JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 4
No. 1
Juni 2018

Pengaruh Metode *Dictionary Lookup* pada *Cleaning* Korpus Terhadap Akurasi Mesin Penerjemah Statistik Indonesia–Melayu Pontianak

M. Dwi Etsa Putra[#1], Herry Sujaini [#2], Novi Safriadi[#3]
[#]*Program Studi Informatika Universitas Tanjungpura*
Jalan Prof. Dr. H. Hadari Nawawi, Pontianak, Kalimantan Barat
[1]putracl@gmail.com
[2]herry_sujaini@yahoo.com
[3]bangnops@gmail.com

Abstrak **- Bahasa Melayu Pontianak merupakan dialek bahasa Melayu yang dituturkan oleh masyarakat Kota Pontianak dan sekitarnya, meskipun masih jauh dari kepunahan namun perlu dilestarikan sebagai tindakan pencegahan agar tidak punah, salah satu upaya pelestarian bahasa daerah yaitu dengan pembuat mesin penerjemah. Mesin Penerjemah Statistik (MPS) adalah sebuah pendekatan mesin penerjemah dengan hasil terjemahan dihasilkan atas dasar model statistik, namun masih terdapat kelemahan yaitu rendahnya tingkat akurasi terjemahan. Proses *cleaning* adalah proses pencarian dan perbaikan (penghapusan) kata atau kalimat yang salah ataupun tidak sesuai dalam rangka meningkatkan tingkat akurasi terjemahan, salah satu metode yang dapat digunakan pada proses *cleaning* adalah metode *dictionary lookup*. Tujuan dari penelitian ini adalah mengetahui pengaruh penerapan metode *dictionary lookup* pada proses *cleaning* korpus terhadap akurasi mesin penerjemah statistik bahasa Indonesia – bahasa Melayu Pontianak. Penelitian menggunakan korpus paralel sebanyak 9157 kalimat. Pengujian dilakukan dengan membandingkan nilai akurasi hasil terjemahan sebelum dan setelah *cleaning* dengan metode *dictionary lookup*. Pengujian dilakukan dengan pengujian otomatis menggunakan *Bilingual Evaluation Understudy* (BLEU). Dari hasil penelitian, penerapan metode *dictionary lookup* pada proses *cleaning* dapat mempengaruhi akurasi MPS, ini terlihat dari terjadinya penurunan sebesar 1,5% pada korpus manual dan penurunan sebesar 6,94% dengan korpus orisinal sementara itu terjadi peningkatan sebesar 2,58% pada korpus clean dic. Berdasarkan hal tersebut penerapan metode *dictionary lookup* pada proses *cleaning* dapat menurunkan nilai akurasi hasil terjemahan.**

Kata kunci **—** ***BLUE, Cleaning, Dictionary Lookup,*** **Mesin Penerjemah Statistik,** ***Moses Decoder***

I. PENDAHULUAN

Terdapat lebih dari 1340 kelompok etnik atau suku bangsa yang ada di Indonesia dari hasil Sensus Penduduk pada tahun 2010, dan ada sekitar 2500 jenis bahasa daerah yang ada di Indonesia[1]. Pada tahun 2015 terdapat Indonesia mempunyai 726 bahasa daerah[2]. Bahasa Melayu Pontianak merupakan dialek bahasa Melayu yang dituturkan oleh masyarakat Kota Pontianak, Kabupaten Kubu Raya dan Kabupaten Mempawah . Meskipun masih jauh dari kepunahan namun tetap perlu dilakukan pelestarian sebagai tindakan pencegahan agar bahasa Melayu Pontianak tidak punah. Salah satu upaya pelestarian bahasa daerah yaitu dengan pembuat mesin penerjemah.

Mesin penerjemah merupakan alat penerjemah otomatis pada sebuah teks dari satu bahasa ke bahasa lainnya. Mesin Penerjemah Statistik (MPS) adalah sebuah pendekatan mesin penerjemah dengan hasil terjemahan dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus teks bilingual atau paralel[3]. Terdapat beberapa penelitan yang dilakukan berkaitan dengan MPS bahasa daerah diantaranya penelitian tentang pengaruh kuantitas korpus terhadap akurasi MPS bahasa Bugis Wajo ke bahasa Indonesia[4], penelitian tentang sistem penerjemah bahasa Jawa-aksara Jawa berbasis *finite state automata* [5], penelitian akurasi penerjemahan bahasa Indonesia-Jawa menggunakan metode statistik berbasis frasa[6], Algortima Pembagian Frasa dalam Kalimat untuk Menigkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Bugis Wajo[7], Algoritma Pembagian Frasa dalam Kalimat Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Jawa Kromo[8], Meningkatkan Akurasi Pada Mesin Penerjemah Bahasa Indonesia Ke Bahasa Melayu Pontianak Dengan Part Of Speech[9], Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke Bahasa Melayu Pontianak[10], namun masih terdapat kelemahan yaitu rendahnya tingkat akurasi terjemahan. Rendahnya tingkat akurasi dapat dipengaruhi oleh korpus yang digunakan sebagai dasar pembuatan MPS.

