parallel corpus*), korpus pemelajar (*learner corpus*), korpus pedagogis (*pedagogic corpus*), korpus historis atau diakronis (*historical or diachronic corpus*), dan korpus monitor (*monitor corpus*) [8]. Berdasarkan jenis korpus tersebut, untuk penelitian ini penulis akan memfokuskan pada korpus paralel.

D. Automatic Evaluation

Sitem evaluasi otomatis yang populer saat ini adalah BLEU (*Bilingual Evaluation Understudy*). BLEU adalah sebuah algoritma yang berfungsi untuk mengevaluasi kualitas dari sebuah hasil terjemahan yang telah diterjemahkan oleh mesin dari satu bahasa alami ke bahasa lain. BLEU mengukur *modifiedn-gramprecission score* antara hasil terjemahan otomatis dengan tejemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty*.

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. Sangat penting untuk diketahui bahwa semakin banyak terjemahan rujukan per kalimatnya, maka akan semakin tinggi nilainya. Untuk menghasikan 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 sebagai berikut [9]:

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

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

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

Keterangan:

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

E. Stemming

Stemming adalah proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya. Algoritma stemming dikembangkan berdasarkan aturan morfologi Bahasa Indonesia yang mengelompokkan imbuhan menjadi awalan *(prefix)*, sisipan *(infix)*, akhiran *(suffix)* dan gabungan awalan akhiran *(confixes)*. Algoritma ini menggunakan kamus kata dasar dan mendukung *recoding*,

ISSN 2460-0741

yakni penyusunan kembali kata-kata yang mengalami proses *stemming* berlebih [10].

III. PEMBAHASAN

A. Data Penelitian

Data penelitian yang digunakan berupa bahasa sehari-hari yang digunakan dalam berbahasa Indonesia dari *Indonesian ASEAN langguage translation public service* dan dokumen-dokumen bahasa Dayak Taman. Data-data yang diperoleh tersebut selanjutnya diolah menjadi korpus teks paralel bahasa Dayak Taman dan bahasa Indonesia. Adapun jumlahnya yaitu 3110 pasangan kalimat korpus paralel bahasa Indonesia dan bahasa Dayak Taman.

B. Implementasi Mesin Penerjemah Statistik Jawa ke Bahasa Indonesia

1. Implementasi SRILM

Model bahasa digunakan sebagai sumber pengetahuan berbasis teks dengan nilai-nilai probabilistik. Penelitian ini menggunakan n-gram sebagai *language model*. Model bahasa dibangun dengan tools SRILM. Model bahasa akan menghasilkan output dengan format file *.lm. Berikut merupakan tabel model bahasa yang dihasilkan oleh SRILM pada mesin penerjemah statistik bahasa Indonesia - Bahasa Dayak Taman, seperti terlihat pada Gambar 2.

```
\data\
ngram 1=3363
ngram 2=10312
ngram 3=3166
\1-grams:
-3.933055 apiangaan -0.08117735
-3.488691 apoang        -0.1804145
-------------------------------------
\2-grams
-1.253341 talalo ambat  -0.3530326
-1.257469 talalo ambato -0.1568878
-------------------------------------
\3-grams
-0.6992335  deka tujuan dapatkaan
-0.8309988  deka tujukuun manyule
```

Gambar 2. *M*odel bahasa bahasa Dayak Taman sebagai *B*ahasa Target

2. 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 *tools* Giza++. Proses pemodelan translasi oleh Giza++ menghasilkan dokumen *vocabulary corpus*, *word alignment* dan *lexical model table*. Dokumen-dokumen tersebut terdapat dalam folder "train" yang didalamnya terdapat 4 file yaitu "corpus, giza.id-dy, giza.dy-id dan model", seperti terlihat pada Gambar 3.

```
1   UNK   0
2   yang  507
3   to    499
4   jo    456
5   aika  288
6   insa  154
7   tiang 153
8   ingko 149
9   indi  112
10  inju  102
```

Gambar 3. Dokumen *vocabulary corpus* bahasa Dayak Taman

Angka 1 sampai 10 pada dokumen vocabulary corpus merupakan uniq id untuk setiap data token, sedangkan angka disebelah kanan token menunjukkan frekuensi kemunculan. Vocabulary corpus yang dihasilkan mesin penerjemah bahasa Indonesia – bahasa Dayak Taman terdiri dari 3342 token untuk korpus bahasa Dayak Taman dan 1961 token untuk bahasa Indonesia.

```
# Sentence pair (1) source length 7 target length 6
alignment score : 0.00662217

apakah kita harus membuat reservasi ?

NULL ({ }) aika ({ 1 }) ingki ({ 2 }) ' ({ }) harus ({ 3 })
maniang ({ 4 }) reservasi ({ 5 }) ? ({ 6 })
```

