TELKOMNIKA, Vol.12, No.3, September 2014, pp. 581~588
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v12i3.79

A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation

Herry Sujaini, Kuspriyanto, Arry Akhmad Arman, Ayu Purwarianti
School of Electrical Engineering and Informatics
Bandung Institute of Technology, Jl. Ganesha No. 10, Bandung, Indonesia
e-mail : herry_sujaini@yahoo.com

Abstract

Part of speech (PoS) is one of the features that can be used to improve the quality of statistical-based machine translation. Typically, language PoS determined based on the grammar of the language or adopted from other languages PoS. This work aims to formulate a model to develop PoS as linguistic factors to improve the quality of machine translation automatically. The model is based on word similarity approach, where we performed word clustering on corpus. The result of word clustering will be defined as PoS set obtained for a given language. The PoS sets resulted by the word clustering were compared to the manually defined PoS set in a machine translation (MT) experiment, the MT experiment employed English as the source language and Indonesian as the target language.

Keywords*: method, part-of-speech, statistical machine translation, moses, word similarity*

1. Introduction

The dream of automatically translating documents between two languages is one of the oldest pursuits of artificial intelligence research. Now, armed with vast amounts of example translations and powerful computers, we can witness significant progress toward achieving that dream. Statistical analysis of bilingual parallel corpora allow for the automatic construction of machine translation systems. Already, for some language pairs, statistical systems are the best machine translation systems currently available.

Statistical Machine Translation is corpus-based and consequently requires a parallel corpus to learn a model [1],[2]. Parallel corpora are different from normal text corpora in that they are not just a collection of texts, but are bilingual or multilingual and structured so that every sentence is linked to its translations.

Some works have shown that the translation quality can be increased by using additional features such as lemma, part of speech (PoS), gender and others. In their research, Koehn and Hoang [3] explained that by adding a factor of part-of-speech in English-German translator system, the quality of the translation was increased from 18.04% to 18.15%. They also showed that by using morphological factors and part-of-speech, the English-Spanish translator system quality was increased from 23.41% to 24.25%.

Youssef et al. [4] examined the factors on adding part-of-speech on statistical translation system for English-Arabic. Research results showed that the addition of a factor of part-of-speech can improve the quality of translation from 0.6095% to 0.6394%. Razavian and Vogel [5] examined the factors on adding to the statistics based interpreter systems, for English-Iraqi interpreter system, the quality of the translation was improved from 15.62% to 16.41%; for the Spanish-English translator system, the quality of the translation was improved from 32.53% to 32.84%; and for Arabic-English translator system, the quality of the translation was improved from 41.70% to 42.74%.

For English-Indonesian, Sujaini et al. [6] conducted a study of the addition of PoS factors based on a statistical translator system factors. The results of these studies indicated that the PoS factor increased the quality of the English-Indonesian translation of 2%, from 31.26% to 33.26%.

Grammatically, words can be divided into two categories: open class and closed class. Open class is a class category which number of words always increases over time, while closed class is a class category whose words are fixed. Grammatically different categories of words, commonly called Part of Speech [1].

Received February 9, 2014; Revised May 14, 2014; Accepted June 10, 2014

PoS functions for natural language processing is to provide some information about a word and the words around it. This applies to general category (noun vs. verb) as well as to more specialized. For example, a set of tags to distinguish between possessive pronouns (my, your, his, her, it) and personal pronouns (I, you, he, she) [7]. While PoS tagging is the process of labeling each word in a sentence with the appropriate tag from a set of PoS [8].

In general, a set of tags encode both the classification of the target feature, tell the user useful information about the grammatical word classes, and predictive features, encoding feature that would be useful in predicting the behavior of other words in the context. Both tasks should overlap, but they are not always identical [9].

PoS generally refers to a class of words used in a particular language and each language has different PoS categories. Classes for the Greek word has been defined by Dionysius Thrax in 100 BC which consists of eight classes of words, namely: noun, verb, pronoun, preposition, adverb, conjunction, particle, and the article. Indonesian class words divided into verbs, adjectives, noun, word numbers, pronouns, adverbs, conjunction, demonstrative, interjection, interogative, articulatory, preposition, and reduplication [10].

