JITE, 5(1) July 2021 ISSN 2549-6247 (Print) ISSN 2549-6255 (Online)

JITE (Journal of Informatics and Telecommunication Engineering)

Available online http://ojs.uma.ac.id/index.php/jite **DOI : 10.31289/jite.v5i1.5308**

Received: 30 May 2021	Accepted: 29 June 2021	Published: 18 July 2021

Minang and Indonesian Phrase-Based Statistical Machine Translation

Muhammad Sandika Alam[1)]*, Arie Ardiyanti Suryani[2)]

[1,2)]Program Studi Teknik Informatika, Universitas Telkom, Indonesia

*Coresponding Email: sandikaalam@student.telkomuniversity.ac.id

Abstrak

Penelitian ini berfokus dalam pembuatan mesin penerjemah statistik bahasa Minang – Indonesia serta melihat seberapa baik hasil terjemahan mesin. Sumber data *training* dan *test* berupa *corpus parallel* dan *corpus monolingual* berasal dari *Wikipedia* bahasa Minang dan *website* berita bahasa Indonesia. Semua eksperimen konfigurasi dilakukan dengan menggunakan data uji sebanyak 600 baris kalimat. Dua skenario pengujian dilakukan yaitu dengan menggunakan skenario yang berdasarkan *monolingual corpus* dan *parallel corpus*. Untuk melihat seberapa baiknya terjemahan akan dilihat dengan menggunakan pengujian otomatis *Bilingual Evaluation Understudy* (BLEU). Hasil pengujian pada 6 konfigurasi menunjukan adanya peningkatan nilai akurasi mesin penerjemah setelah jumlah *corpus monolingual* dan *parallel* ditambahkan. Pada skenario pertama konfigurasi 3 dan 2 meningkat sebesar 3,6% konfigurasi 2 dan 1 meningkat sebesar 2,59%. Pada skenario dua konfigurasi 5 dan 4 meningkatkan sebesar 0,44% konfigurasi 4 dan 1 meningkat sebesar 0,06%. Hasil penelitian menunjukan bahwa pada pengujian skenario pertama memiliki dampak yang signifikan dibandingkan dengan pengujian scenario kedua dalam segi terjemahan. Kurangnya sumber *corpus* menjadi masalah dalam membuat mesin penerjemah statistik.

Kata Kunci: bahasa minang, bahasa Indonesia, mesin penerjemah statistik, bilingual evaluation understudy

Abstract [Font: Cambria, size, 9, Italic - Bold]

This research focuses on making a phrase based statistical machine translation for Minang – Indonesian language as well as seeing how well the machine translation results. The source of training and test data in the form of parallel corpus and monolingual corpus that taken from Minang Wikipedia language and Indonesian news website. All configuration experiments were carried out using 600 lines of sentences. Two test case scenario were tested in this research that based on the monolingual corpus and parallel corpus. To see how well the translation will be seen by using automatic testing Bilingual Evaluation Understudy (BLEU). The test results on 6 configurations show an increase in the accuracy of the translator machine after the number of monolingual and parallel corpus is added. In the first scenario configurations 3 and 2 increased by 3.6%, configurations 2 and 1 increased by 2.59%. In the scenario of two configurations 5 and 4 it increases by 0.44%, configurations 4 and 1 increase by 0.06%. The result showed that the testing for the first scenario have a significant impact compare to the second scenario in terms of translation. The lack of corpus resources is a problem in building phrase-based statistical machine translation.

Keywords: minang language, indonesian language, phrase based statistical machine translation, bilingual evaluation understudy

How to Cite: Alam, M. S., & Suryani, A. A.. (2021). Minang and Indonesian Pharase-Based Statistical Machine Translation. *JITE (Journal Of Informatics And Telecommunication Engineering).* 5 (1): 216-224

I. PENDAHULUAN

Bahasa merupakan salah satu cara manusia untuk dapat mengungkapkan sebuah gagasan, ide atau menyampaikan perasaan kepada orang lain. Dengan adanya bahasa, dua individu atau lebih dapat mengespresikan berbagai ide, perasaan, arti dan pengalaman (Fridani, Lara; Dhieni, 2014). Indonesia memiliki sangat banyak bahasa yang beragam, dari daerah timur hingga barat. Salah satu contohnya adalah bahasa Minang, bahasa ini berasal dari daerah Sumatera Barat.

Bahasa Minang sudah mulai kurang digunakan, menurut Lembaga Kerapatan Adat Alam Minangkabau (LKAAM) banyak warga yang berdarah Minangkabau tidak mengajari anaknya sendiri untuk berbahasa Minang, tetapi mengajarkan bahasa Indonesia dan Inggris ("Penutur Bahasa Minang Dikhawatirkan Berkurang," n.d.). Hal ini dapat membuat bahasa Minang menjadi bahasa daerah yang terancam punah.