Gambar 4. Dokument alignment bahasa Indonesia - bahasa Dayak Taman

Pada Gambar 4 dokumen alignment Bahasa Indonesia–bahasa Dayak Taman terdapat tiga baris kalimat. Baris pertama berisi letak kalimat target (1) dalam korpus, panjang kalimat sumber (7), panjang kalimat target (7) dan skor *alignment*. Baris kedua merupakan bahasa sumber dan baris ketiga merupakan alignment kalimat bahasa target terhadap kalimat bahasa sumber. Kata "aika" ({ 1 }) memiliki makna bahwa kata "aika" pada kalimat bahasa target, di-align ke kata kedua pada kalimat bahasa sumber yaitu "ingki".

C. Pengujian Hasil Terjemahan Mesin Translasi

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 *tagging* kata per kata dengan mengambil kata dasar dan imbuhan.

Langkah pada pengujian otomatis, korpus yang akan diuji terlebih dahulu melalui langkah translasi otomatis yang akan memberikan output berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin. Korpus uji yang digunakan pada tahap ini berjumlah 3110.

ISSN 2460-0741

Setelah membuat output berupa hasil translasi 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.

```
yosep@yosep-K43SD:~$
~/NLP/mosesdecoder/scripts/generic/multi-bleu.perl
~/NLP/st/id-dy.lowercased.dy < ~/NLP/st/outputcoba.dy
BLEU = 80.17, 89.4/82.8/77.1/72.3 (BP=1.000,
ratio=1.000, hyp_len=21672, ref_len=21662)
yosep@yosep-K43SD:~$
```

Gambar 5. Tampilan nilai dari outputcoba.dy

Berdasarkan Gambar 5 diperoleh nilai awal dari outputcoba.dy sebesar 80.17%.

D. Stemming Bahasa Dayak Taman

1. Stemming Imbuhan

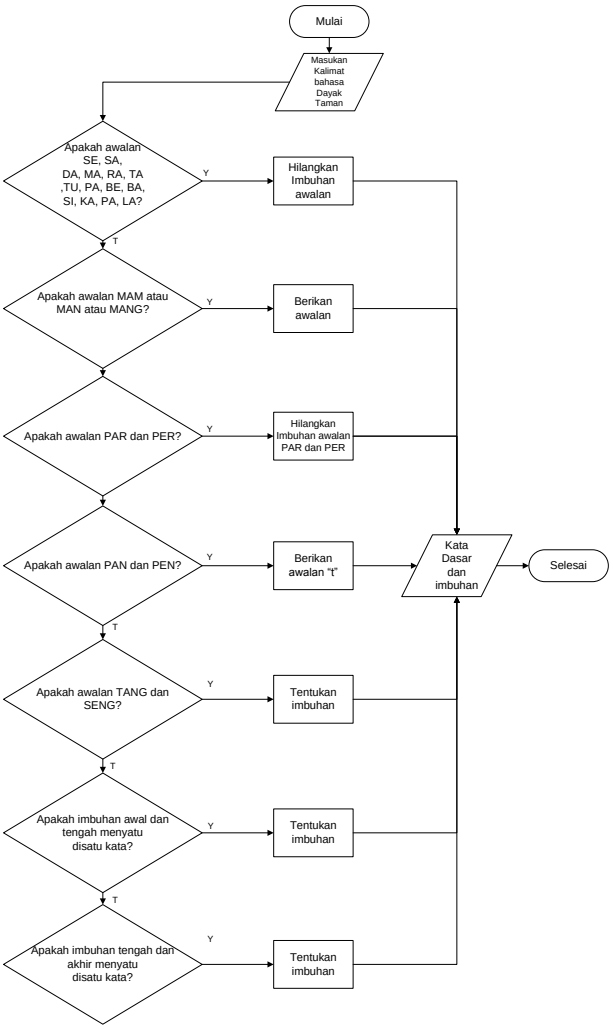
Imbuhan dalam bahasa Dayak Taman secara keseluruhan hampir sama dengan imbuhan bahasa indonesia hanya saja ada beberapa imbuhan yang berbeda. Adapun imbuhan dalam bahasa Dayak Taman yaitu a, i, u, e, o, m, n, k, s se, sa, da, ma, ra, ta, tu, pa, be, ba, si, ka, pa, la, un, on, ng, an, pi, de, du, to, mo, ngi, ang, aka, kan, min, mam, man, pam, per, kir, ter, tang, mang, dan seng. *Flowchart stemming* imbuhan dapat dilihat pada Gambar 6.