PoS for various languages have been developed for the computerization, one of which is the Penn Treebank by LINC Laboratory, Computer and Information Science, University of Pennsylvania [11]. They divided English words into 48 PoS. Previously, Francis [12] divided the English words used for 87 PoS in the Brown corpus. Additionally Garside et al. [13] divided the English words into a 146 PoS for C7 tagset.

Various sets of Indonesia PoS has been used in the research field of natural language processing, including through the PAN Localization Project, specifically for PoS Indonesia has been developed specifically to be translated into English in 2009 [14], the PoS based on the Penn Treebank POS tag set [11] consists of 29 PoS tags. Pisceldo et al. [15] defined 37 tags for Indonesia. Wicaksono and Purwarianti [16],[17] in their work using 35 tag tagset modification results produced by Adriani, [14] and Pisceldo et al. [15]. Lastly, Larasati et al. [18] uses only 19 tags in their work.

Several other works also showed variations in the amount tagset used in a variety of languages. For the Arabic, Hajic et al. [19], using 21 tags in the Arabic Treebank data and tools. Brants et al. [20] used 54 tags to build the TIGER treebank in German. Simov et al. [21] used 54 tags to build a corpus of Bulgarian. Csendes et al. [22] used 43 tags to build a treebank Szeged in Hungarian. Civit and M.A. Mart [23] used 47 tags to build a Spanish treebank in Spanish. For developed part-of-speech tagger, Avontuur et al. [24] used 25 tags for Dutch, Singha et al. [25] used 97 tags for Manipuri, Neunerdt et al. [26] used 54 tags for German.

In this article, we propose a method to determine a set of PoS automatically by using word similarity approach for Indonesian. The contributions of this research are a novel method for developing a language PoS automatically and an alternative Indonesian Sets PoS to be used in statistical machine translation.

2. Developing Part-of-Speech Set Method

The input of this method is mono corpus that contains a collection of sentences. The output of this method is a PoS set. Models to determine computationally PoS Set consists of 4 (four) steps of the process, namely: computing word similarity, word clustering, visualization cluster, and PoS categorization as shown in Figure 1.

Step 1: Computing word similarity

At this step, mono corpus processed using Extended Word Similarity Based (EWSB) algorithm which has been developed and presented by Sujaini et al. [23]. The mutual information between w1 and w2 is defined as :

$$I(t, w_1, r, w_2) = log\ \frac{Cnt(t,w_1,r,w_2).Cnt(t,*,r,*)}{Cnt(t,w_1,r,*).Cnt(t,*,r,w_2)} \quad (1)$$

and the word similarity between w1 and w2 is defined as :

$$sim(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)}[I(w_1,r,w)+I(w_2,r,w)]}{\sum_{(r,w)\in T(w_1)}[I(w_1,r,w)]+\sum_{(t,r,w)\in T(w_2)}[I(w_2,r,w)]} \quad (2)$$

The output of this step is a list of word pairs along with the similarity value.

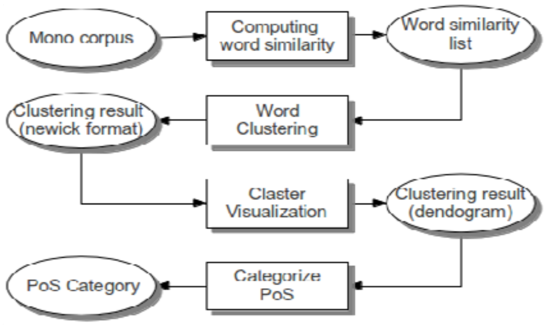

Figure 1. Block Diagram of Determination Part of Speech Set Model

Step 2 : Word clustering

Word clustering process at this step using Agglomerative and customized approach to get the history of clustering in Newick format. Adopted in 1986, Newick format (Newick notation) is a way to represent graph-theoretical trees by using parentheses and commas [24].

