JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989

DOI: 10.26418/justin.v8i3.39810
Vol. 08, No. 3, Juli 2020

Perbandingan Nilai Akurasi Terhadap Penggunaan *Part of Speech Set* pada Mesin Penerjemah Statistik

Eric Dharmawan[a1], Herry Sujaini[a2], Hafiz Muhardi[a3]

[a]*Program Studi Sarjana Informatika Fakultas Teknik Universitas Tanjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak 78124

[1]ericdharmawan.ed@student.untan.ac.id
[2]herry_sujaini@yahoo.com
[3]hafizm@informatika.untan.ac.id

Abstrak

Part of speech pada mesin penerjemah statistik sebagai faktor tambahan sudah beberapa dilakukan terhadap bahasa daerah di Indonesia. *Part of speech* (PoS) untuk bahasa Indonesia pula sudah banyak dikembangkan oleh beberapa peneliti sebelumnya. Penelitian ini menganalisa pengaruh penggunaan dua *tagset* PoS berbeda terhadap hasil terjemahan mesin penerjemah. *Tagset* PoS yang digunakan adalah milik Wicaksono dan Dinakaramani. Mesin penerjemah dibangun dengan korpus paralel Bahasa Indonesia dan Bahasa Melayu Putussibau yang sudah ditandai dengan *tagset* PoS. Proses pengujian menggunakan 2 cara yaitu pengujian otomatis menggunakan *tools* BLEU dan pengujian manual yang dinilai oleh penutur bahasa terhadap hasil terjemahan mesin penerjemah. Hasil pengujian otomatis dengan skenario kedua menunjukkan penerjemahan dengan menambahkan faktor PoS dapat meningkatkan akurasi hasil terjemahan, namun dapat pula menurunkan hasil terjemahan yang dapat disebabkan oleh kuantitas atau kualitas dari korpus *traning*. Selain itu menunjukkan pula persentase peningkatan akurasi yang signifikan pada korpus *training* 5500 terjadi pada Mesin2 (*tagset35*) dengan peningkatan 14,73%, kemudian Mesin1 (*tagset23*) 11,31%, dan disusul oleh Mesin3 (*notagset*) 8,76%. Hasil pengujian dengan skenario pertama dan uji manual mendapatkan bahwa Mesin1 memiliki akurasi terjemahan lebih baik dibandingkan Mesin2. Dengan uji BLEU Mesin1 memiliki akurasi terjemahan (42,39) dan Mesin2 dengan akurasi terjemahan (41,61). Sedangkan untuk uji manual oleh Sigit Heru nilai akurasi Mesin1 (87,47%) dan Mesin2 (83,29%), kemudian oleh Titin Rahayu nilai akurasi Mesin1 (90,91%) dan Mesin2 (86,57%).

Kata kunci: Mesin penerjemah statistik, *Part of speech*, *Tagset* PoS, Korpus paralel

Comparison of The Accuracy Value Toward Using Part of Speech Sets on Statistical Machine Translation

Abstract

Part of speech on the statistical machine translator as an additional linguistic factor applied several times on regional languages in Indonesia. Part of speech (PoS) for bahasa Indonesia also developed by few previous researchers. This research analyzes the effect of using two PoS tags set on the results of machine translation. The PoS tags set used from Wicaksono and Dinakaramani. The machine translator was built using Indonesian parallel corpus and Putussibau Malay language that was marked with PoS tag set. The testing process uses 2 ways, automatic assessment using BLEU tools and manual testing assessed by language speakers on the translation machine translation results. Automatic test results with the second scenario show translation by adding PoS factors can improve the accuracy of the translation results, but can reduce the results of the translation which can be caused by quantity or quality of the corpus training. In addition, it also showed a significant percentage increase in accuracy in the 5500 training corpus occurred in machine 2 (tagset35) 14.73%, then machine 1 (tagset23) 11.31%, and followed by machine 3 (without tagset) 8.76%. The results of first scenario and manual test show that machine 1 own better translation accuracy than machine 2. With the BLEU test, machine 1 has translation accuracy (42.39) and machine 2 with translation accuracy (41.61). for the manual test by Sigit Heru the accuracy percentage of machine 1 (87.47%) and machine 2 (83.29%), then by Titin Rahayu the accuracy percentage of machine 1 (90.91%) and machine 2 (86.57%).

Keywords: Statistical machine translator, *Part of speech*, *Tagset* PoS, Parallel corpus

I. PENDAHULUAN

Bahasa digunakan seseorang untuk menyampaikan ide, mengenalkan diri, dan menceritakan pengalamannya kepada orang lain. Interaksi antar manusia yang baik terjadi jika diantara keduanya memahami apa yang disampaikan. Seperti yang kita ketahui, di Indonesia terdiri dari beragam suku dan budaya, ini berbanding lurus dengan keberagaman bahasa yang dimiliki bangsa Indonesia. Penelitian terhadap jumlah bahasa daerah di Indonesia mendapatkan hasil yang berbeda-beda, perbedaan ini terjadi karena adanya perbedaan metode penelitian yang dilakukan. Dalam artikel yang dimuat pada situs Kemendikbud menyatakan bahwa Badan Bahasa Kemendikbud telah memetakan dan memverifikasi terdapat 652 bahasa daerah di Indonesia, dimana data yang digunakan berasal dari hasil pemetaan bahasa daerah dari tahun 1991 sampai 2017.

