Analysis of Extended Word Similarity Clustering based Algorithm on Cognate Language

Arif B. Putra N
Computer Science/Informatics Department,
Engineering Faculty, Universitas Tanjungpura
Pontianak, Indonesia

Herry Sujaini
Computer Science/Informatics Department,
Engineering Faculty, Universitas Tanjungpura
Pontianak, Indonesia

***Abstract* — Extended Word Similarity Based (EWSB) Clustering is a clustering algorithm based on the similarity value of word that derived from the results of computation from a corpus. One of the benefits of clustering results by the algorithm is to improve the accuracy of the translation of a statistical machine translation. From the results of previous studies, the algorithm can improve the accuracy of machine translation of English into Indonesian, wherein the algorithm is applied to the Indonesian language as the target language.**

English and Indonesian is a language that does not cognate. This paper discusses the results of the use of EWSB algorithms for statistical machine translation of the cognates for Indonesian to Malay Pontianak, where the algorithm applied to the Malay Pontianak as the target language. Research obtains the results that the EWSB algorithm effective if used in Pontianak Malay language as the language of the target with an increase in the value of BLEU by 2,48%.

Keywords—Extended Word Similarity Based, Statistical Machine Translation, Clustering.

I. Introduction

This Machine translation (MT) is a machine that can make the process of automatically translating from one language to another. MT has practical utility because it can help people to communicate with each other have different languages. This issue becomes even more important at this time of globalization, when the translation manually by humans who have limited resources and expensive, MT has the potential to increase efficiency. Additionally, communication media such as email, SMS, BBM, social media and video conferencing, today has become increasingly varied and almost instantaneous and has become an integral part of human activity.

One approach to machine translation is to use a statistical approach that uses the concept of probability, usually called statistical machine translation (SMT). For each sentence pair (s, t), is given a P (t | s) shall be interpreted as a probability distribution where SMT will generate t in the target language when given s in the source language. SMT has been widely used in various applications, such as common multi-language translator like Google translator, Bing translator and others.

Some studies MT in several languages, has shown that the accuracy of the MT will be better with the addition of features such as lemma, class of words (part-of-speech/PoS), gender and others. The features of these languages can be induced in the process of "training" or added in parallel corpus, as linguistic information.

Koehn and H. Hoang [1] explained that by adding the Post factor in English-German translation system (751 088 sentence) can improve the accuracy of the translation from 18.04% to 18.15%. Whereas in English-Spanish translation system (40,000 words) generated 23.41% without adding factor, increased to 24.25% with the added factor of morphology and PoS. Youssef et al. [2] conducted a study of the addition of PoS factor on statistics-based translation system, for English-Arabic translator system (68 685 words). Research results show that the addition of PoS factors can improve the accuracy of the translation from 60.95% to 63.94%. In the study, Nedjo1 A.T. and Degen, H. [3] added factor PoS system Oromo-English translator (13 633 words) can improve the accuracy of the translation from 2.56% to 2.88%. Razavian et al. [4] conducted a study of the factors adding to the statistics-based translation system, for system-Iraqi translator English (650,000 words) can improve the accuracy of the translation from 15.62% to 16.41%, for a Spanish-English translator system (1,200,000 sentence) can improve the accuracy of the translation from 32.53% to 32.84%, and for the system of Arabic-English translators (3,800,000 words) can improve the accuracy of the translation from 41.70% to 42.74%.

For Indonesian, research conducted by Sujaini et al [5] showed that the use of Extended Word Similarity Based (EWSB) Algorithm Clustering on statistical machine translation English-Indonesia can improve the accuracy of 2.07%, but these studies have not shown the influence EWSB against the cognate language translation. While it has been proven that EWSB worked well for machine translation with cognate language, but it is not known the extent to which performance of EWSB if used in machine translation with cognate languages.

In this study, we use the Indonesian language as the source language and Malay Pontianak as the target language. Malay Pontianak language is one of the local languages which are in the province of West Kalimantan, Indonesia. The language spoken by the Malay people in the city of Pontianak In most vocabulary, Malay Pontianak is very similar to Indonesian, because it is rooted in Indonesian language.

II. PART OF SPEECH INDUCTION WITHOUT SUPERVISION

Translator machine is a machine that does the translation automatically, where a computer takes over all the work of translation. Obviously, the computer will work faster and cheaper than human. In the last two decades, it is seen that research in the field of MP leads to a translation model is built automatically from the parallel corpus. Models are usually called statistical machine translation (SMT) is using statistical techniques approach.