Dalam penelitian ini mesin penerjemah berperan sebagai solusi dalam masalah ini. Penelitian yang telah dilakukan dalam kasus menerjemahkan bahasa Minang dan bahasa Indonesia menggunakan metode *rule-based*, di dalam penelitian tersebut menunjukan bahwa ketepatan hasil terjemahan adalah 97% (Soyusiawaty, 2008). Hasil terjemahan pada *rule-based* memiliki akurasi yang baik dikarenakan bahasa terjemahannya berdasarkan dari seorang ahli bahasa, sehingga hasil mesin terjemahan ini bahasanya sangat alami. Selain metode *rule-based,* ada juga metode lain yang dapat melakukan penerjemahan yaitu metode statistik.

Mesin penerjemah statistik tidaklah membutuhkan seorang ahli bahasa karena hasil terjemahan dihasilkan berdasrkan probabilitas yang dihitung di dalam mesin terjemahan, sehingga mesin terjemahan statistik jauh lebih murah dari segi biaya dibadingkan dengan metode *rule-based* yang menggunakan seorang ahli bahasa untuk membentuk mesin penerjemahnya.

II. STUDI PUSTAKA

Penerjemah statisik atau *Statistical Translation* merupakan penerjemah yang menggunakan pendekatan statistik dalam menghitung probabilitas kalimat yang akan diterjemahkan. Penerjemah statistik menghasilkan terjemahan yang lebih baik *inrerlinear translation* atau penerjemahan kata demi kata dengan syarat *parallel corpus* yang digunakan memiliki kualitas yang bagus serta jumlah yang cukup banyak (Permata, Abidin, & Ariyani, 2020). Dalam mesin penerjemah statistik terdapat tiga komponen arsitektur yang berperan yaitu*: language model, translation model, dan decoder. Language model* berfungsi untuk mencari kefasihan (*fluency)* terjemahan dalam *language model statistic*, memberikan probabilitas untuk setiap kata yang menunjukan seberapa besar kemungkinan kata tersebut akan terjadi selanjutnya. Salah satu pendekatan *language model* adalah *n-gram model* yang akan memprediksi kata berikutnya dalam urutan tersebut (Hidayat, Sjaini, & Dwinyoto, 2015). *Transltaion model* berfungsi untuk memasangkan input teks dalam bahasa sumber dengan teks dalam bahasa target. Tujuan utama dari *translation model* adalah untuk mencari ketepatan dalam penerjemahan (Nugroho Aditya, Adji Bharata, & Hantono S, 2015).

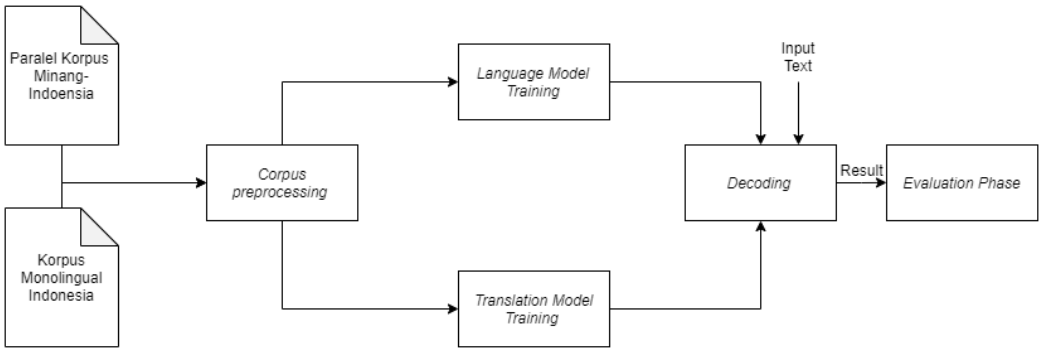
Pada 2016 Yosep Jarob, Henry Sujaini, Nofi Safriadi melakukan penelitian dengan judul "Uji Akurasi Penerjemah Bahasa Dayak Taman demgan Penandaan Kata Dasar dan Imbuhan" (Jarob, Sujaini, & Safriadi, 2016). Penulis melakukan penelitian ini untuk membantu orang berkomunikasi dengan bahasa yang berbeda dengan menrancang mesin penerjemah statistik menggunakan Moses. Dalam penelitian ini digunakan data penlitian yang didapatkan dari dokumen-dokumen bahasa Dayak Taman. Data tersebut lalu diolah menjadi *parallel corpus* bahasa Indonesia – Dayak Taman sebanyak 3110 pasang kalimat dan menghasilkan akurasi terjemahan sebesar 80.17% (Jarob et al., 2016).

Pada 2015 Andri Hidayat, Henry Sujainin, Rudy Dwinyoto melakukan penelitian dengan judul "Aplikasi Penerjemah Duah Arah Bahasa Indonesia – Bahasa Melayu Sambas Berbasis Web dengan Megggunakan *Decoder Moses"*(Hidayat et al., 2015). Penelitian ini dilakukan untuk mengatasi keterbatasan dalam menguasasi berbagai macam bahasa daerah di Indonesia. Penelitian ini menggunakan 1197 pasang kalimat atau *parallel corpus* untuk dilakukan pengujian pada mesin statisiik penelitian ini mendapatkan 58.50% untuk bahasa Indonesia ke Bahasa Melayu Sambas dan 63.76% untuk bahas Melayu Sambas ke bahasa Indonesia (Hidayat et al., 2015).