Keberagaman Bahasa daerah yang dimiliki tidak memungkinkan untuk dikuasai seseorang secara keseluruhan. Pada setiap bahasa daerah memiliki kesulitan-kesulitan tersendiri, tidak terkecuali bahasa Melayu Putussibau, selain memiliki dialek yang sulit dipahami, susunan kata dalam kalimatnya pula sulit untuk dipahami. Bahasa Melayu Putussibau memiliki keunikan, terutama pada kata yang menyatakan kepemilikan suatu benda. Susunan katanya berbeda dengan bahasa Indonesia. Seperti kalimat "Sebatang pohon tua yang rindang meneduhiku" dalam bahasa Melayu Putussibaunya adalah "Sepon kayu tua yang daun ya lobat mayong aku", perbedaan kata dalam bahasa Indonesia dan bahasa Melayu Putussibau dapat dilihat pada kata yang digaris bawahi. Dengan begitu untuk membantu dalam memahami makna sebuah kalimat bahasa Melayu Putussibau, dibuatlah mesin penerjemah bahasa Indonesia ke bahasa Melayu Putussibau.

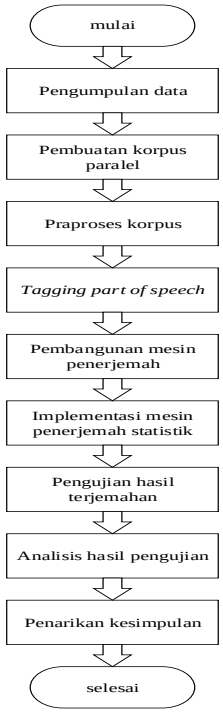
Mesin penerjemah adalah alat penerjemah otomatis pada sebuah teks dari satu bahasa (bahasa sumber) ke bahasa lain (bahasa target) [1]. Mesin penerjemah memiliki keterbatasan dalam menerjemahkan suatu bahasa, akibatnya bahasa yang diterjemahkan belum akurat ada resiko berkurangnya arti dan maksud dari sebuah kalimat. Salah satu cara menghasilkan terjemahan yang optimal adalah dengan menerapkan konsep penerjemahan secara statistik atau yang disebut mesin penerjemah statistik atau *Statistical Machine Translation*. Mesin penerjemah statistik merupakan pendekatan mesin penerjemah dengan hasil terjemahan yang dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus paralel, atau biasa disebut dengan korpus bilingual dari dua bahasa berbeda [2].

Salah satu faktor tambahan yang digunakan untuk meningkatkan hasil terjemahan adalah menggunakan *Part of Speech* (PoS). *Part of Speech* merupakan kategori kata-kata berdasarkan sifat tata bahasa, kata-kata yang memiliki sifat gramatikal serupa seperti sintaks dan morfologi di kelompokkan ke dalam tag PoS yang sama [3]. Penelitian dengan penambahan PoS sudah banyak dilakukan salah satunya yang dilakukkan oleh [4], pada penelitiannya didapatkan bahwa akurasi terjemahan mengalami peningkatan sebesar 0,6% dengan perhitungan BLEU dan meningkat 21,67% berdasarkan penilaian ahli bahasa. Berbagai *tagset Part of Speech* bahasa Indonesia telah dikembangkan oleh peneliti-peneliti selama 11 tahun (sejak tahun 2008) *tagset* bahasa Indonesia pula belum terstandarisasi dengan *tagset* yang bervariasi mulai dari 16 tag sampai 37 tag [5]. Beberapa penelitian yang membangun *tagset* PoS bahasa Indonesia, diantaranya penelitian yang dilakukan oleh [6] dengan 23 *tagset* PoS, oleh [7] dengan 2 tipe *tagset* PoS yang terdiri dari 37 dan 25 *tagset*, kemudian penelitian dengan 35 *tagset* oleh [8], kemudian oleh [9] dengan 19 *tagset* PoS, [10] dengan 23 *tagset* PoS, dan [11] dengan 29 *tagset* PoS.

Berdasarkan penjelasan yang sudah dipaparkan, penelitian ini akan membuat mesin penerjemah dengan tambahan fitur PoS, dimana *tagset* PoS yang digunakan adalah hasil pengembangan yang dilakukan [8] dan [10], pemilihan kedua *tagset* ini karena ada 2 bentuk *tagset* yang spesifik kelas katanya dan bentuk lebih umum kelas katanya, berdasarkan kedua *tagset* tersebut penelitian ini akan menganalisa jenis *tagset* mana memiliki nilai akurasi lebih baik pada mesin penerjemah bahasa Indonesia ke bahasa Melayu Putussibau.

II. METODE PENELITIAN

Penelitian dilakukan dengan beberapa langkah-langkah sistematis, berikut langkah penelitian diperlihatkan pada Gambar. 1.



Gambar. 1 Metode Penelitian

A. Pengumpulan Data

Data yang dikumpulkan digunakan untuk membuat korpus paralel, korpus paralel merupakan dua file dokumen text yang saling berhubungan dimana dokumen text pertama berisikan kumpulan kalimat bahasa sumber dan dokumen kedua berisikan kumpulan kalimat bahasa terjemahan [12]. Korpus yang dikumpulkan merupakan korpus dalam bahasa Indonesia. Data korpus dapat berasal dari buku-buku, teks berita, dan lainnya. Pada penelitian ini

korpus didapat dari novel Andrea Hirata yaitu Laskar Pelangi, artikel cerita, percakapan sehari-hari, dan sumber lainnya. Data yang terkumpul 6.000 baris korpus bahasa Indonesia, kemudian diterjemahkan kedalam bahasa Melayu Putussibau oleh seorang ahli bahasa Melayu untuk menjadi korpus paralel.