2. Stemming Menunjukan Orang

Dalam bahasa Dayak Taman kata-kata yang menunjukan orang biasanya langsung berada pada satu kata tersebut, kata-kata yang menunjukan orang letaknya bisa diawal, tengah, dan akhir kata. Dalam poin ini kata dasar dapat kita temukan dengan menghilangkan kata yang menunjukan orang tetapi kata yang menunjukan orang bisa juga menjadi kata dasar, karena kata yang menunjukan orang bukan termasuk imbuhan jika kata yang menunjukan orang bukan merupakan kata dasarnya. Adapun kata-kata yang menunjukan orang didalam bahasa Dayak Taman yaitu ko, ku, iyak, ak, pam, nam, mau, nau, ngau, ngam, nu, ngu, araina, ingkam, ingko, ingki, kam, ki, ka, dan kin. *Flowchart stemming* menunjukan orang dapat dilihat pada Gambar 7.

3. Stemming Menunjukan Kata “ini” atau “itu”

Dalam bahasa Dayak Taman kata-kata yang menunjukan kata “ini” dan kata “itu” biasanya langsung berada pada satu kata tersebut, kata-kata yang menunjukan orang letaknya bisa tengah, dan akhir kata. Dalam poin ini kata dasar dapat kita temukan dengan menghilangkan kata yang menunjukan kata “ini” dan kata “itu” serta memperhatikan kembali morfologi imbuhannya.



Gambar 6. *Flowchart stemming* imbuhan

Flowchart stemming menunjukan kata “ini” ata “itu” dapat dilihat pada Gambar 8.

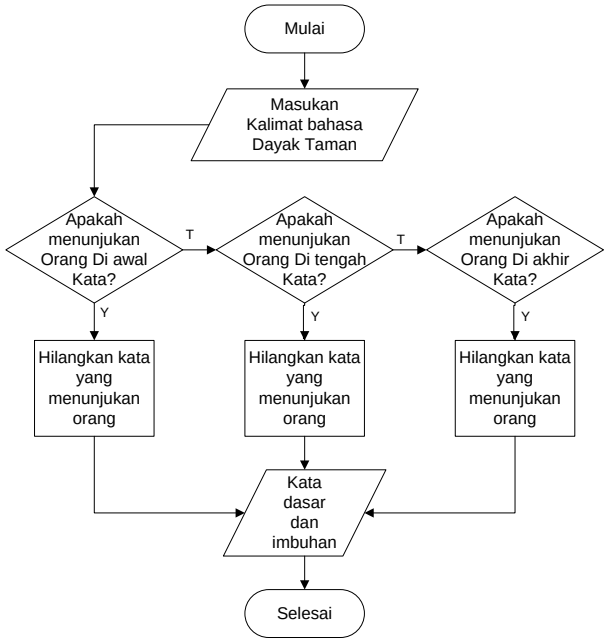
E. Pengujian Ulang Hasil Terjemahan Mesin Translasi

berikutnya adalah menguji kembali hasil terjemahan mesin translasi bahasa Jawa-Bahasa Indonesia yang telah melewati perbaikan *lexical model*. Langkah pengujian dilakukan dengan cara melakukan pengujian otomatis yang akan memberikan output berupa korpus dalam bahasa target yang telah diterjemahkan oleh mesin dan pengujian oleh ahli bahasa.

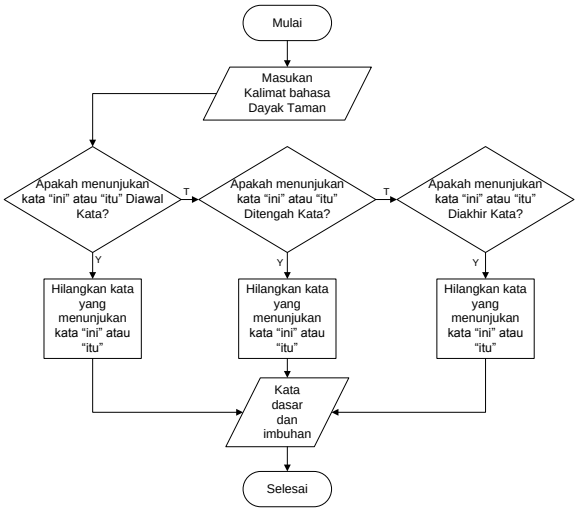
1. Pengujian Otomatis

Pengujian dilakukan dengan cara membandingkan nilai BLEU hasil terjemahan otomatis dari mesin penerjemah Bahasa Indonesia-bahasa Dayak Taman sebelum dan setelah melewati tahap *tagging* kata per kata dengan mengambil kata dasar dan imbuhan.