Untuk memperbaiki korpus dapat dilakukan dengan memfilter kalimat-kalimat yang berkualitas dari sebuah

korpus parallel[11],menambah kuantitas kalimat pada korpus[12][13], atau perbaikan proses cleaning[14]. Proses cleaning adalah proses pencarian dan perbaikan (penghapusan) kata atau kalimat yang salah ataupun tidak sesuai[15]. Proses cleaning yang disediakan oleh mosesdecoder hanya menghapus kalimat yang terlalu panjang, serta yang kalimat kosong.[16] Salah satu metode pengecekan yang dapat digunakan pada proses cleaning adalah metode dictionary lookup. Dictionary lookup merupakan metode yang melakukan pencarian secara sederhana untuk melihat apakah input kata yang dimasukkan terdapat dalam kamus atau daftar kata yang ada, jika tidak ada maka kata tersebut dianggap sebagai kata yang salah[17].

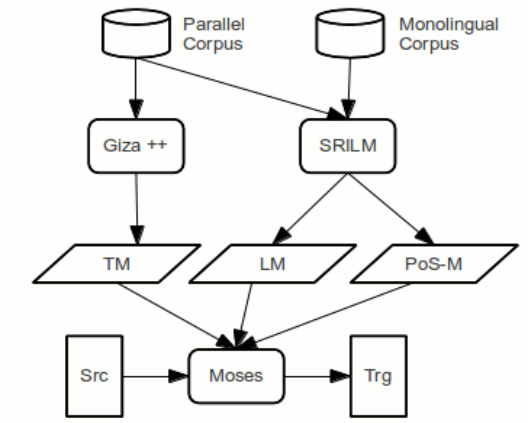
Berdasarkan uraian diatas, maka dilakukan penelitian untuk mengetahui pengaruh penerapan metode dictionary lookup pada proses cleaning korpus terhadap akurasi mesin penerjemah statistik bahasa Indonesia – bahasa Melayu Pontianak.

II. TINJAUAN PUSTAKA

A. *Mesin Penerjemah Statistik*

Mesin penerjemah statistik (MPS) merupakan salah satu jenis mesin penerjemah dengan menggunakan pendekatan statistik.[18] Pendekatan statistik yang digunakan adalah konsep probabilitas. Setiap pasangan kalimat (S,T) akan diberikan sebuah P(T|S) yang diinterpretasikan sebagai distribusi probabilitas dimana sebuah penerjemah akan menghasilkan T dalam bahasa target ketika diberikan S dalam bahasa sumber[3]. Salah satu MPS yang populer saat ini adalah Moses.

Moses meupakan software gratis yang merupakan implementasi dari Mesin Penerjemah Statistik. Moses digunakan untuk melatih model statistik teks terjemahan dari bahasa sumber ke bahasa target. Saat melakukan penerjemahkan bahasa, Moses membutuhkan korpus dalam dua bahasa, bahasa sumber dan bahasa target.



Gambar. 1. Arsitektur mesin pnerjemah statistik Moses

Secara umum, arsitektur mesin penerjemah statistik Moses ditunjukkan pada Gambar 1[19].