B. Pembuatan Korpus Paralel

Korpus merupakan kumpulan kalimat-kalimat yang disusun secara teratur, korpus biasanya digunakan sebagai sumber penelitian pada bidang kebahasaan dan sastra. Penelitian ini menggunakan dua korpus bahasa, yang pertama adalah korpus bahasa Melayu Putussibau dan korpus bahasa Indonesia. Kedua korpus diselaraskan antara korpus bahasa Indonesia dan korpus bahasa Melayu Putussibau dari urutan setiap baris kalimatnya. Penyelarasan ini kemudian membuat kedua korpus menjadi korpus paralel yang akan digunakan pada mesin penerjemah statistik. Sebelum menjadi korpus paralel yang berasal dari sumber data yang sudah dipaparkan sebelumnya, data korpus tersebut diperoleh dalam bahasa Indonesia yang kemudian akan dilakukan penerjemahan ke bahasa Melayu Putussibau oleh ahli bahasa. Setelah itu melakukan penyelarasan dari segi urutan bahasa sumber dan target agar korpus paralel yang dihasilkan.

C. Praproses Korpus

Praproses korpus merupakan proses yang biasa dilakukan pada penelitian-penelitian yang berkenaan dengan *natural language* seperti pada [13] praposes yang dilakukan adalah *sentence splitting*, *remove punctuation*, dan *tokenisasi*. Pada penelitian [14] praproses yang dilakukan dengan *tokenisasi*, *case folding*, dan *filtering*. Sama seperti pada penelitian ini pra prose yang dilakukan adalah *cleaning*, tokenisasi, dan *lowercasing*. Proses *cleaning* merupakan proses menghilangkan spasi berlebih dan menghilangkan tanda titik di akhir kalimat, ini sama seperti pada proses *remove punctuation*. Kemudian proses selanjutnya adalah mentokenisasi korpus, dimana korpus akan ditokenisasi setiap katanya ataupun tanda baca. Selanjutnya adalah proses *lowercasing* atau *case folding*, proses ini mengubah huruf kapital menjadi huruf kecil secara keseluruhan.

D. Tagging Part of Speech

Tagging part of speech adalah proses menandai kata perkata pada seluruh kalimat yang ada pada korpus bahasa target yaitu bahasa Melayu Putussibau. Proses pemberian tanda dilakukan dengan metode *word based*. Pada dasarnya pemberian tanda dilakukan secara manual, namun dengan banyaknya jumlah korpus yang akan ditandai, penelitian ini dibantu dengan penggunaan I-postagger dimana 10% dari jumlah korpus bahasa Melayu Putussibau akan digunakan sebagai korpus *training* pada I-postagger. Jumlah korpus yang digunakan pada penelitian ini sebanyak 6.000 kalimat, jadi akan ada 600 kalimat yang ditandai secara manual, kemudian 600 kalimat yang sudah ditandai akan digunakan sebagai korpus *training* pada I-postagger. Selanjutnya akan diambil 100 kalimat untuk dilakukan penandaan otomatis menggunakan I-postagger, kemudian hasil penandaan akan dikoreksi kembali, lalu hasil koreksian akan dimasukkan ke korpus *training* pada I-postagger untuk di *training* lagi, begitu selanjutnya hingga korpus tertandai seluruhnya.

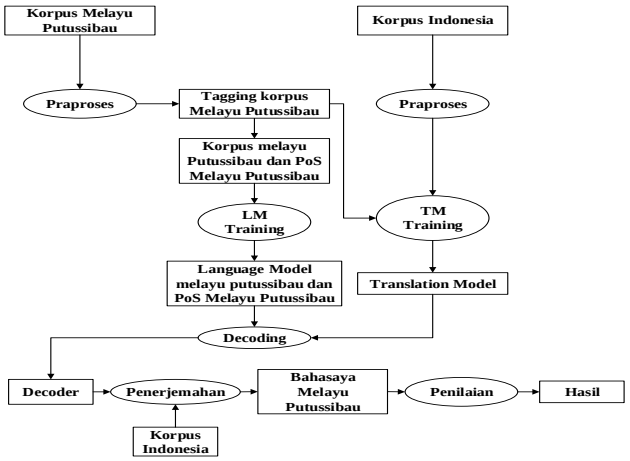
E. Pembangunan Mesin Penerjemah

Pembangunan mesin penerjemah statistik pada penelitian ini dilakukan pada komputer dengan system operasi Ubuntu. Kemudian, proses pembangunan mesin penerjemah perlu untuk melakukan penginstalan beberapa perangkat lunak yang diperlukan. Adapun perangkat lunak yang digunakan dalam mesin penerjemah statistik adalah sebagai berikut:

- KenLm untuk pemodelan bahasa,
- GIZA++ untuk pemodelan translasi,
- Moses SMT untuk *decoding* dan
- BLEU untuk pengujian otomatis.

F. Implementasi Mesin Penerjemah Statistik

Pengimplementasian mesin penerjemah pada penelitian ini berdasarkan rancangan arsitektur mesin penerjemah statistik yang terdiri dari beberapa proses yaitu, tahap persiapan data, *tagging* PoS pada korpus, pemodelan bahasa, pemodelan *translasi*, *decoding* dan pengujian hasil terjemahan. Adapun implementasi sistem mesin penerjemah statistik ini berdasarkan arsitektur yang diperlihatkan pada Gambar. 2.