Pada 2015 Mohammad Anugrah Sulaeman, Ayu Purwarianti melakukan penelitian dengan judul "*Development of Indonesia-Japanese Statistical Machine Translation Using Lemma Translation and Additonal Post-Process"* (Sulaeman & Purwarianti, 2015). Data yang dikumpulkan untuk penelitian tersebut berasal dari JLPT (*Japanese Language Proficiency Test*) sebanyak 1132 kalimat. Penelitian tersebut lebih fokus pada mengatasi kata-kata yang tidak diketahui, masalah penyusunan ulang kalimat dengan menggunakan Lemma dan POSTAG. Untuk hasil terjemahan tanpa menggunaka Lemma dan POSTAG adalah sebesar 6.53% untuk Jepang-Indonesia dan 13.69% untuk Indonesia-Jepang(Sulaeman & Purwarianti, 2015).

III. METODE PENELITIAN

Pada penelitian ini dibangun mesin penerjemah statistik bahasa Minang dan bahasa Indonesia. Gambar model yang dibangun dapat dilihat pada Gambar 1.



Gambar 1 Arsitektur Mesin Penerjemah Statistik Bahasa Minang – Indonesia

Terdapat beberapa fase yang akan dilakukan oleh mesin penerjemah statisik Minang – Indonesia. *corpus preprocessing, language model training, translation model training, decoding,* dan terakhir *evaluation phase* (Raju & Raju, 2016).

A. Corpus Preprocessing

Corpus Preprocessing merupakan tahapan awal dari mesin penerjemah statisik Minang-Indonesia, tahapan ini merupakan persiapan *corpus* yang nanti akan digunakan pada mesin penerjemah statistik. Pada tahapan ini *corpus* akan dilakukan pembersihan. Pembersihan disini dilakukan agar korpus dapat digunakan untuk melatih model terjemahan atau bahasa (Firdaus, Suryani, & Ramadhani, 2017). Dalam proses *preprocessing* hal pertama yang dilakukan adalah melakukan tokenisasi dimana proses ini melakukan pemisahan antara kata dengan kata lain atau tanda baca.

B. Language Model, Translation Model, Decoding

Language model digunakan untuk menghasilkan kefasihan (*fluency*) terjemahan, memberikan probabilitas untuk setiap kata yang menunjukan seberapa besar kemungkinan kata tersebut akan terjadi selanjutnya. Untuk menciptakan *language model* digunakan sebuah *tool* yang bernama IRSTLM. Data yang digunakan dalam pembuatan *language model* berupa *monolingual corpus* bahasa Indonesia yang sudah dilakukan *preprocessing corpus* sebelumnya. Untuk membuat *language model* digunakan metode yang bernama *n-gram* yaitu Unigram, Bigram, Trigram (Apriani, Sujaini, & Safriadi, 2016).

$$\begin{aligned}&\text{Unigram (1-gram)} : P(w1\,), P(w2) \ldots P(wn)\\ &\text{Bigram (2-gram)} : P(w1), P(w2\,|w1\,) \ldots P(wn|w(n-1))\\ &\text{Trigram (3-gram)} : P(w1,n) = P(w1), P(w2|w1),\\ &P(w3|w1,2) \ldots P(wn|wn-2\,,n-1)\end{aligned} \quad (1)$$

Disini kalimat *corpus monolingual* diubah kedalam *n-gram* tersebut sehingga mesin penerjemah mengetahui kalimat apa yang memungkinkan terjadi selanjutnya. Semakin tinggi nilai probabilitas yang dihasilkan menunjukan bahwa kalimat yang dibentuk dengan baik (Tanuwijawa & Maruli Manurung, n.d.). Dalam membuat *n-gram* pertama harus dilakukan *data training* terlebih dahulu, berikut rumus dalam menghitung prediksi probabilitas *Trigram*.

$$p(w_3|w_1,w_2) = \frac{Count(w_1,w_2,w_3)}{\sum_w Count(w_1,w_2,w)} \quad (2)$$

Dimana p(W3|W1,W2) probabilitas kalimat yang dicari dengan mencari berapa banyak kemunculan dengan kalimat yang sama dibagi dengan jumlah kata yang mengandung w_1 dan w_2.

Translation model digunakan untuk memasangkan input teks dalam bahasa sumber dengan teks bahasa target. Disini kalimat dibagi-bagi menjadi potongan kata-kata untuk menerjemahkan setiap frasa ke tujuan, dan *reordering.* Tugas utama dari translation model adalah membuat padanan kata yang akan digunakan dalam terjemahan. Di tahap ini akan dilakukan analisis dari semua kemungkinan pola kalimat target(Singvongsa & Seresangtakul, 2016).

Proses training dari *translation model* akan menggunakan *tool* yang berada di *moses translation* bernama GIZA++. Data yang digunakan dalam melakukan pelatihan data berupa *parallel corpus* bahasa Minang-Indonesia.