Agglomerative algorithms which have been adjusted to obtain the results of the Newick format is as follows :

1. Initialize each unique word (token) as a cluster
2. Calculate the similarity between two clusters
3. Sort ranking between all pairs of clusters based on similarity, then combine the two top clusters
4. Add clusters are combined in Newick format
5. Stop until it reaches a single cluster, if not, return to step 2.

To calculate the similarity between two clusters in step 2, we used the formula in equation (3) [23]:

$$sim(C_1, C_2) = \frac{1}{N_1 * N_2} \sum_{w_1 \in C_1} \sum_{w_2 \in C_2} sim\,(w_1, w_2) + \frac{\lambda}{N_1 + N_2} \qquad (3)$$

where N_1 and N_2 denote the numbers of words in the classes, C1 and C2 , respectively. Jeff et al. [25] added the term $\frac{\lambda}{N_1+N_2}$ to the class similarity computation, tending to have a higher priority for smaller classes to be merged. In our experiments we set $\lambda \approx 0$.

Step 3 : Cluster Visualization

Results of hierarchical clustering illustrated with a dendogram, where the dendrogram is a curve that describes the cluster grouping. At this stage, Newick format generated in the previous stage be used as input to obtain a visualization cluster dendogram. We use “Dendroscope” to describe clusters that can be accessed at http://www-ab2.informatik.uni-tuebingen.de/software/dendroscope/.

Step 4 : PoS categorization

The last process of this model is the PoS categorization manually processed by the dendogram visualization. The output of this process is the grouping and naming PoS.

3. Determining Indonesian PoS Set

The purpose of this experiment is to determine the set of Indonesian PoS computationally through computational results. In this experiment, we use a 171K sentences Indonesian corpus which has 3,4 M tokens (114 K unique tokens).

We have experimented to determine the set of PoS with two (2) ways, namely clustering words with each category separately conducted PoS and word clustering as a whole. In separate ways, we classify certain words that fit the category. PoS categories used are :

verbs, nouns, adjectives, numerals, adverbs, conjunctions and other categories. We have chosen some appropriate and varies words from a list of unique token (uni-gram) for each category. As an example, we computed the words similarity against words in verbs category, the results of the second step from computational process produces an output word similarity list (20 highest scores) can be seen in Table 1.

Tabel 1. Word Similarity Scores for Verbs Category

No	Word 1	Word 2	Word Similarity Score
1	*membaik (getting better)*	*melemah (weakened)*	0.1257617113
2	*menguat (strengthened)*	*melemah (weakened)*	0.0984328977
3	*membaik (getting better)*	*menguat (strengthened)*	0.0810526508
4	*dilakukan (do)*	*dilaksanakan (implemented)*	0.0801780730
5	*mengatakan (say)*	*menyatakan (state)*	0.0738027565
6	*dilakukan (do)*	*digunakan (used)*	0.0725432867
7	*memberikan (provide)*	*memberi (give)*	0.0692650245
8	*berdiri (stand)*	*duduk (sit)*	0.0629038361
9	*dibuat (made)*	*dilaksanakan (implemented)*	0.0606981494
10	*membaik (getting better)*	*memburuk (deteriorate)*	0.0597877822
11	*diatur (regulated)*	*dilaksanakan (implemented)*	0.0562082758
12	*digunakan (used)*	*dibuat (made)*	0.0550282651
13	*dilaksanakan (implemented)*	*berkembang (thrive)*	0.0543281608
14	*digunakan (used)*	*ditemukan (found)*	0.0517041060
15	*digunakan (used)*	*dilaksanakan (implemented)*	0.0505865651
16	*diatur (regulated)*	*dibuat (made)*	0.0496155258
17	*dilakukan (do)*	*dibuat (made)*	0.0486961755
18	*duduk (sit)*	*tidur (sleep)*	0.0473421617
19	*dilakukan (do)*	*diberikan (given)*	0.0457435151
20	*bergerak (move)*	*berkembang (thrive)*	0.0446583898

From the results of the above process, we have processed the next step, ie grouping of words to obtain the cluster results in Newick format, word similarity clustering results for verb PoS categories are:

(((((((((((diberikan),((ditemukan),(((dibuat),(((dilakukan),(dilaksanakan)),(digunakan))),(diatur)))),((bergerak),(berkembang))),(((bermain),(bertemu)),(((berdiri),(duduk)),((makan),(tidur))))),((mandi),(minum))),(terbawa)),((((memburuk),((menguat),((melemah),(membaik)))),((mengecil),(melambat))),((membesar),(memudar)))),(terpakai)),(terdengar)),((ialah),(((((adalah),(merupakan)),(((((memberikan),(mendapatkan)),(mempunyai)),(menggunakan)),((membuat),(melakukan))))),((((mengatakan),(menyatakan)),(melihat)),(merasa))),(mengalami)))),(((((ingin),(((akan),(dapat)),(harus))),(sudah)),(boleh)),(mesti)))

Furthermore, we have a PoS verbs visualization with Dendroscope software, visualization is obtained as shown in Figure 2.

Figure 2. Dendogram Visualization of Verbs Category Clustering Results

Next, we determined the PoS for verbs categories based by dendogram visualization, verbs which are already in use PoS based grammar is: VBT (transitive verb), VBI (intransitive verb) and MD (modal). We can see that the MD (*akan, dapat, harus*, etc.) and VBT form a separate group (*membuat, melakukan, merupakan*, etc.), While the VBI dispersed into several groups. VBI spread with groups of passive verbs (*dibuat, digunakan, dilaksanakan*, etc.), Which has the meaning of the verb "to be" (*melemah, membaik, mengecil*, etc.), and other VBI scattered. Based on the foregoing, we provided recommendations for verbs PoS set on the results of computational such as Table 2. In the same way, we also have to experiment with other types of words in order to obtain a set PoS for Indonesian as in Table 3.

Tabel 2. PoS Set Recommended for Verbs Category

No	Words Examples	PoS Tag	Description
1	*dapat, akan, ingin, sudah*	MD	Modal
2	*mengatakan , melakukan membuat , melihat*	VBT	Transitive
3	*duduk, minum, mandi, berkembang, terpakai*	VBI	Intransitive
4	*digunakan, dibuat, diatur, dilaksanakan*	VBI1	passive verbs
5	*melemah, membaik, mengecil, memudar*	VBI2	meaning of the verb "to be"

Table 3. Indonesian PoS Set Recommended by Computational Based

No	Tag	Description	Word Examples
1	OP	Opening parenthesis	({ [
2	CP	Closing parenthesis	) }]
3	GM	Slash	/
4	;	Semicolon	;
5	:	Colon	:
6	“	Quotation	“ ’
7	.	Sentence terminator	. ? !
8	,	Comma	,
9	-	Dash	-
10	...	Ellipsis	...
11	JJ1	Adjectives 1	*panjang, kuat, indah, besar*
12	JJ2	Adjectives 2	*genap, buntu, negatif*
13	RB	Adverbs	*sekedar, hampir, tidak*
14	RB1	Adverbs 1	*sangat, amat, cukup, paling*
15	NN	Common Noun	*mobil, air, negara*
16	NNP	Proper nouns	*tvri, jokowi, persib*
17	NNG	Genitive nouns	*bukunya, hatinya*
18	VBI	Intransitive Verb	*duduk, pergi, makan*
19	VBI1	Intransitive Verb 1	*dibuat, diambil*
20	VBI2	Intransitive Verb 2	*mengecil, menguat*
21	VBT	Transitive Verb	*membeli, memukul*
22	IN	Preposition	*di, ke, dari*
23	MD	Modal	*akan, harus*
24	CC	Coor - conjunction	*dan, atau, ketika, jika*
25	DT	Determiner	*ini, itu*
26	UH	Interjections	*wah, aduh, oi*
27	CDO	Ordinal numerals	*pertama, kedua*
28	CDC	Collective numerals	*berdua, bertiga*
29	CDP	Primary numerals	1, 2, 3
30	CDP1	Primary numerals 1	*satu, dua*
31	CDP2	Primary numerals 2	*puluh, ribu, juta*
32	CDP3	Primary numerals 3	1990, 2001, 2013
33	CDI	Irregular numerals	*beberapa*
34	PRP	Personal pronoun	*saya, kamu*
35	WP	WH-pronouns	*apa, siapa*
36	PRN	Number pronouns	*kedua-duanya*
37	PRL+	Locative Proper nouns/pronouns	*sini, situ, Jakarta, Bali*
38	SYM	Symbols	@#$%^&
39	RP	Particles	*pun, kah*
40	FW	Foreign words	foreign, Word
41	ART	Articles	*sang, si, para*
42	COP	Copula	*adalah, bukan, merupakan*