Gambar. 2 Arsitektur mesin penerjemah statistik indonesia-melayu putussibau

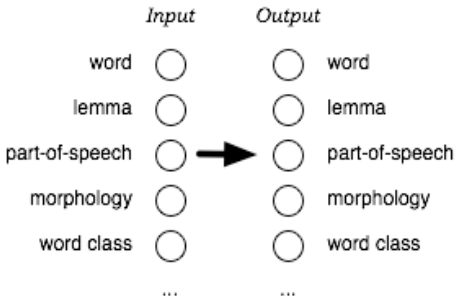
Contoh dari korpus yang sudah di-*tagging* PoS diperlihatkan pada Gambar. 3.

```
gerak|gerak|nn sidak|sidak|prp tuk|tuk|pr nyuruh|nyuruh|md
tesaban|tesaban|vb

tamak|tamak|vb dengan|dengan|in ingan|ingan|nn gendang|gendang|nn
dan|dan|cc tarian|tarian|nn yang|yang|sc saja|saja|rb dinamis|dinamis|jj ,|,|z
penonton|penonton|nn pun|pun|rp tesaban|tesaban|vb

penonton|penonton|nn tesaban|tesaban|vb nerimak|nerimak|vb
sajian|sajian|nn musik|musik|nn etnik|etnik|nn ngehentak|ngehentak|jj
yang|yang|sc nak|nak|neg diduga-duga|diduga-duga|vb
```

Gambar. 3 Contoh korpus *tagging* pos

Setelah melakukan praproses dan *tagging* PoS korpus, kemudian akan dilakukan penyesuain format pelatihan

untuk penerjemahan berbasis faktor. Dimana pada bagian sisi korpus bukan hanya berisikan informasi kata saja, namun berisikan informasi terkait *lemma*, PoS, dan morfologi. Model penerjemahan faktor diperlihatkan pada Gambar. 4 [15]. Pemanfaatan model terjemahan faktor pada penelitian-penelitian mesin penerjemah sudah banyak dikembangkan diantaranya pada penelitian [16]–[19].



Gambar. 4 Model penerjemahan faktor [15]

Dokumen yang sudah di-*tagging* kemudian akan dilakukan pemisahan ke dalam 2 file. File pertama berisikan korpus Bahasa Melayu Putussibau saja dan file kedua berisikan tag PoS saja. Selanjutnya akan dilakukan proses pemodelan bahasa menggunakan *tools* KenLM. Hasilnya adalah model bahasa Melayu Putussibau dan Pos model, contoh hasil pemodelan bahasa diperlihatkan pada Gambar. 5.

```
\data\
ngram 1=7998
ngram 2=38665
ngram 3=53716
ngram 4=53508
ngram 5=49193

\1-grams:
-3.0897884          selepas                 -0.12743756
----------------------------------------------------------
\2-grams:
-3.0122144          , selepas               -0.03021513
----------------------------------------------------------
\3-grams:
-2.1538653          sekulah , selepas       -0.008024926
----------------------------------------------------------
\4-grams:
-1.8863553          ke sekulah , selepas    -0.0037472197
----------------------------------------------------------
\5-grams:
-1.6681552          seru ke sekulah , selepas
```

Gambar. 5 Contoh hasil model bahasa

Pada proses pemodelan *translasi* dokumen korpus yang digunakan adalah korpus paralel bahasa Indonesia dan bahasa Melayu Putussibau yang sudah di-*tagging*. Selanjutnya yang akan menghasilkan model *translasi*. Contoh hasil tabel model *translasi* diperlihatkan pada Gambar. 6.

```
! ini kesempatan ||| !|z tuk|pr kesempatan|nn ||| 1 0.412789 1 0.82911 ||| 0-0 1-1 2-2 ||| 1 1 1 ||| |||
! ini ||| !|z tuk|pr ||| 1 0.412789 1 0.82911 ||| 0-0 1-1 ||| 1 1 1 ||| |||
! puji bu mus ||| !|z puji|nn bu|nnp mus|nnp ||| 1 0.524752 1 0.823968 ||| 0-0 1-1 2-2 3-3 ||| 1 1 1 ||| |||
! puji bu ||| !|z puji|nn bu|nnp ||| 1 0.524752 1 0.83414 ||| 0-0 1-1 2-2 ||| 1 1 1 ||| |||
! puji ||| !|z puji|nn ||| 1 0.524752 1 0.898305 ||| 0-0 1-1 ||| 1 1 1 ||| |||
! ||| !|z ||| 0.981132 0.524752 0.981132 0.898305 ||| 0-0 ||| 53 53 52 ||| |||
! ||| ya|prp !|z ||| 1 0.524752 0.0188679 0.577741 ||| 0-1 ||| 1 53 1 ||| |||
```

Gambar. 6 Contoh hasil model *translasi*

Kemudian, model bahasa Melayu Putussibau, PoS model, dan model translasi, bersama-sama dilakukan proses *decoding* yang akan menghasilkan *decoder*. Setelah *decoder* terbentuk akan melakukan penerjemahan dari bahasa Indonesia ke bahasa Melayu Putussibau, yang kemudian akan dilakukan pengujian terhadap hasil terjemahan.

G. Pengujian Hasil Terjemahan

Pengujian hasil terjemahan dilakukan dengan dua cara yaitu otomatis dan manual. Pengujian otomatis dilakukan dengan menggunakan BLEU, serta menggunakan metode K-*fold Cross Validation* dalam penghimpunan nilai akurasi secara keseluruhan. Penggunaan metode K-*fold Cross Validation* bertujuan agar nilai yang dihasilkan lebih akurat. Pengujian secara manual adalah metode pengujian yang memiliki tingkat akurasi paling baik, namun akan memakan waktu yang lama karena dilakukan secara manual oleh ahli bahasa.