Decoding merupakan tahapan dimana mesin penerjemah akan melakukan uji coba ke pada mesin penerjemah statistik bahasa Minang – Indoneisa. Tahapan ini digunakan untuk menemukan teks atau kalimat bahasa target yang memiliki probabilitas paling besar berdasarkan faktor dari model terjemahan (f|e) dan model bahasa P(e) yang sebelumnya sudah dilatih (Dharmawan, Sujaini, & Muhardi, 2020). Nantinya bahasa sumber atau bahasa Minang akan di proses sehingga nanti mesin akan mengeluarkan terjemahan ke dalam bahasa target yaitu bahasa Indonesia. berikut adalah rumus dalam menghitung probabilitas (Turitzin, n.d.)

$$\acute{e} = arg_e \max P(e|f) = arg_e \max P(f|e)\, P(e) \quad (3)$$

C. *Evaluation Phase*

Evaluation Phase merupakan tahapan terakhir dari mesin penerjemah statistik bahasa Minang dan bahasa Indonesia. Di tahapan ini akan dilakukan dengan menggunakan BLEU-metric (*Bilingual Evaluation Understudy*) yang merupakan pengujian otomatis. Nilai BLEU-metric didapatkan dari perkalian *brevity penalty* berdasarkan rata – rata geometri dari *modified precision score.* Semakin banyak terjemahan rujukan perkalimat, maka akan semakin tinggi nilainya. Panjang kalimat terjemahan harus mendekati panjang kalimat referensi dan memiliki urutan yang sama (Mandira, Sujaini, & Putra, 2016). Berikut adalah rumus BLEU:

$$BP_{BLEU} = \begin{cases} 1, & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)}, & if\ c \leq r \end{cases}$$

$$BLEU = BP_{BLEU} \cdot e^{\sum_{n=1}^{N} W_n \, Log \, P_n} \quad (4)$$

Nilai Standart untuk N adalah 4 dan c merupakan jumlah kata dari translasi otomatis dan r merupakan kata referensi.

IV. HASIL DAN PEMBAHASAN

A. *Corpus Preparation*

Penelitian ini meggunakan *parallel corpus* yang dikumpulkan dari website *Wikipedia* bahasa Minang. Korpus yang digunakan berjumlah sebanyak 1300 baris kalimat dimana 700 korpus digunakan sebagai data training dan 600 digunakan sebagai data testing. selain *parallel corpus* digunakan *juga monolingual corpus* yang bersumber dari webstie berita bahasa Indonesia yaitu cnnindonesia.com dan kompas.com yang dikumpulkan sebanyak 3000 korpus. Pengumpulan data dilakukan secara manual seperti

menerjemahkan kalimat dalam korpus bahasa minang yang telah dikumpulkan ke bahasa Indonesia dan pengumpulan *monolingual corpus* sebanyak 3000 baris kalimat.

Tabel 1. Contoh Tokenisasi

No	Kalimat	Tokenisasi
1	Pada musim liburan, air terjun ini dapat dikunjungi hingga ratusan penggunjung.	{Pada musim liburan} {,} {air terjun ini dapat dikunjungi hingga ratusan penggunjung} {.}
2	Pemandian ini dikelolah oleh BUMNag (Badan Usaha Milik desa) setempat.	{Pemandian ini dikelolah oleh BUMNag} {(} {Badan Usaha Milik desa} {)} {setempat} {.}
3	Oleh orang yang tua-tua, mengobrol sambil memesan teh telur bisa sampai begadang; orang toko mau juga membuka kedainya sampai menjelang subuh.	{Oleh orang yang tua-tua} {,} {mengobrol sambil memesan teh telur bisa sampai begadang} {;} {orang toko mau juga membuka kedainya sampai menjelang subuh} {.}
4	"Jayakarta" berarti "kemenangan", "kesempurnaan" atau "kemenangan yang sempurna".	{"} {Jayakarta} {"} {berarti} {"} {kemenangan} {"} {,} {"} {kesempurnaan} {"} {atau} {"} {kemenangan yang sempurna} {"} {.}

Dapat dilihat pada Tabel 1 bahwa kata – kata akan dipisahkan dari tanda baca sepert koma, titik, kurung buka dan tutup sertpa kutip dua. Untuk kutip dua dipisahkan dengan cara ditulis sebagai " Setelah proses tokenisasi selesai dilakukan proses *truecasing* dimana mengubah huruf besar atau kecil yang memungkinkan.

Tabel 2 Contoh *Truecasing*

No	Tokenisasi	*Truecasing*
1	Pada musim liburan , air terjun ini dapat dikunjungi hingga ratusan penggunjung .	**pada** musim liburan , air terjun ini dapat dikunjungi hingga ratusan penggunjung .
2	Pemandian ini dikelolah oleh BUMNag (Badan Usaha Milik desa) setempat .	**pemandian** ini dikelolah oleh BUMNag (Badan Usaha Milik desa) setempat .
3	Oleh orang yang tua-tua , mengobrol sambil memesan teh telur bisa sampai begadang ; orang toko mau juga membuka kedainya sampai menjelang subuh .	**oleh** orang yang tua-tua , mengobrol sambil memesan teh telur bisa sampai begadang ; orang toko mau juga membuka kedainya sampai menjelang subuh .
4	" Jayakarta " berarti " kemenangan " , " kesempurnaan " atau " kemenangan yang sempurna " .	" Jayakarta " berarti " kemenangan " , " kesempurnaan " atau " kemenangan yang sempurna " .