Initial research of SMT started by Brown et al. [6] with a word-based models, the process of translating word by word. This model has largely been replaced by more complex models, but still used as a basis for other models such as word alignment. Zens et al. [7] and Koehn et al. [8] proposed a model based phrases translated sentences based on consecutive words in the source sentence to the corresponding word in the target language. The phrase the term phrase in this case simply means adjacent words, not the actual phrase in terms of grammar. Initially, the models stem phrase-based from research by Och and Weber [9]; Och et al. [10]; and Och and Ney [11]. The pair [12] proposed the use of the phrase in a word decoding model-based. Accordingly, the use of log-linear models proposed by Och and Ney [13].

In general, the architecture of statistical machine translation as shown in Figure 1. The primary data source used was a parallel corpus and monolingual corpus. The training process on the parallel corpus produce a translation model (TM). The training process in the target language parallel corpus, coupled with the monolingual target language corpus, generating language model (LM). While the features of the model (FM) generated from the target language on the parallel corpus, that every word has been characterized by linguistic features such as PoS, lemma, gender, the process of forming the word (morpheme) and others. TM, LM and FM results of the above process is used to generate decoder. Furthermore, the decoder is used as a machine translator to generate the target language from an input sentence in the source language.

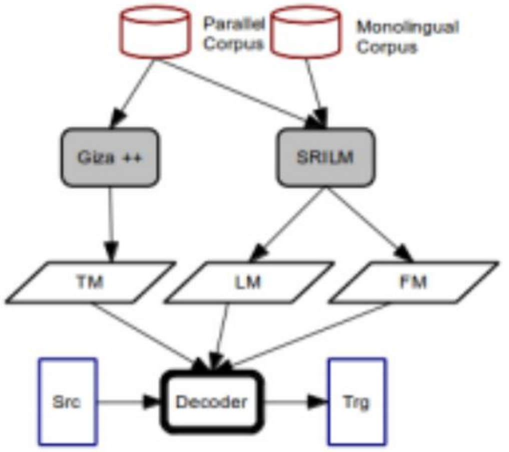

Fig. 1. Architecture Engineering Translators Statistics

If seen from MAPS architecture, it is clear that key data used to produce models in MPS is parallel corpus. Monolingual corpus can be obtained from the parallel corpus in the target language although usually propagated again from other sources.

The position of the experimental induction of word class without the supervision of machine translation as an instrument of experimentation can be seen in Figure 2.

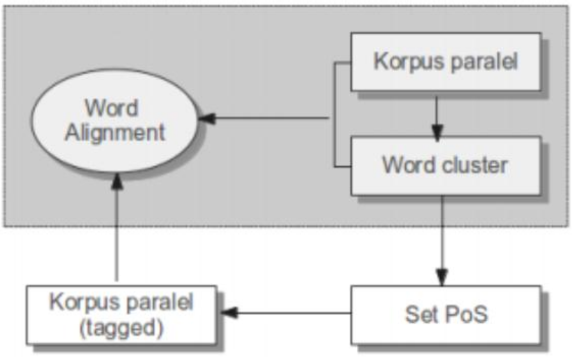

Fig. 2. Position Experiments on machine translators Statistics

Word alignment is one of the important processes in machine translation, there are two ways the use of information lexical to improve the accuracy of the word alignment, the first way is to use the word cluster is generated from a corpus of parallel automatically, while the second way is to use a corpus that has characterized each he said with PoS corresponding to these words. These experiments use the first approach is to look at the effectiveness of the algorithm EWSB in English as the target language.

Several approaches to the study of induction PoS has been done, such as the use of "Class-based n-grams" [14] which uses bigram models; "Class-based n-grams with morphology" [15], which uses a model that is similar to a class-based n-gram and clustering kind words; "Chinese Whispers graph clustering" [16] which induces value | C | with a clustering algorithm called "Chinese Whispers" based on contextual similarity; "Bayesian HMM with Gibbs sampling" [17] which is based on standard HMM for PoS tagging; "Sparsity regularization posterior-HMM" [18] which uses a Bayesian approach; "Feature-based HMM" [19] which uses the structure of the HMM models with standard and "Extended Word Similarity Based Clustering" [5] which uses n-gram approach in the process of clustering. In the research on the use of algorithms Extended Word Similarity Based (EWSB) on machine translation of English into Indonesia, Sujaini et al. reported that the use of these algorithms can improve translation accuracy of 2.07%.

Agglomerative approach to the use of hierarchical clustering algorithm for clustering purposes as indicated by the word Sujaini et al. [5] as follows:

1. initialize each unique word (token) as one cluster,
2. calculate the similarity between two clusters,
3. sort the rankings between all pairs of clusters based on similarity, and then combine the two top clusters.
4. stop until the desired number of clusters, if not, go back to step 2.