1) Pengujian Otomatis: Pengujian dengan pengujian otomatis menggunakan metode K-*fold cross validation* dimana dilakukan sebanyak 12 kali iterasi/perulangan dengan korpus *training* dan korpus uji berbeda-beda yaitu, sebanyak 5500 *training* dan 500 korpus uji. Korpus uji dan korpus *training* total 6000 baris kalimat, dimana dari 6000 dipecah ke-12 bagian, masing-masing bagian berisi 500 kalimat. Kemudian pada pengujian otomatis pula dilakukan skenario pengujian dengan membuat 3-*fold* korpus *testing*, dimana pada 1-*fold* korpus *testing* itu akan diujikan terhadap mesin penerjemah dengan korpus *training* yang ditambahkan secara berkala dari 500 baris kalimat hingga 5500 kalimat sebagai korpus *traning* mesin penerjemah. Pengujian ini dilakukan terhadap Mesin1 (*tagset23*), Mesin2 (*tagset35*), dan Mesin3 (*notagset*).

2) *Pengujian Manual:* Pengujian manual dilakukan oleh ahli bahasa Melayu Putussibau. Dalam pelaksanaan pengujian ini akan digunakan 100 kalimat uji yang akan di ambil dari korpus paralel. Pemilihan 100 kalimat akan dilakukan dengan tetap berdasarkan pembagian *fold* pada metode K-*fold Cross Validation* sebelumnya, dimana akan dicari nilai akurasi penerjemahan terbaik dari setiap *fold* terhadap nilai akurasi mesin penerjemah statistik (MPS) dengan *tagset* PoS 23 (Mesin1) dan *tagset* PoS 35 (Mesin2). Kemudian akan didapat 2 *fold* dengan nilai akurasi terbaik dari masing-masing Mesin1 dan Mesin2. Masing-masing *fold* berisi 500 baris kalimat, yang kemudian akan dilakukan K-*fold* kembali terhadap *fold* yang didapat sebelumnya untuk mendapatkan 100 baris kalimat uji terbaik yang akan digunakan sebagai korpus uji oleh ahli bahasa. Pengujian manual dihitung dengan persamaan 1 [20].

$$P = \frac{C}{R} 100\% \quad (1)$$

Keterangan:
P = persentase peningkatan akurasi
C = jumlah kata yang tepat diterjemahkan menurut penilaian ahli bahasa
R = jumlah kata hasil terjemahan

H. Analisis Hasil Pengujian

Analisis hasil dan pengujian dilakukan untuk mengetahui pengaruh *tagset Part of Speech* Dinakaramani dan Wicaksono terhadap hasil terjemahan pada mesin penerjemah Statistik bahasa Indonesia ke bahasa Melayu Putussibau.

Selain itu proses analisis pula melibatkan beberapa penelitian-penelitian sebelumnya sebagai referensi untuk melakukan analisis hasil penelitian.

Perhitungan nialai rata-rata, penggunaan grafik, dan penggunaan persamaan matematis juga digunakan sebagai pendekatan dalam melakukan analisis.

I. Penarikan Kesimpulan

Penarikan kesimpulan dirumuskan berdasarkan analisis hasil pengujian, penarikan kesimpulan mengacu pada tujuan dari penelitian yang dilakukan. Dari penelitian yang telah dilakukan didapatkan sebuah pengetahuan baru yang dapat dijadikan sebagai bahan penelitian selanjutnya.

III. HASIL DAN PEMBAHASAN

Pengujian dilakukan dengan 2 cara yaitu pengujian otomatis menggunakan *tools* BLEU dan pengujian yang dilakukan oleh ahli Bahasa Melayu Putussibau.

Pengujian otomatis dilakukan dengan dua skenario, yang pertama ada dengan uji k-*fold cross validation* dimana terdapat 12-*fold* pengujian. Hasil pengujian k-*fold* diperlihatkan pada Tabel I.

TABEL I
NILAI AKURASI 12-FOLD

Iterasi Fold	**Mesin1 (*Tagset23*)**	**Mesin2 (*Tagset35*)**
1	40,96	40,74
2	41,74	41,3
3	38,09	37,83
4	38,28	38,17
5	40,26	39,24
6	41,26	39,82
7	33,22	32,57
8	40,79	40,9
9	48,9	47,62
10	49,48	47,29
11	47,57	46,53
12	48,11	47,28
Rata-rata	42,39	41,61

Kemudian pada pengujian otomatis pula dilakukan skenario pengujian dengan membuat 3-*fold* korpus *testing*, dimana pada 1-*fold* korpus *testing* itu akan diujikan terhadap mesin penerjemah dengan korpus *training* yang ditambahkan secara berkala dari 500 baris kalimat hingga 5500 kalimat sebagai korpus *traning* mesin penerjemah. Pengujian ini dilakukan terhadap Mesin1 (*tagset23*), Mesin2 (*tagset35*), dan Mesin3 (*notagset*). Hasil pengujian dengan penambahan jumlah korpus *training* diperlihatkan pada Tabel II.

TABEL III
NILAI AKURASI RATA-RATA

Jumlah Korpus	**Mesin1 (*Tagset23*)**	**Mesin2 (*Tagset35*)**	**Mesin3 (*Notagset*)**
500	36,17	34,83	43,83
1000	36,2	35,06	45,21
1500	36,37	36,99	45,81
2000	37,79	37,53	46,33
2500	38,96	39,23	46,59
3000	38,79	38,48	46,4
3500	39,13	38,69	46,62
4000	38,9	38,75	46,74
4500	38,99	38,66	46,9
5000	39,72	39,89	47,37
5500	40,26	39,96	47,67
Rata-rata	38,30	38,01	46,32

Berdasarkan nilai-nilai pada Tabel II dibuat perhitungan persentase peningkatan yang dihitung dengan persamaan 2.