Sumber data utama yang dipergunakan adalah *parallel corpus* (korpus paralel) dan *monolingual corpus* (monolingual korpus). Proses *training* terhadap korpus paralel menggunakan GIZA++ menghasilkan *translation model* (TM). Proses *training* terhadap bahasa target pada korpus paralel ditambah dengan monolingual korpus bahasa target menggunakan SRILM menghasilkan *language model* (LM), sedangkan *PoS model* (PoS-M) dihasilkan dari bahasa target pada korpus paralel yang setiap katanya sudah ditandai dengan PoS. *TM, LM dan PoS-M* digunakan untuk menghasilkan *decoder* Moses.[19] Selanjutnya Moses digunakan sebagai mesin penerjemah untuk menghasilkan bahasa target dari *input* kalimat dalam bahasa sumber.[19]

B. *Cleaning*

Data *cleaning*, data *cleansing*, atau data *scrubbing* adalah proses pencarian dan perbaikan (atau penghapusan) data yang salah ataupun kurang tepat dari kumpulan data, table, atau database kemudian menentukan jenis kesalahan apa yang terjadi. Setelah diketahui jenis kesalahan yang terjadi pada data yang salah maka proses berikutnya adalah melakukan penggantian, perbaikan, atau penghapusan data yang salah, hal ini dilakukan untuk meningkatkan kualitas data yang digunakan[15].

C. *Dictionary Lookup*

Dictionary lookup merupakan metode yang melakukan pencarian secara sederhana untuk melihat apakah input kata yang dimasukkan berada dalam kamus atau daftar kata yang ada. Jika tidak ada maka kata tersebut dianggap sebagai kata yang salah. Metode *dictionary lookup* merupakan metode yang sering digunakan dalam menentukan *non-word error*[17].

D. *Dictionary Lookup*

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. Ide utama dibalik ini adalah "semakin dekat terjemahan sebuah mesin dengan terjemahan manusia, maka akan semakin baik[20].

BLEU mengukur *modified n-gram precission score* antara hasil terjemahan otomatis dengan tejemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty*. *Berikut adalah rumus yang digunakan dalam perhitungan nilai BLEU [21]:*

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

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

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

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)

p_n = *jumlah n-gram hasil terjemahan yang sesuai dengan rujukan dibagi jumlah n-gram hasil terjemahan*

III. PEMBAHASAN

A. *Data Penelitian*

Data penelitian yang digunakan adalah Korpus paralel bahasa Indonesia – bahasa Melayu Pontianak yang terdiri atas 9.153 baris kalimat, monolingual korpus bahasa Indonesia yang terdiri atas 82.170 baris kalimat dengan komposisi kumpulan berita online sebanyak 42.382 baris kalimat serta Kamus Besar Bahasa Indonesia 39.788 baris kalimat. Monolingual korpus ini diolah menjadi daftar kata atau kamus yang terdiri dari 97837 kata unik.

B. *Penerapan Metode Dictionary Lookup pada Proses Cleaning*

Monolingual korpus yang telah diolah menjadi kamus digunakan sebagai dasar perbaikan yang akan dilakukan, sementara korpus orisinal (sebelum penerapan metode dictionary lookup) digunakan sebagai korpus yang akan di cleaning. Korpus orisinal yang telah di cleaning, ini nantinya akan disebut dengan korpus clean dic (setelah penerapan metode dictionary lookup) akan digunakan pada proses pengujian. Adapun perbandingan kalimat sebelum dan setelah perbaikan dapat dilihat pada Tabel I.

TABEL I

PERBANDINGAN KALIMAT SEBELUM DAN SETELAH PENERAPAN METODE DICTIONARY LOOKUP PADA PROSES CLEANING

Kalimat Ke-	Kalimat Korpus Setelah Proses *Cleaning* dengan Metode *Dictionary Lookup*	Kalimat Korpus Sebelum Proses *Cleaning* dengan Metode Dictionary Lookup
1	saya berangkat **tadi** malam	saya berangkat **sadi** malam
2	saya **berangkat tadi** m alam	saya **beangkat sadi** malam
3	mereka semua **belum** mandi	mereka semua **belim** mandi
4	**bagi** tubuh korban terbakar habis	**bagia** tubuh korban terbakar habis
5	**apa** yang ada di pikiran **sayi**	**ap** yang ada di pikiran **say**

C. *Implementasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Melayu Pontianak*

Pemodelan bahasa oleh SRILM (Standarford Research Institute Language Modelling) dilakukan pada bahasa target dan menghasilkan tabel model bahasa dengan n-gram data. Model bahasa n-gram memiliki nilai probabilitas dalam bahasa target. Model bahasa dibangun dengan tools SRILM.