Dapat dilihat pada Tabel 2 bahwa huruf besar dan kecil yang memungkinkan seperti pada kalimat "Pemandian ini dikelolah oleh BUMNag (Badan Usaha Milik desa) setempat." kata "Pemandian" disini diubah menjadi huruf kecil karena pada korpus yang diolah lebih mungkin kata tersebut dalam huruf kecil daripada huruf besar. Proses terakhir dalam *corpus preprocessing* adalah *cleaning* dimana pada proses ini akan dilakukan penghilangan terhadap *empty space* yang berlebih dan memotong kalimat yang panjang.

B. *Language Model* dan *Translation Model Training*

Pada tahapan ini *language model* akan menghasilkan output file lm. dimana dalam file ini merupakan tabel model bahasa yang dihasilkan oleh IRSTLM. Gambar 2 adalah contoh output file yang dihasilkan.

```
...
-0.342043       orang lainnya meninggal
-0.140842       orang dinyatakan sembuh
-0.128073       orang meninggal dunia
-0.573963       dengan jumlah kasus
-0.0304773      dengan protokol kesehatan
...
```

Gambar 2 Tabel Language Model

Gambar 2 menunjukan bahwa kemungkinan kemunculan kata orang diikuti dengan kalimat lainnya meninggal adalah 10^(-0.342043) = 2.1981, kemunculan kata orang diikuti dengan dinyatakan sembuh adalah 10^(-0.140842) = 1.3831.

Setelah *language model* proses selanjutnya adalah membuat *translation model.* translation model dibuat menggunakan tool GIZA++ proses pelatihan menggunakan *corpus parallel* bahasa Minang – Indonesia yang berjumlah sebanyak 700 kalimat.

```
1       UNK     0
2       ,       856
3       .       702
4       yang    391
5       dan     390
...
```

Gambar 3 Vocabulary Corpus Bahasa Indonesia

Pada Gambar 3 merupakan isi dari *Vocabulary corpus* Bahasa Indonesia. Angka 1 sampai 5 pada dokumen ini merupakan *uniq id* untuk setiap kata, sedangkan angka disebelah kanan menunjukan kemunculan frekuensi kata(Mandira et al., 2016).

```
...
pangsit mie 0.1666667
pangsit pangsit 1.0000000
mandorong mendorong 1.0000000
minik menit 1.0000000
rentannyo rentannya 1.0000000
...
```

Gambar 4 Tabel lexical model

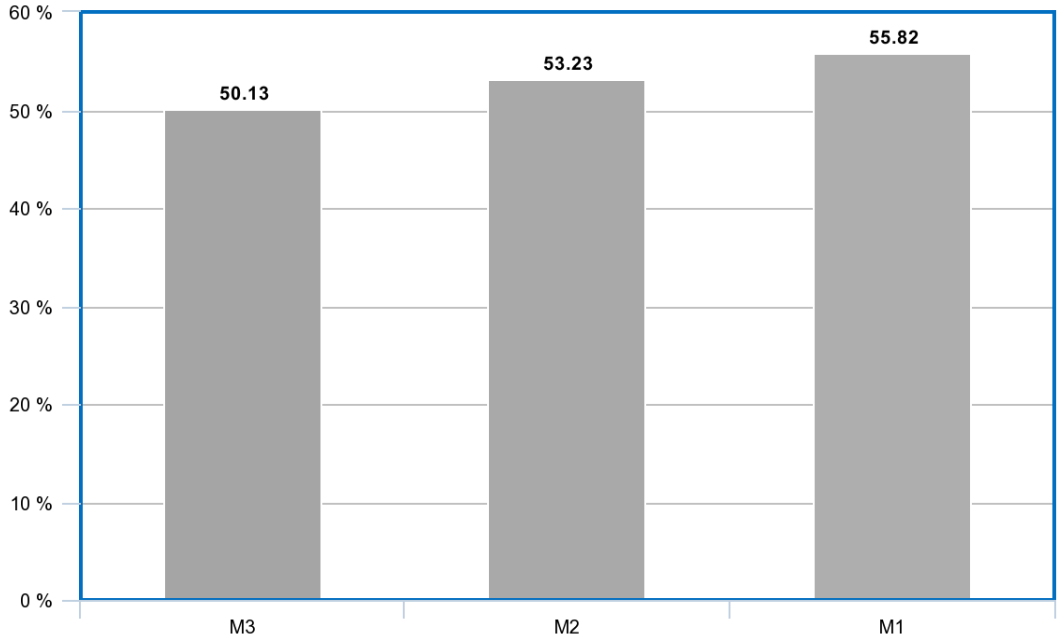
Pada Gambar 4 merupakan tabel *lexical model* pada mesin penerjemah statistik bahasa Minang-Indonesia. Proses ini akan menghasilkan tabel translasi *lexical model* dimana terdiri dari kata-kata dari bahasa sumber yaitu bahasa Minang dan memiliki makna dalam bahasa Indonesia ataupun sebaliknya. Proses *Translation Model* juga dilakukan *alignment* dimana kalimat pada bahasa sumber dipasangkan dengan bahasa target, hasil *alignment* ini dapat kita lihat pada Gambar 5.

```
Babeda-beda tapi tatok ciek ) ||| Berbeda-beda tapi tetap satu ) ||| 1 0.16 1 0.711462 ||| 0-0 1-1 2-2 3-3 4-4 ||| 1 1 1 ||| |||
Babeda-beda tapi tatok ciek ||| Berbeda-beda tapi tetap satu ||| 1 0.16 1 0.711462 ||| 0-0 1-1 2-2 3-3 ||| 1 1 1 ||| |||
Babeda-beda tapi tatok ||| Berbeda-beda tapi tetap ||| 1 0.4 1 0.782609 ||| 0-0 1-1 2-2 ||| 1 1 1 ||| |||
Babeda-beda tapi ||| Berbeda-beda tapi ||| 1 1 1 0.782609 ||| 0-0 1-1 ||| 1 1 1 ||| |||
```