4. Experiments on SMT

The purpose of this experiment is to compare the accuracy of the translation system that uses PoS computational results compared with translation system with PoS determined by grammar based. In addition, we also compared the results of the translation without PoS features. For PoS determined based grammar, in this work used the Wicaksono's PoS and hereinafter called Grammar PoS

We used several instruments in this experiment, Moses [1] as machine translators, SRILM [26] to building language and PoS models, Giza++ [27] for word alignment process, and Grammar Postagger for PoS tagging. Furthermore, we use the BLEU method [28] for scoring the translation results. We used a parallel corpus for training the translation model and mono corpus for training the language model. We used "Identic" Parallel corpus [29] that contains 27K sentence pairs of English-Indonesian. While mono corpus used is the same as that used in the experiments at 170 K sentence clustering.

We tested the factor-based statistical machine translation by marking the PoS (postagging) against English-Indonesian parallel corpus. Test sentences totaling 1,500 sentences consisting of 5 test groups, each consisting of 300 sentences with word length 10, 15, 20, 25 and 30 (reference sentence).

The BLEU score of the experiment results of conducted in MPS can be seen in Table 4. The increase in the BLEU score of the translation results using computational PoS and Grammar PoS of the translation results without using PoS illustrated in Figure 3.

From Table 4. we can see that the translation accuracy using Grammar PoS better than without PoS. While the use of PoS of computing results can also improve the accuracy of the translation results as compared to the use of Grammar PoS.

The increase in accuracy due to the use of PoS features better on short sentences. The best enhancement to the translation by computing PoS of 8.89% on a corpus containing sentences with 10 words long, while the lowest increase of 1.57% occurs at the E corpus containing sentences with 30 words long. When compared with the use of Grammar PoS, SMT with computational PoS results to increase average accuracy of 4.13%. The increase in average accuracy of the translation use grammar PoS on without PoS is 2.23%.

Tabel 4. BLEU score of Grammar and Computational PoS

Corpus	Base (no PoS)	Grammar PoS	Computational PoS
A	56.93	57.90	61.99
B	47.86	49.06	51.38
C	44.98	46.94	48.56
D	43.52	44.92	46.56
E	55.39	55.44	56.26
Average	49.74	50.85	52.95

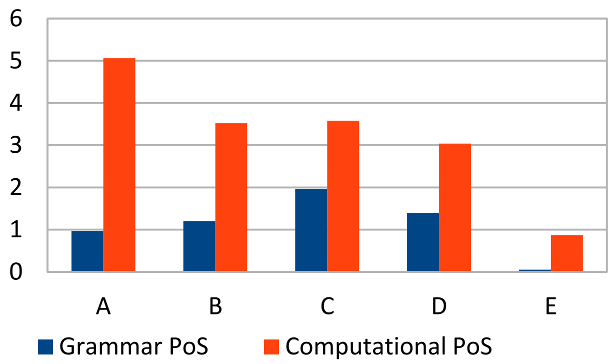

Figure 3. Graph Translation Accuracy Againts Without PoS

The BLEU score examples of each group form source sentences in English, a reference translation, translation with grammar PoS and computing PoS has increased, fixed, and decreased accuracy can be seen in Table 5.

Based on the experimental results, we can conclude that the use of sets of computationally generated PoS can reduce weaknesses determined PoS set based grammar so as to improve the quality of statistical machine translation. This is because the determination of grammar PoS is generally based on the function and meaning, and it does not guarantee similarity of distribution of words in a sentence to the words in the same category PoS.