ISSN 2460-0741



Gambar 7. *Flowchart stemming* menunjukan orang



Gambar 8. *Flowchart stemming* menunjukan kata "ini" atau "itu"

```
yosep@yosep-K43SD:~$
~/NLP/mosesdecoder/scripts/generic/multi-bleu.perl
~/NLP/hasil/id-dy.lowercased.dy <
~/NLP/hasil/output.dy
BLEU = 80.46, 83.5/81.7/79.9/76.8 (BP=1.000,
ratio=1.185, hyp_len=101599, ref_len=85733)
yosep@yosep-K43SD:~$
```

Gambar 9. Tampilan nilai BLEU setelah *tagging*

Berdasarkan Gambar 9 diketahui sebelum *tagging* kata per kata dengan mengambil kata dasar dan imbuhan, nilai BLEU pada korpus uji 3110 sebesar 80.17% dan setelah *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya nilai BLEU sebesar 80.46%. Terdapat peningkatan nilai BLEU sebesar 0.36% dari perbandingan sebelum dan setelah *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya.

2. Pengujian Ahli Bahasa

Pengujian ahli bahasa dilakukan terhadap hasil terjemahan mesin penerjemah statistik bahasa Indonesia-bahasa Dayak Taman dengan mengambil kalimat yang mengalami perubahan pada hasil terjemahan otomatis yang terdapat pada korpus uji 3110 sebelum dan setelah dilakukan *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya sebanyak 20 kalimat. Ahli bahasa menilai apakah hasil terjemahan lebih baik, sama, atau lebih buruk berdasarkan tingkat akurasi terjemahan kata. Perhitungan akurasi dilakukan dengan Persamaan berikut :

$$P = \frac{C}{R} 100\% \quad (4)$$

Keterangan:

P = Persentase akurasi

C = Jumlah kata yang diterjemahkan dengan tepat menurut penilaian dari ahli bahasa

R = Jumlah kata hasil terjemahan

Akurasi ahli bahasa dapat dilihat pada Tabel 1.

TABEL 1 TABEL AKURASI AHLI BAHASA

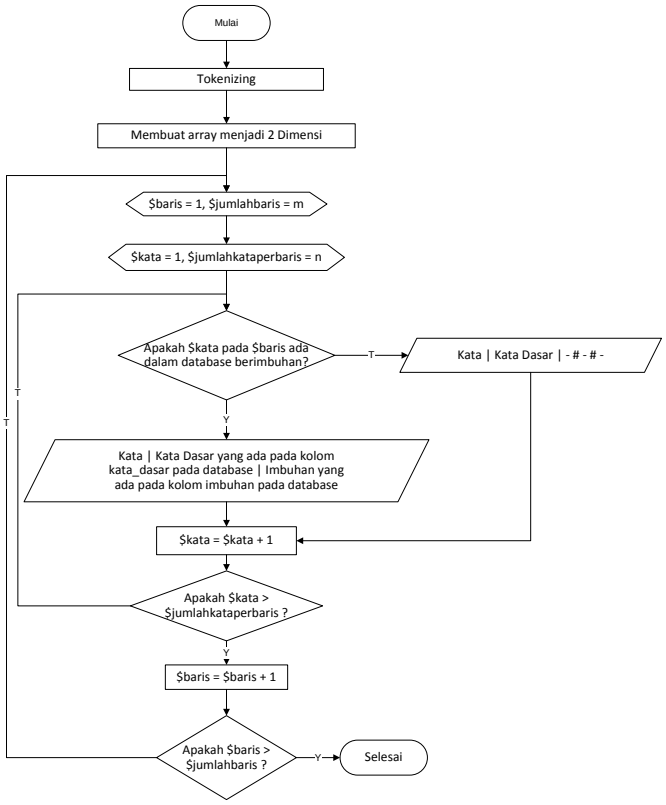
Kalimat Hasil Terjemahan	Ahli Bahasa	C,R	$P = \frac{C}{R} 100\%$
Sebelum *tagging* kata per kata dengan mengambil kata dasar dan imbuhan	H.B. Dailang	C = 107, R = 135	79,26%
Setelah *tagging* kata per kata dengan mengambil kata dasar dan imbuhan	H.B. Dailang	C = 129, R=135	95.56%