Gambar 5 Tabel Frasa

C. Analisis

Untuk melihat seberapa baik mesin Penerjemah statistik bahasa Minang dan bahasa Indonesia, sejumlah pengujian akan dilakukan. Pengujian pertama dilakukan dengan melihat seberapa efisien *parallel corpus* dalam mesin penerjemah statistik bahasa Minang dan Indonesia. disini akan dilakukan pengujian dengan menggunakan banyak *parallel corpus* yang bervariasi dan *monolingual corpus* sebanyak 3000. Pengujian kedua dilakukan dengan menggunakan banyak *monolingual corpus* yang bervariasi dan *parallel corpus* sebanyak 700. Pengujian dilakukan dengan menggunakan data *testing* sebanyak 600 *corpus parallel* banyaknya.



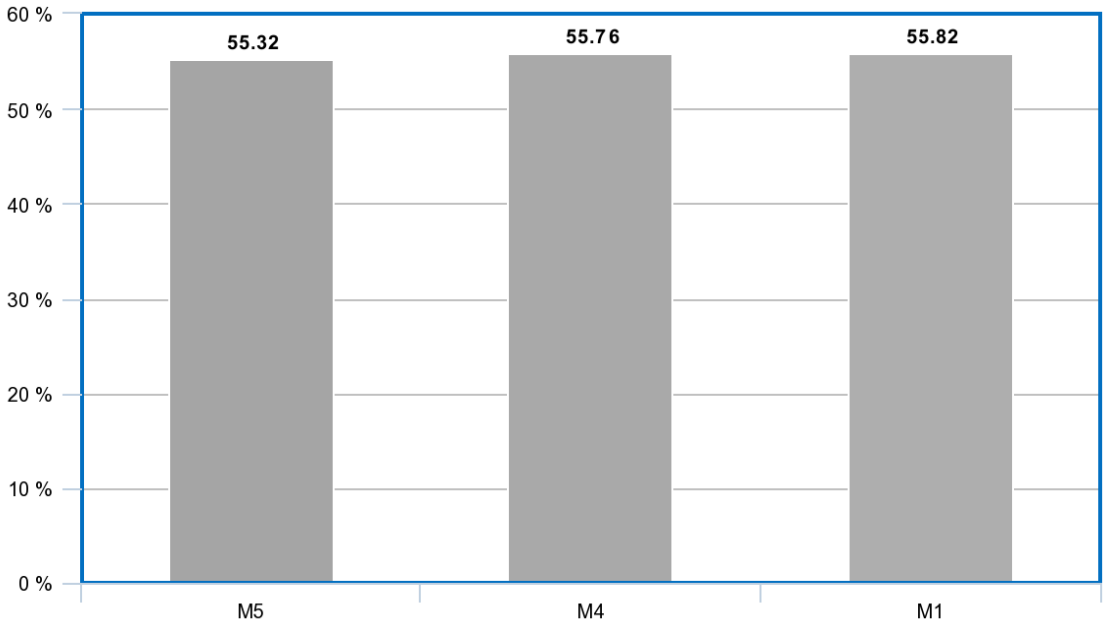
Gambar 6 Hasil Akurasi Pengujian Pertama

Untuk hasil pengujian pertama dapat dilihat pada Gambar 5. Dari hasil pengujian dapat dilihat dampak *parallel corpus* yang digunakan, semakin banyak yang digunakan maka semakin baik juga akurasi yang dihasilkan. Dari pengujian tersebut M3 menggunakan *parallel corpus* sebanyak 250 dan menghasilkan akurasi terjemahan sebesar 50.13%. M2 menggunakan *parallel corpus* sebanyak 500 dan menghasilkan akurasi terjemahan sebesar 53.23%. Kemudian M3 menggunakan *parallel corpus* sebanyak 700 dan menghasilkan akurasi terjemahan sebesar 55.82%.