Model bahasa akan menghasilkan output dengan format file *.lm. Gambar 2 merupakan tabel model bahasa yang dihasilkan oleh SRILM pada mesin penerjemah statistik Bahasa Indonesia ke bahasa Melayu Pontianak. Pemodelan bahasa oleh SRILM dapat dilihat pada Gambar 2.

```
\data\
ngram 1=8150
ngram 2=35095
\1-grams:
-4.930465       a.yani
-4.930465       aade
-------------------------------------------
\2-grams:
-3.007215       agek itu
-3.007215       agek jadi
```

Gambar. 2. Tabel Model Bahasa Dengan Bahasa Melayu Pontianak Sebagai Bahasa Target

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 dan word alignment. Dokumen-dokumen tersebut terdapat dalam folder "train" yang didalamnya terdapat 4 file yaitu "corpus, giza.id-my, giza.my-id dan model". Pemodelan translasi oleh Giza++ dapat dilihat pada Gambar 3.

```
1       UNK     0
2       ,       4231
3       aku     3070
4       yang    1444
```

Gambar. 3. Dokumen *vocabulary corpus* bahasa Melayu Pontianak

Gambar 3 merupakan isi dari dokumen vocabulary corpus. Angka 1 sampai 4 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 bahasa Indonesia ke Bahasa Melayu Pontianak terdiri dari 7486 token untuk Bahasa Indonesia dan 8231 token untuk korpus bahasa Melayu Pontianak. Dokumen alignment dapat dilihat pada Gambar 4.

```
# Sentence pair (18) source length 7 target length 7 alignment score : 0.000156307
kau tanyakan parit di tempat aku ?
NULL ({ }) kau ({ 1 }) nanyak-ek ({ 2 }) paret ({ 3 }) tang ({ 4 }) tempat ({ 5 }) aku ({ 6 }) ? ({ 7 })
```

Gambar. 4. Dokumen *Alignment* bahasa Indonesia – bahasa Melayu Pontianak

Gambar 4 merupakan dokumen alignment Bahasa Indonesia ke bahasa Meayu Pontianak terdapat tiga baris kalimat. Baris pertama berisi letak kalimat target (18) dalam korpus, panjang kalimat sumber (7), panjang kalimat target (7) dan skor alignment 0.000156307. Baris kedua merupakan bahasa sumber dan baris ketiga merupakan alignment kalimat bahasa target terhadap

kalimat bahasa sumber. Kata “paret” ({ 3 }) memiliki makna bahwa kata “paret” pada kalimat bahasa target, di-align ke kata keenam pada kalimat bahasa sumber yaitu “parit”.

Implementasi dilakukan dengan dua jenis korpus yang berbeda, implementasi ini menghasilkan dua mesin penerjemah yang berbeda. Korpus yang pertama adalah korpus paralel bahasa Indonesia yang nantinya disebut sebagai korpus orisinal. Korpus orisinal ini akan digunakan untuk membuat MPS 1, korpus orisinal ini kemudian di cleaning. Korpus orisinal yang telah di cleaning ini nantinya akan disebut dengan korpus clean dic dan digunakan untuk membuat MPS 2.

D. 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)*. Pada peneltian ini dilakukan dua jenis pengujian otomatis.

1. Pengujian Otomatis 1

Pengujian otomatis 1 dilakukan menggunakan 3 jenis kalimat uji yaitu kalimat uji orisinal dan kalimat uji clean dic yang masing-masing terdiri dari 9157 baris kalimat, serta kalimat uji manual yang terdiri atas 32 baris kalimat hasil terjemahan ahli bahasa. Hasil pengujian otomatis 1 dapat dilihat pada tabel II.

Dari 3 jenis kalimat uji, kalimat uji orisinal dan manual menunjukkan penurunan nilai BLEU masing-masing 7,07% dan 1,5%, sementara pada kalimat uji clean dic terjadi peningkatan nilai BLEU sebesar 2,51%.

TABEL II

PERBANDINGAN NILAI BLEU PENGUJIAN OTOMATIS I

Kalimat Uji	MPS1	MPS2	Selisih (MPS2 – MPS1)
Orisinal	88,48	81,41	-7,07
Clean Dic	82,84	85,35	2.51