Tabel 5. BLEU score for Grammar and Computational PoS Used

No		Sentences	BLEU Score (%)
1	Input	did i not just say i 'm saving the film ?	
	Ref	bukan kah saya sudah bilang untuk menghemat film nya ?	
	Grammar	apa kah saya tidak hanya bilang aku untuk menghemat film nya ?	37.70
	Komp	bukan kah saya sudah bilang untuk menghemat film nya ?	100.00
2	Input	the challenges to meet the investment needs will come from the government itself	
	Ref	tantangan pemenuhan kebutuhan investasi itu justru berasal dari pemerintah sendiri	
	Grammar	tantangan pemenuhan kebutuhan investasi itu akan datang dari pemerintah sendiri	52.54
	Komp	tantangan pemenuhan kebutuhan investasi itu akan datang dari pemerintah sendiri	52.54
3	Input	proven that all policies are aimed for successing liberalization implementation	
	Ref	terbukti bahwa segala kebijakan ditujukan untuk menyukseskan berlaku nya liberalisasi	
	Grammar	terbukti bahwa segala kebijakan ditujukan untuk menyukseskan berlaku nya liberalisasi	100.00
	Komp	terbukti bahwa semua kebijakan ditujukan untuk menyukseskan berlaku nya liberalisasi	70.71

5. Conclusion

Models to determine computationally PoS Set consists of 4 (four) steps of the process, namely: computing word similarity, word clustering, visualization cluster, and PoS categorization. From experiment result, we recommended 42 tags Indonesian PoS for machine translation. The average of increase in accuracy of the translation use grammar PoS on without PoS is 2.23%. The use of PoS computing results can improve the accuracy of 6.45% compared to a translation without PoS. When compared with the use of PoS grammar, usage PoS computing results can improve the accuracy of about 4.13%. Accuracy of PoS use both grammar PoS and PoS TB results are low at long sentences (30 words).

References

[1] Koehn P. Statistical Machine Translation. New York: Cambridge University Press. 2010.

[2] Peng L. A Survey of Machine Translation Methods. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(12): 7125-7130.

[3] Koehn P, Hoang H. *Factored Translation Models*. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague. 2007: 868-876.

[4] Youssef I, Sakr M, Kouta M. Linguistic Factors in Statistical Machine Translation Involving Arabic Language. *IJCSNS International Journal of Computer Science and Network Security*. 2009; 9(11): 154-159.

[5] Razavian, N.S, Vogel S. *Fixed Length Word Suffix for Factored Statistical Machine Translation*. Proceedings of the ACL 2010 Conference Short Papers. Uppsala. 2010: 147-150.

[6] Sujaini H, Kuspriyanto, Arman A.A, and Purwarianti A. *Pengaruh Part-Of-Speech pada Mesin Penerjemah Bahasa Inggris-Indonesia Berbasis Factored Translation Model*, SNATI-2012. Yogyakarta. 2012: H77-H82.

[7] Jurafsky D, Martin H. Speech and Language Processing, New Jersey: Parson International Edition. 2009.

[8] Raja F, Tasharofi S, Oroumchian F. *Statistical POS Tagging Experiments on Persian Text, Proceedings of the Second Workshop on Computational Approaches to Arabic Script-based Languages*. California. 2007 : 128-133.