Tabel 3 Contoh kalimat pengujian pertama

No	Bahasa Minang	M3	M2	M1	Bahasa Indonesia
1	organisme multisel dan banyak **jinih** organisme sel **tungga**	organisme multisel dan banyak **jinih** organisme sel **tunggal**	organisme multisel dan banyak **jinih** organisme sel **tunggal**	organisme multisel dan banyak **jenis** organisme sel **tunggal**	organisme multisel dan banyak **jenis** organisme sel **tunggal**
2	**budayo basifaik** kompleks, abstrak, dan **lueh**.	**budaya basifaik** kompleks , abstrak , dan **lueh** .	**budaya basifaik** kompleks , abstrak , dan **lueh** .	**budaya basifaik** kompleks , abstrak , dan **luas** .	**budaya bersifat** kompleks, abstrak, dan **luas**.
3	Batolak dari teori evolusi, **inyo** mengasumsikan bahwa **satiok** manusia **tumbuah**	Batolak dari teori evolusi , **dia** mengasumsikan bahwa **satiok** manusia **tumbuah**	Batolak dari teori evolusi , **dia** mengasumsikan bahwa **setiap** manusia **tumbuh**	Batolak dari teori evolusi , **dia** mengasumsikan bahwa **setiap** manusia **tumbuah**	Bertolak dari teori evolusi , **dia** mengansumsikan bahwa **setiap** manusia **tumbuh**

Dari Tabel 3 dapat dilihat hasil terjemahan kalimat pada setiap mesin. Dari hasil pengujian pertama dapat dilihat pada percobaan kalimat nomor 1, M1 berhasil menerjemahkan kalimat pertama dengan sempurna sedangkan pada M2 dan M3 masih belum berhasil menerjemahkan kata “jinih” yang berarti “jenis” dalam bahasa Indonesia. Pada percobaan kalimat nomor 2 dapat dilihat setiap mesin berhasil menerjemahkan kata Budaya akan tetapi pada kata “basifaik” setiap mesin tidak berhasil menerjemahkan. Dapat dilihat juga pada kata “lueh” mesin M1 berhasil menerjemahkan kata dengan benar sedangkan pada M2 dan M3 masih belum. Pada percobaan kalimat nomor 3 dapat dilihat setiap mesin berhasil menerjemahkan kata “inyo” yang berarti “dia” dalam bahasa Indonesia, tetapi untuk kata “satiok” M3 belum benar dalam menerjemahkan kata tersebut. Untuk kata “tumbuah” mesin M2 berhasil menerjemahkan kata dengan benar yaitu ”tumbuh” dalam bahasa Indonesia sedangkan M1 dan M3 masih belum benar dalam terjemahan.



Gambar 7 Hasil Akurasi Pengujian Kedua

Untuk hasil penerjemahan percobaan kedua dapat dilihat pada Gambar 6. Dari hasil pengujian dapat dilihat bahwa pengaruh *monolingual corpus* tidak terlalu berdampak besar namun semakin banyak korpus yang digunakan hasil akurasi akan meningkat. Dari percobaan ini M1 menggunakan *monolingual corpus* sebanyak 3000, menghasilkan akurasi terjemahan sebesar 55.82%. M4 menggunakan *monolingual corpus* sebanyak 2000, menghasilkan akurasi terjemahan sebesar 55.76%. kemudian M5 menggunakan *monolingual corpus* sebanyak 1000 dan menghasilkan terjemahan sebesar 55.32%.

Tabel 4 Contoh kalimat pengujian kedua

No	Bahasa Minang	M5	M4	M1	Bahasa Indonesia
1	kabudayoan nan dipunyoi dek masyarakaik tu **surang**.	kebudayaan yang dipunyoi oleh masyarakat itu **Seorang** .	kebudayaan yang dipunyoi oleh masyarakat itu **Seorang** .	kebudayaan yang dipunyoi oleh masyarakat itu **sendiri** .	kebudayaan yang dipunyai oleh masyarakat itu **sendiri**.
2	kabudayoan sangaik arek hubungannyo **jo** masyarakaik.	kebudayaan sangat erat hubungannyo **dan** masyarakat .	kebudayaan sangat erat hubungannyo **dan** masyarakat .	kebudayaan sangat erat hubungannyo **dan** masyarakat .	kebudayaan sangat erat hubunganya **dengan** masyarakat.
3	babarapo anak sungai akan bagabuang untuak **mambantuak** sungai utamo.	beberapa anak sungai akan bagabuang untuk **membentuk** sungai utama .	beberapa anak sungai akan bagabuang untuk **membentuk** sungai utama .	beberapa anak sungai akan bagabuang untuk **membuat** sungai utama .	beberapa anak sungai akan bergabung untuk **membentuk** sungai utama.

Dari tabel 4 kita dapat melihat hasil terjemahan pada setiap mesin. Pada percobaan kalimat 1 dapat dilihat terdapat perbedaan pada penerjemahan kata “surang” untuk mesin M5 dan M4, mesin menerjemahkan kata tersebut dengan kata “seorang” sedangkan terjemahan yang tepat adalah “sendiri”. Pada percobaan kalimat 2 setiap mesin menerjemahkan kalimat “jo” dengan arti “dan”, hasil terjemahan yang tepat seharusnya adalah “dengan”. Pada percobaan kalimat 3 dapat dilihat M1 menerjemahkan kata “mambantuak” dengan arti membuat sedangakn mesin M4 dan M5 menerjemahkan dengan benar yaitu “membentuk”.