2. Pengujian Otomatis 2

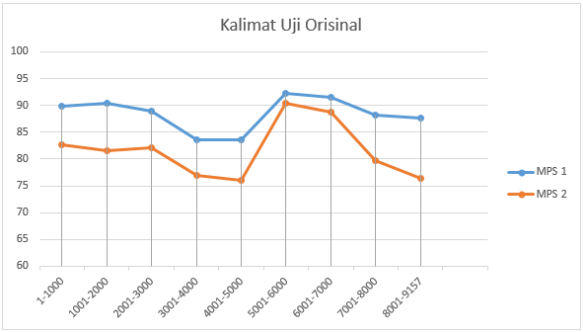
Pengujian otomatis 2 dilakukan dengan 2 buat kalimat uji yaitu kalimat uji orisinal dan clean dic namun berbeda dengan pengujian otomatis 1, pada pengujian otomatis 2 9157 kalimat yang ada di bagi menjadi 9 segmen yang terdiri atas 1000 kalimat dan segmen ke-9 berisi 1157 kalimat. Hasil pengujian otomatis 2 untuk pengujian dengan kalimat uji orisinal dapat dilihat pada Tabel 3 dan Gambar 5.

Pada Tabel III terlihat bahwa nilai BLEU dengan kalimat uji orisinal terjadi penurunan sebesar 6,82%. Terjadinya penurunan disebabkan oleh penggunaan korpus orisinal sebagai korpus referensi perhitungan BLEU.

TABEL III

NILAI BLEU PENGUJIAN OTOMATIS 2 DENGAN KALIMAT UJI ORISINAL

Baris ke-	Orisinal		Selisih (MPS2 – MPS1)
	MPS1	MPS2	
1-1000	89,86	82,61	-7,25
1001-2000	90,33	81,56	-8,77
2001-3000	89,01	82,06	-6,95
3001-4000	83,5	77,02	-6,48
4001-5000	83,65	76,09	-7,56
5001-6000	92,28	90,4	-1,88
6001-7000	91,57	88,8	-2,77
7001-8000	88,16	79,77	-8,39
8001-9157	87,69	76,34	-11,35
Rata-rata	88,45	81,63	-6,82



Gambar. 5. Grafik Nilai BLEU dengan Kalimat Uji Orisinal Pengujian Otomatis 2

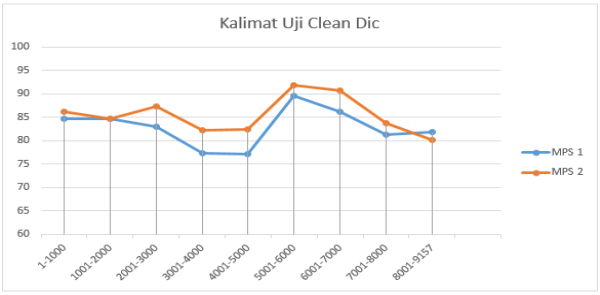
TABEL IV

NILAI BLEU PENGUJIAN OTOMATIS 2 DENGAN KALIMAT UJI CLEAN DIC

Baris ke-	Clean Dic		Selisih (MPS2 – MPS 1)
	MPS1	MPS2	
1-1000	84,67	86,25	1,58
1001-2000	84,7	84,76	0,06
2001-3000	82,98	87,38	4,4
3001-4000	77,34	82,34	5
4001-5000	77,12	82,49	5,37
5001-6000	89,66	91,87	2,21
6001-7000	86,27	90,81	4,54
7001-8000	81,32	83,68	2,36
8001-9157	81,78	80,12	-1,66
Rata-rata	82,87	85,52	2,65

Gambar 5 menampilkan perbandingan nilai BLEU antara MPS 1 dan MPS 2, dari gambar tersebut terlihat bahwa baris ke 5001-6000 serta baris ke 6001-7000, kedua MPS memiliki Nilai BLEU yang tidak terlalu jauh berbeda, hal ini menunjukkan bahwa hasil terjemahan

otomatis pada kedua baris tersebut memiliki banyak kemiripan, sehingga menyebabkan Nilai BLEU yang tidak terlalu jauh berbeda.



Gambar. 6. Grafik Nilai BLEU dengan Kalimat Uji Clean Dic Pengujian Otomatis 2

Tabel IV menampilkan hasil pengujian otomatis 2 dengan kalimat uji clean dic, terlihat bahwa pengujian dengan kalimat uji clean dic terjadi peningkatan nilai BLEU sebesar 2,65%.

Dari Gambar 6 terlihat pada Baris ke 1001-2000 nilai BLUE MPS 1 dan MPS 2 berada di titik yang sama, yang berarti terjemahan otomatis MPS 1 dan MPS 2 menghasilkan kalimat yang sama sehingga menghasilkan nilai akurasi yang sama. Sementara pada baris ke 8001-9157 nilai BLUE MPS 1 lebih besar MPS 2, hal ini di sebabkan oleh hasil terjemahan otomatis MPS 1 lebih mirip dengan kalimat referensi jika dibandingkan dengan hasil terjemahan otomatis MPS 2.