[9] Manning C.D, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press. 1999.
[10] Waridah E. EYD dan Seputar Kebahasa-Indonesiaan. Jakarta: Kawan Pustaka. 2008.
[11] Marcus M.P, Marcinkiewicz M.A, Santoroni B. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics - Special issue on using large corpora: II*. 1993; 19(2): 313-330.
[12] Francis W.N. A Tagged Corpus – Problems and Prospects. In: Greenbaum, S., Leech, G., and Svartvik, J. *Editors*. Studies in English Linguistics for Randolph Quirk. London: Longman; 1979: 192-209.
[13] Garside R, Leech G, McEnery A. Corpus Annotation : Linguistic Information from Computer Text Corpora. London: Longman. 1997.
[14] Adriani M, Riza H. *Research Report on Local Language Computing: Development of Indonesian Language Resources and Translation System.* PAN Localization, 102042 . 2008.
[15] Pisceldo F. Adriani M, Manurung R. *Probabilistic Part of Speech Tagging for Bahasa Indonesia*. Third International Wokshop on Malay and Indonesian Language Engineering. Singapore. 2009.
[16] Wicaksono A.F, Purwarianti A. *HMM Based Part-of-Speech Tagger for Bahasa Indonesia. The 4th International Malindo Wokshop*. Jakarta. 2010: 94-100.
[17] Purwarianti A, Saelan A, Afif I, Ferdian F, Wicaksono A.F. Natural Language Undertanding Tools, with Low Language Resource in Building Automatic Indonesian Mind Map Generator". *International Journal on Electrical Engineering and Informatics*. 2013: 5(3): 256-269.
[18] Larasati S.D, Kuboň V, Zeman D, Indonesian Morphology Tool (MorphInd): *Towards an Indonesian Corpus. SFCM 2011. Springer CCIS proceedings of the Workshop on Systems and Frameworks for Computational Morphology*. Zurich. 2011: 119-129.
[19] Hajic O, Smrz P, Zemanek J.S, Beska E. *Prague Arabic dependency treebank: Development in Data and Tools*. Network for Euro-Mediterranean Language Resources (NEMLAR). Cairo. 2004.
[20] Brants S, Dipper S, Hansen S, Lezius W, Smith G. *The TIGER Treebank*. Workshop on Treebanks and Linguistic Theories. Sozopol. 2002: 24-41.
[21] Simov K, Osenova P, Kolkovska S, Balabanova E, Doikoff D, Ivanova K, Simov A, Kouylekov M. *Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. European Language Resources Association LREC*. Canary Islands. 2002.
[22] Csendes D, Csirik J, Gyimóthy T, Kocsor A. *The Szeged Treebank*. Proceedings of the 8th International Conference on Text, Speech and Dialogue. Karlovy Vary. 2005: 123-131.
[23] Civit M, Mart M.A. Building cast3lb: A Spanish treebank. *Research on Language & Computation*. 2004: 2(4): 549–574.
[24] Avontuur T, Balemans I, Elshof L, Noord N.V, Zaanen M.V, Developing a part-of-speech tagger for Dutch tweets. *Computational Linguistics in the Netherlands Journal*. 2012: 2: 34–51.
[25] Singha K.R, Purkayastha B.S, Singha K.D, Part of Speech Tagging in Manipuri: A Rule-based Approach. *International Journal of Computer Applications*. 2012: 51(14): 31-36.
[26] Neunerdt M, Reyer M, Mathar R. A POS Tagger for Social Media Texts trained on Web Comments. *Research journal on Computer science and computer engineering with applications*. 2013: 1(48): 61-68.
[27] Sujaini H, Kuspriyanto, Arman A.A, Purwarianti A, Extended Word Similarity Based Clustering on Unsupervised PoS Induction to Improve English-Indonesian Statistical Machine Translation. *16th ORIENTAL COCOSDA/CASLRE-2013*. Gurgaon. 2013: 47-48.
[28] Felsenstein J. Inferring Phylogenies. Sunderland: Sinauer Associates, Inc. 2004.
[29] Jeff M.A, Matsoukas S, Schwartz R. *Improving Low-Resource Statistical Machine Translation with a Novel Semantic Word Clustering Algorithm*. Proceedings of the MT Summit XIII. Xiamen. 2011: 352-359.
[30] Stolcke A, Zheng J, Wang W, Abrash V. *SRILM at Sixteen: Update and Outlook. IEEE Automatic Speech Recognition and Understanding Workshop*. Waikoloa. 2011.
[31] Och F.J, and Ney H. A Systematic Comparison Of Various Statistical Alignment Models. *Journal Computational Linguistics*. 2003: 29(1): 19-51.
[32] Papineni K, Roukos S, Ward T, Zhu W.J. *BLEU: A Method For Automatic Evaluation of Machine Translation*. ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 311-318.
[33] Larasati S.D, IDENTIC Corpus : *Morphologically Enriched Indonesian-English Parallel Corpus*. LREC, Europan Language Resources Association ELRA. 2012: 902-906.