F. Proses Tagging Bahasa Dayak Taman

Algoritma *tagging* kata per kata dengan mengambil kata dasar dan imbuhan yang digunakan pada bahasa Dayak Taman dijelaskan pada Gambar 10. Sistem di mulai dengan mempersiapkan korpus bahasa Dayak Taman yang melalui proses *tokenizing* yaitu akan dilakukan proses pemecahan kalimat yang akan menghasilkan korpus tanpa karakter selain huruf dan angka, kemudian hasil dari *tokenizing* tersebut dibuat menjadi *array* 2 dimensi, *array* dimensi pertama digunakan untuk jumlah *array* baris, sedangkan *array* dimensi kedua adalah *array* kata per kata pada baris korpus. Selanjutnya sistem akan menyiapkan jumlah baris pada korpus, pada tahap ini akan dilakukan persiapan untuk setiap jumlah korpus dan akan dihitung per baris korpus. Dan akhirnya sistem akan

ISSN 2460-0741

menyiapkan jumlah kata pada jumlah kata per baris yang kata per kata dari setiap baris korpus akan dilakukan proses *tagging* kata per kata dengan mengambil kata dasar dan imbuhan yang sudah disiapkan dari *database*



Gambar 10. Algoritma *tagging* bahasa Dayak Taman

IV. KESIMPULAN

A. *Kesimpulan*

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

1. Mesin penerjemah statistik dapat diimplementasikan untuk menerjemahkan bahasa Indnesia ke bahasa Dayak Taman.
2. Berdasarkan hasil penelitian, proses *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya dapat meningkatkan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia – bahasa Dayak Taman.
3. Persentase peningkatan nilai akurasi terjemahan mesin penerjemah bahasa Indonesia – bahasa Dayak Taman yang dicapai sebesar 0.36% pada pengujian otomatis oleh BLEU.
4. Penilaian yang dilakukan oleh ahli bahasa menghasilkan peningkatan akurasi hasil terjemahan sebesar 20.57%.

B. saran

Beberapa saran yang dapat diberikan sebagai pengembangan dari penelitian ini adalah sebagai berikut.

1. Perlu penambahan jumlah korpus untuk meningkatkan kualitas terjemahan mesin penerjemah statistik.
2. Perlu adanya teknik lain dalam mencari kata dasar dan imbuhannya yang akan di *tagging* seperti dapat dibuat *thesaurus* bahasa Dayak Taman untuk menentukan kata dasar dan imbuhannya.
3. Perlu dilakukan penelitian lanjutan untuk melakukan analisis dalam menghasilkan terjemahan bahasa Indonesia ke bahasa Dayak Taman dengan mempertimbangkan hubungan antar frase dalam kalimat.
4. Perlu dilakukan pengujian terhadap korpus yang bukan merupakan bagian dari korpus paralel untuk mengetahui tingkat akurasi.
5. Agar dapat mengimplementasikan mesin penerjemah statistik ke dalam bahasa daerah yang lain dengan melakukan proses *tagging* kata per kata dengan mengambil kata dasar dan imbuhannya.

DAFTAR PUSTAKA

[1] Akuntono, Indra. *Jumlah Ragam Bahasa di Indonesia*. Jakarta : Dikti.2012.

[2] Triyono, Sulis. Pembahasan Hasil Penelitian: Pergeseran Bahasa Daerah Akibat Kontak Bahasa Melalui Pembauran. Yogyakarta, LITERA Vol 5 No 1. 2006.

[3] Hadi, Ibnu. Uji Akurasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Melayu Sambas dan Bahasa Melayu Sambas ke Bahasa Indonesia. Pontianak: JUSTIN Vol 3 No 1. 2014.

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

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

[6] Koehn, Philipp. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic. 2007.

[7] McEnery, T., et al. Corpus-Based Language Studies: An Advanced Resource Book. Oxon: Routledge. 2006.

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

[9] Tanuwijaya, Hansel., Manurung, Hisar Maruli. Penerjemahan Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan Word Reordering dan Phrase Reordering. Jakarta, Jurnal ilmu Komputer dan Informasi Vol 2 No 1. 2009.

[10] Nazief, Bobby dan Mirna Adriani, Confix-Stripping: Approach to Stemming Algorithm for Bahasa Indonesia, Fakulty of Computer Science University of Indonesia. 2007.

ISSN 2460-0741