IV. KESIMPULAN

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

1. Mesin penerjemah statistik dapat diimplementasikan untuk menterjemahkan bahasa Indonesia ke bahasa Melayu Pontianak.
2. Terjadi penurunan nilai akurasi MPS dengan kalimat uji (manual) yaitu sebesar 1,5%.
3. Terjadi penurunan nilai akurasi MPS dengan kalimat uji (orisinal) yaitu sebesar 6,94%.
4. Terjadi kenaikan nilai akurasi MPS dengan kalimat uji (clean dic) yaitu sebesar 2,65%.
5. Berdasarkan hasil penelitian, penerapan metode *dictionary lookup* pada proses cleaning korpus dapat mempengaruhi akurasi terjemahan MPS bahasa Indonesia ke bahasa Melayu Pontianak. Pada penelitian ini, penerapan metode *dictionary lookup* pada proses *cleaning* mengakibatkan penurunan akurasi mesin penerjemah statistik.

REFERENSI

[1] Badan Pusat Statistik 2011. Kewarganegaraan, Suku Bangsa, Agama, dan Bahasa Sehari-hari Penduduk Indonesia Hasil Sensus Penduduk 2010.

[2] Moseley, Christopher. 2010. Atlas of the World's Languages in Danger of Disappearing, UNESCO Publishing Vol 3.

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

[4] Apriani, T., Pengaruh Kuantitas Korpus Terhadap Akurasi Mesin Penerjemah Statistik Bahasa Bugis Wajo ke Bahasa Indonesia, Jurnal Sistem dan Teknologi Informasi (JustIN), Vol. 1, No. 1, hal. 1-6, 2016.

[5] Yohanes, B.W., Robert, T., dan Nugroho, S., Sistem Penerjemah Bahasa Jawa-Aksara Jawa Berbasis Finite State Automata, Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), Vol. 6, No. 2, hal. 127-132, Mei 2017.

[6] Nugroho, R.A., Adji, T.B. & Hantono, B.S., Penerjemahan Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Berbasis Frasa, Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015), 2015, hal. 51.

[7] Mulyana. 2018. Algoritma Pembagian Frasa dalam Kalimat Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Bugis Wajok. Jurnal Sistem dan Teknologi Informasi (JustIN), Vol. 6, No. 2, hal. 39-48, 2018.

[8] 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.

[9] Indrayana, Danny. 2016. Meningkatkan Akurasi Pada Mesin Penerjemah Bahasa Indonesia Ke Bahasa Melayu Pontianak Dengan Part Of Speech. Pontianak, JEPIN Vol 1 No 1 2016.

[10] Sujaini, H., Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke Bahasa Melayu Pontianak, Jurnal Teknik Elektro (ELKHA), Vol. 6, No. 2, hal. 38-44, Oktober 2014.

[11] Sujaini, Herry. dan Arif B.P.N. 2015. Strategi Memperbaiki Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik. Jakarta, Seminar Nasional Teknologi Informasi XI.

[12] Yıldız, E., Tantuğ, A.C., & Diri, B., The Effect of Parallel Corpus Quality vs Size in English-to-Turkish SMT, Sixth International Conference on Web services & Semantic Technology (WeST 2014), 2014, hal. 21-30.

[13] Maheshwar, S. & Sharma, H., Improvements in Corpus Quality for Statistical Machine Translation, IJSRD - International Journal for Scientific Research & Development, Vol. 2, No, 5, hal. 2321-0613, 2014.

[14] *Xu, Hainan and Koehn, Philipp* (2017): Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[15] Devi. Sapna, dan Kalia, Arvind. 2015. Study of Data Cleaning & Comparison of Data Cleaning Tools. IJCSMC, Vol. 4, Issue. 3, March 2015.

[16] Koehn, Philipp. 2016. MOSES Statistical Machine Translation User Manual dan Code Guide. The University of Edinburgh.

[17] Maghfira, Tusty Nadia. 2017. Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance. Malang, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 1, No. 6, Juni 2017

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

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

[20] Manindra, Soni. 2016. Perbaikan Probabilitas Lexical Model untuk Meningkatkan Akurasi Mesin Penerjemah Statistik. Pontianak, JEPIN Vol 2 No 1 2016.

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