$$\%\ kenaikan = \frac{a - b}{b} \cdot 100\% \qquad (2)$$

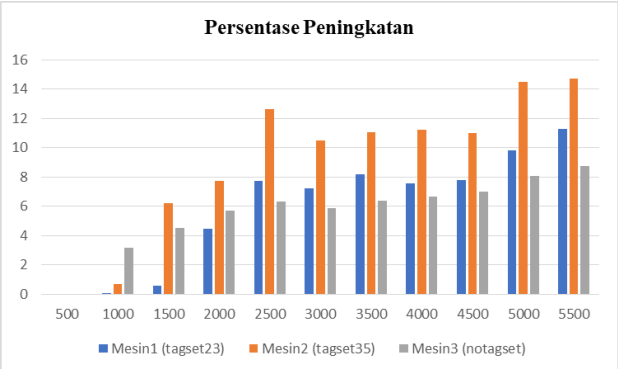
Keterangan :
a = nilai akhir
b = nilai awal

Persentase peningkatan akurasi pada tiap penambahan korpus *training* dari 500 sampai 5500 diperlihatkan pada Tabel III.

TABEL IIIII
PERSENTASE PENINGKATAN AKURASI

Jumlah Korpus	**Mesin1 (*tagset23*)**	**Mesin2 (*tagset35*)**	**Mesin3 (*notagset*)**
500	0%	0%	0
1000	0,08%	0,66%	3,15%
1500	0,55%	6,20%	4,52%
2000	4,48%	7,75%	5,70%
2500	7,71%	12,63%	6,30%
3000	7,24%	10,48%	5,86%
3500	8,18%	11,08%	6,37%
4000	7,55%	11,25%	6,64%
4500	7,80%	11,00%	7,00%
5000	9,81%	14,53%	8,08%
5500	11,31%	14,73%	8,76%

Nilai-nilai hasil persentase peningkatan pada Tabel III dapat ditampilkan dalam bentuk grafik untuk menganalisa peningkatannya. Grafik peningkatan persentase diperlihatkan pada Gambar. 7.



Gambar. 7 Grafik persentase peningkatan

Dari Tabel I, dapat kita lihat bahwa pengujian k-*fold* terhadap Mesin1 dan Mesin2 dengan BLEU memperlihatkan hasil akurasi terjemahan Mesin1 dengan nilai rata-rata (42,39) dan Mesin2 dengan nilai rata-rata (41,61).

Kemudian Pada Tabel II, memperlihatkan Mesin3 yaitu mesin penerjemah statistik (MPS) tanpa tambahan *tagset* PoS memiliki nilai akurasi rata-rata sebesar (46,32), lebih baik akurasi terjemahnnya Jika dibandingkan dengan Mesin1 dan Mesin2. Secara teori dengan penambahan *tagset* PoS pada sisi korpus target dapat meningkatkan akurasi terjemahan lebih baik dibandingkan dengan MPS tanpa tambahan faktor *tagset* PoS. Seperti halnya pada penelitian [4] MPS Indonesia-Melayu Pontianak terjadi peningkatan 0,6% oleh uji BLEU, Kemudian pada penelitian [21] terjadi peningkatan 2% oleh uji BLEU pada Bahasa Inggris-Indonesia, dan penelitian oleh [15] Bahasa German-Inggris mengalami peningkatann 0,86 oleh uji BLEU. Namun, penurunan akurasi ini dapat terjadi seperti pendapat yang disampaikan oleh [22] dalam bukunya yang berjudul "*Statistical Machine Translation*" beliau menyatakan bahwa, dasar dari anotasi sintaksis adalah memberikan tanda/label pada setiap kata dalam suatu kalimat dengan *part of speech*. Tanda untuk anotasi sintaksis tidak muncul dengan sendirinya, namun harus ditambahkan sehingga diperlukannya *tools* otomatis untuk membantu dalam memberikan tanda. Penggunaan *tools* otomatis dilakukan karena memberikan tanda pada setiap kata secara manual akan sangat memakan biaya yang besar dan waktu yang lama. Penggunaan *tools* otomatis sangat membantu, tetapi *tools* otomatis juga memiliki tingkat *error* yang signifikan (kemampuan *parsing* sering tidak lebih baik dari 90%), jadi penambahan faktor dapat berguna secara teori, namun dapat juga tidak berguna dalam peraktek. Karena dengan penambahan faktor akan membuat model semakin komplek dan akan semakin sulit dalam pencarian model penerjemahan, karena opsi terjemahan akan semakin banyak. Penurunan nilai akurasi itu pula dapat disebabkan oleh kuantitas korpus *training* ataupun kualitas dari korpus itu sendiri. Ini dapat kita lihat pada penelitian [21], di dalam penelitiannya terdapat percobaan yang dilakukan terhadap perbedaan jumlah korpus *training*, hasil penelitiannya diperlihatkan pada Tabel IV. Dapat dilihat bahwa pada jumlah kalimat *training* 1000 dan 5000, kualitas terjemahan lebih baik terjadi pada *surface* atau tanpa *tagset* PoS.

TABEL IVV
HASIL SCORING BLEU [17]

Jumlah kalimat	Surface (%)	Surface + PoS(%)
1000	31,63	31,61
2000	31,3	31,7
5000	31,86	31,56
10000	31,51	31,84
12000	32,21	32,71

Dari Tabel III memperlihatkan peningkatan persentase nilai akurasi seiring dengan bertambahnya korpus *training*. Dimana peningkatan tertinggi terjadi pada mesin penerjemah dengan faktor tambahan *tagset* PoS pada korpus uji 5500, Mesin1 memiliki persentase peningkatan sebesar 11,31%, pada Mesin2 persentase peningkatannya sebesar 14,73%, dan Mesin3 dengan peningkatan 8,67%. Bisa dilihat pula pada Gambar. 7, persentase peningkatan terus mengalami kenaikan signifikan untuk kedua *tagset* PoS. Pada jumlah korpus *traning* tertentu mesin penerjemah dengan tambahan PoS akan memiliki nilai akurasi terjemahan lebih baik dibandingkan dengan mesin penerjemah tanpa faktor tambahan, seperti pada penelitian [21] dimana pada jumlah korpus *traning* 10.000 keatas akurasi terjemahan mesin dengan faktor tambahan konsisten lebih tinggi dari mesin penerjemah tanpa tambahan faktor.