V. KESIMPULAN

Bedasarkan hasil pengujian yang dilakukan dapat diasumsikan bahwa mesin penerjemah statistik Minang- Indonesia berbasis frasa menunjukan bahwa masih banyak kesalahan dalam menerjemahkan kalimat dalam experimen ini. Konfigurasi terbaik menunjukan akurasi sebesar 55.82% berdasarkan niali BLEU *score.* Kurangnya *parallel corpus* dan *monolingual corpus* dan tidak konsisntennya dalam mengetik membuat hasil terjemahan tidak akurat. Untuk penelitian selanjutnya disaraknkan untuk meningkatkan kualitas *parallel corpus* dan *monolingual corpus* bahasa Minang – Indonesia serta menambahkan PoS Tag dalam pembuatan mesin penerjemah bahasa Minang - Indonesia.

DAFTAR PUSTAKA

Apriani, T., Sujaini, H., & Safriadi, N. (2016). Pengaruh Kuantitas Korpus Terhadap Akurasi Mesin Penerjemah Statistik Bahasa Bugis Wajo ke Bahasa Indonesia. *JUSTIN (Jurnal Sistem Dan Teknologi Informasi)*, *4*(1), 168–173.

Dharmawan, E., Sujaini, H., & Muhardi, H. (2020). *Perbandingan Nilai Akurasi Terhadap Penggunaan Part of Speech Set pada Mesin Penerjemah Statistik Comparison of The Accuracy Value Toward Using Part of Speech Sets on Statistical Machine Translation*. *08*(3), 250–256. https://doi.org/10.26418/justin.v8i3.39810

Firdaus, A., Suryani, A. A., & Ramadhani, K. N. (2017). *Pengumpulan Korpus Paralel Bahasa Indonesia-Sunda dari Wikipedia Menggunakan Metode Pointwise Mutual Information Indonesian-Sundanese Parallel Corpus Retrieval from Wikipedia Using Pointwise Mutual Information Method*. *4*(3), 4859–4865.

Fridani, Lara; Dhieni, N. (2014). Hakikat Perkembangan Bahasa Anak. *Metode Pengembangan Bahasa*, 1–28.

Hidayat, A., Sjaini, H., & Dwinyoto, R. . *Aplikasi Penerjemah muka Bahasa Indonesia – Bahasa Melayu Sambas Berbasis Web dengan Menggunakan Decoder Moses*. 0–5.

Jarob, Y., Sujaini, H., & Safriadi, N. (2016). Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman Dengan Penandaan Kata Dasar Dan Imbuhan. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *2*(2), 78–83. https://doi.org/10.26418/jp.v2i2.16520

Mandira, S., Sujaini, H., & Putra, A. B. (2016). Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *2*(1), 3–7. https://doi.org/10.26418/jp.v2i1.13393

Nugroho Aditya, R., Adji Bharata, T., & Hantono S, B. Penerjemahan Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Berbasis Frasa. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, *2015*(Sentika).

Penutur Bahasa Minang Dikhawatirkan Berkurang. (n.d.). Retrieved December 1, 2019, from https://www.harianhaluan.com/news/detail/64741/penutur-bahasa-minang-dikhawatirkan-berkurang

Permata, P., Abidin, Z., & Ariyani, F. (2020). Efek Peningkatan Jumlah Paralel Korpus Pada Penerjemahan Kalimat Bahasa Indonesia ke Bahasa Lampung Dialek Api. *Jurnal Komputasi*, *8*(2), 41–49. https://doi.org/10.23960/komputasi.v8i2.2613

Raju, B. N. V. N., & Raju, M. S. V. S. B. (2016). Statistical Machine Translation System for Indian Languages. *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, 174–177. https://doi.org/10.1109/IACC.2016.41

Singvongsa, K., & Seresangtakul, P. (2016). Lao-Thai machine translation using statistical model. *2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016*. https://doi.org/10.1109/JCSSE.2016.7748893

Soyusiawaty, D. E-Translator With Rule Based Indonesia – Minang Dan Minang – Indonesia. *Jurnal Informatika*, *2*(2), 234–247. https://doi.org/10.26555/jifo.v2i2.a5255

Sulaeman, M. A., & Purwarianti, A. (2015). Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process. *Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015*, (i), 54–58. https://doi.org/10.1109/ICEEI.2015.7352469

Tanuwijawa, H., & Maruli Manurung, H. (n.d.). Penerjemahan Dokumen Inggris - Indonesia Menggunakan Mesin Penerjemahan Statistik Dengan Word Reordering dan Pharase Reordering. *Jurnal Ilmu Komputer Dan Informasi*, *2*.

Turitzin, M. (n.d.). *Statistical Machine Translation of French and German into English Using IBM Model 2 Greedy Decoding*.