Pengujian manual dilakukan oleh ahli bahasa Melayu Putussibau. Dalam pelaksanaan pengujian ini akan digunakan 100 kalimat uji yang akan di ambil dari korpus paralel. Pemilihan 100 kalimat akan dilakukan dengan tetap berdasarkan pembagian *fold* pada metode K-*fold Cross Validation* sebelumnya, dimana akan dicari nilai akurasi penerjemahan terbaik dari setiap *fold* terhadap nilai akurasi mesin penerjemah statistik (MPS) dengan *tagset* PoS 23 (Mesin1) dan *tagset* PS 35 (Mesin2). Kemudian akan didapat 2 *fold* dengan nilai akurasi terbaik dari masing-masing Mesin1 dan Mesin2. Masing-masing *fold* berisi 500 baris kalimat, yang kemudian akan dilakukan K-*fold* kembali terhadap *fold* yang didapat sebelumnya untuk mendapatkan 100 baris kalimat uji terbaik yang akan digunakan sebagai korpus uji oleh ahli bahasa, hasil uji diperlihatkan pada Tabel V.

TABEL V
HASIL PENILAIAN MANUAL OLEH PENUTUR BAHASA

No.	Penutur Bahasa	C		R		P= C/R * 100%	
		tag 23	*tag 35*	*tag 23*	*tag 35*	*Mesin1 (tag23)*	*Mesin2 (tag35)*
1	Sigit Heru P.	649	618	742	742	87,47%	83,29%
2	Titin Rahayu	670	638	737	737	90,91%	86,57%

Penilaian manual menurut penutur bahasa Melayu Putussibau Mesin1 lebih tinggi dibandingkan dengan Mesin2. Dimana menurut Sigit Heru nilai akurasi terjemahan Mesin1 adalah 87,47% dan Mesin2 adalah 83,29%. Sedangkan menurut Titin Rahayu terjemahan dengan Mesin1 adalah 90,91% dan Mesin2 adalah 86,57%. Perbedaan kualitas hasil terjemahan antar Mesin1 dan Mesin2 diperlihatkan pada Tabel VI.

TABEL VV
PERBANDINGAN HASIL TERJEMAHAN MESIN PENERJEMAH

Surface	*Tagset 23*	*Tagset 35*
aku mauk seperempat ons parfum chanel no 19	aku|prp mauk|md seperempat|UNK|UNK |UNK ons|nnd parfum|UNK|UNK|UN K chanel|nnp no|UNK|UNK|UNK 19|cd	aku|prp mauk|md ons|nn seperempat|UNK|UNK| UNK parfum|UNK|UNK|UN K chanel|nnp no|UNK|UNK|UNK 19|cdp
kami mauk duduk dampen jalan	kami|prp mauk|rb duduk|vb di|in dampen|jj jalan|nn	kami|prp mauk|rb duduk|vbi di|in jalan|nn dampen|jj
talah nak nuan ngabar aku langsung selepas nuan nyemaet ya ?	talah|wh nak|neg nuan|prp madah|vb aku|prp copat|jj antik|sc nuan|prp nyemait|vb yak|pr ?|z	talah|wp nak|neg nuan|prp madah|vbt aku|prp copat|jj antik|sc nuan|prp nyemait|vbt yak|dt ?|.
ada nak nuan dua agik yang macam tuk ?	ada|vb nak|neg nuan|prp dua|cd agik|rb yang|sc macam|in tuk|pr ?|z	apa|wp nuan|prp agik|rb bisik|vbt dua|cdp yang|sc semacam|nn tuk|dt ?|.

Surface	*Tagset 23*	*Tagset 35*
talah nak aku mesan makan malam di kapal ?	talah|wh nak|neg aku|prp mesan|vb makan|vb malam|nn di|in kapal|nn ?|z	talah|wp nak|neg aku|prp mesan|vbt kapal|nn di|in makan|vbt malam|nn ?|.

Dari Tabel VI perbandingan hasil terjemahan Mesin1 dan Mesin2 dengan kalimat referensi bahasa Melayu Putussibau dari ahli bahasa, bisa dilihat bahwa pada Mesin1 menerjemahkan kalimat uji lebih mirip/sesuai dengan kalimat referensi. Dengan itu kualitas Mesin1 lebih baik dibandingkan dengan Mesin2. Selain itu tedapat tanda UNK (*Unknown*) pada kata-kata di Tabel VI yang artinya kata tersebut tidak diketahui atau pada korpus training tidak ada referensinya terhadap *surface*, *lemma*, dan *part of speech*-nya.

IV. KESIMPULAN

Dari penelitian yang dilakukan dapat disimpulkan bahwa mesin penerjemah statistik bahasa Indonesia-Melayu Putussibau, dengan faktor tambahan *tagset* PoS milik Dinakaramani (23 *tagset*) menghasilkan akurasi terjemahan yang lebih baik dibandingkan dengan mesin penerjemah dengan faktor tambahan *tagset* PoS milik Wicaksono (35 *tagset*). Selain itu, penggunaan faktor tambahan dalam membangun mesin penerjemah statistik dapat meningkatkan nilai akurasi mesin penerjemah, namun dapat pula menyebabkan penurunan kualitas terjemahan diantaranya dapat disebabkan oleh kuantitas korpus *training* ataupun kualitas dari korpus. Kemudian, mesin penerjemah statistik dengan faktor tambahan memiliki persentase peningkatan akurasi yang lebih baik dibandingkan dengan mesin penerjemah statistik tanpa faktor tambahan. Seiring dengan bertambahnya jumlah korpus *training* juga dapan menyebabkan nilai akurasi terjemahan ikut meningakat.

DAFTAR PUSTAKA

[1] A. Setiawan, H. Sujaini, and A. B. Pn, "Implementasi Optical Character Recognition (OCR) pada Mesin Penerjemah Bahasa Indonesia ke Bahasa Inggris," *J. Sist. dan Teknol. Inf.*, vol. 5, no. 2, pp. 135–141, 2017.

[2] H. Sujaini, "Peningkatan Akurasi Penerjemah Bahasa Daerah dengan Optimasi Korpus Paralel," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 1, 2018, doi: 10.22146/jnteti.v7i1.394.

[3] A. F. Abka, "Evaluating the use of word embeddings for part-of-speech tagging in Bahasa Indonesia," *Proceeding - 2016 Int. Conf. Comput. Control. Informatics its Appl. Recent Prog. Comput. Control. Informatics Data Sci. IC3INA 2016*, pp. 209–214, 2017, doi: 10.1109/IC3INA.2016.7863051.

[4] D. Indrayana, H. Sujaini, and N. Safriadi, "Meningkatkan Akurasi Pada Mesin Penerjemah Bahasa Indonesia Ke Bahasa Melayu Pontianak Dengan Part Of Speech," vol. 3, no. 1, pp. 1–5, 2016.

[5] M. Kamayani, "Perkembangan Part-of-Speech Tagger Bahasa Indonesia," *J. Linguist. Komputasional*, vol. 2, no. 2, p. 34, 2019, doi: 10.26418/jlk.v2i2.20.

[6] A. Purwantiari and T. Suhardijanto, "INACL POS Tagging Convention Konvensi Pelabelan Kelas Kata INACL / MALKIN," 2017.

[7] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic Part of Speech Tagging for Bahasa Indonesia," *Proc. 3rd Int. MALINDO Work. Coloca. event ACL-IJCNLP*, 2009.

[8] A. F. Wicaksono, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia HMM Based Part-of-Speech Tagger for Bahasa Indonesia," no. January 2010, 2014.

[9] S. D. Larasati, V. Kuboň, and D. Zeman, "Indonesian morphology tool (MorphInd): Towards an Indonesian corpus," *Commun. Comput. Inf. Sci.*, vol. 100 CCIS, pp. 119–129, 2011, doi: 10.1007/978-3-642-23138-4_8.

[10] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," *Proc. Int. Conf. Asian Lang. Process. 2014, IALP 2014*, pp. 66–69, 2014, doi: 10.1109/IALP.2014.6973519.

[11] S. Fu, N. Lin, G. Zhu, and S. Jiang, "Towards Indonesian Part-of-Speech Tagging : Corpus and Models," *Proc. Lr. 2018 Work. Belt Road Lr.*, vol. 1, pp. 2–7, 2018.

[12] V. Mitra, H. Sujaini, and A. B. P. Negara, "untuk Korpus Paralel Indonesia - Inggris dengan Metode HTML DOM," *J. Sist. dan Teknol. Inf.*, vol. 5, no. 1, pp. 1–6, 2017.

[13] K. M. Lelywiary, C. J. S.; Widowati, S.; & L, "Deteksi Pola Ambiguitas Struktural pada Spesifikasi Perangkat Lunak menggunakan Pemrosesan Bahasa Alami," vol. 4, pp. 51–64, 2019, doi: 10.21108/indojc.2019.4.3.355.

[14] K. E. Dewi, N. I. Widiastuti, and E. Rainarli, "Evaluasi Sentence Extraction pada Peringkasan Dokumen Otomatis," no. September, pp. 8–12, 2017.

[15] P. Koehn and H. Hoang, "Factored translation models," *EMNLP-CoNLL 2007 - Proc. 2007 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, no. June, pp. 868–876, 2007.

[16] V. M. Sánchez-Cartagena, N. Ljubešić, and F. Klubička, "Dealing with data sparseness in SMT with factored models and morphological expansion: A case study on Croatian," *Proc. 19th Annu. Conf. Eur. Assoc. Mach. Transl. EAMT 2016*, vol. 4, no. 2, pp. 354–360, 2016.

[17] P. Bhattacharyya, "Role of Morphology Injection in SMT : A Case Study," vol. 17, no. 1, 2017.

[18] H. Thu, Z. Aye, C. Ding, W. P. Pa, and K. T. Nwet, "English-to-Myanmar Statistical Machine Translation Using a Language Model on Part-of-Speech in Decoding," *Fifteenth Int. Conf. Comput. Appl. (ICCA 2017)*, 2017.

[19] J. Tiedemann *et al.*, "Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools," vol. 2, pp. 391–398, 2016, doi: 10.18653/v1/w16-2326.

[20] S. Mandira, H. Sujaini, and A. B. Putra, "Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik," *J. Edukasi dan Penelit. Inform.*, vol. 2, no. 1, pp. 3–7, 2016, doi: 10.26418/jp.v2i1.13393.

[21] H. Sujaini, A. A. Arman, and A. Purwarianti, "Pengaruh Part-of-Speech Pada Mesin Penerjemah Bahasa Inggris-Indonesia Berbasis Factored Translation Model," vol. 2012, no. Snati, pp. 15–16, 2012.

[22] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2010.