To compute the similarity between two clusters in step 2, use average linkage clustering method.

III. METHOD

Research data which is used is a parallel corpus Indonesian-English by 12 K sentences and monolingual corpus Malay Pontianak at 50 K sentence. While the process of clustering algorithms performed on 12 K EWSB Malay Pontianak sentence taken from the parallel corpus.

The research instruments used are as follows: (1) Moses: used as machine translation, (2) SRILM: used to build language models, (3) Giza ++: used for word alignment, (4) BLEU: used for the assessment of the results of translation and (5) Perl: used to build the program from the algorithm EWSB. Translator system built using algorithms EWSB on word clustering process compared to using a machine translator MKCS (reference algorithm of GIZA ++) on word clustering process as a baseline.

The accuracy of the translation system is measured using BLEU. In this experiment used 12 K sentences are divided into six fold, namely: Fold 1: sentence No. 1-2000, fold 2: sentence No. 2001-4000, fold 3: sentence No. 4001-6000, fold 4: sentence No. 6001-8000, 5 fold: no sentence 8001-10000, and fold 6: No. 10001-12000 sentence.

IV. RESULT AND DISCUSSION

The experiments that have been conducted, obtained the test results of each test group are shown in Table I.

TABLE 1 BLEU VALUE FROM TRANSLATION ACCURACY TEST RESULTS

Group Test	Corpus (fold)	Sentence Test (fold)	BLEU Score (%)	
			baseline	EWSB
A	2,3,4,5,6	1	68.27	70.39
B	1,3,4,5,6	2	74.10	75.62
C	1,2,4,5,6	3	74.21	74.66
D	1,2,3,5,6	4	73.95	75.85
E	1,2,3,4,6	5	71.82	74.34
F	1,2,3,4,5	6	70.22	72.45

From Table 1, it can be calculated the average value of BLEU, which represents the accuracy of the interpretation system. The system produces an average baseline value of BLEU by 72.10%, while the use of EWSB algorithm produces an average BLUE value of 73.89%. This shows that the use EWSB improve translation accuracy of 2.48%. Although not significant, the results showed that the algorithm EWSB quite effective if used in the Malay Pontianak language.

Some examples of using the translation results as a baseline comparison and EWSB algorithm can be seen in Table II.

TABLE II. COMPARISON OF TRANSLATION USING MKCLS AND EWSB

No		Sentence
1	Input	***saya hanya mengambil sedikit saja***
	Ref	*kamék cume ngambék sikit jak bah*
	BL	*aku cume ngambék sikit jak (BLEU= 54.75 %)*
	EWSB	*kamék cume ngambék sikit jak (BLEU= 81.87 %)*
2	Input	***alasan penundaan karena para peserta yang hadir dalam rapat tersebut dalam kondisi mengantuk***
	Ref	*alasan penundaan kerne para peserta nang datang dalam rapat tu dalam kondisi mengantok*
	BL	*alasan penundaan kerne para peserta yang datang dalam rapat tu dalam kondisi mengantok (BLEU = 76.12 %)*
	EWSB	*alasan penundaan kerne para peserta nang datang dalam rapat tu dalam kondisi mengantok (BLEU = 100.00 %)*
3	Input	***terkadang juga memang agak sulit membedakan antara kata dasar dengan kata berimbuhan***
	Ref	*tekadang ugak iye payah gak mbedakan antare kate dasar ngan kate berimbuhan*
	BL	*tekadang ugak mang agak payah gak mbedakan antare kate dasar ngan kate berimbuhan (BLEU = 70.16 %)*
	EWSB	*tekadang ugak iye payah gak mbedakan antare kate dasar ngan kate berimbuhan (BLEU = 100.00 %)*
4	Input	***anaknya tidak sekolah lagi karena tidak ada biaya***
	Ref	*anaknye tadak sekolah agik karne tadak ade duét*
	BL	*anaknye tadak sekolah agik karne tak ade kalok biaye (BLEU = 46.71%)*
	EWSB	*anaknye tadak sekolah agik karne tak ade duét (BLEU = 59.46 %)*
5	Input	***kalau memakai kaca mata barulah dia kelihatan ganteng***
	Ref	*kalok makék kace mate baroklah die keliatan lawar*
	BL	*kalok makék kace mate baroklah die keliatan lawar ke (BLEU = 86.33 %)*
	EWSB	*kalok makék kace mate baroklah die keliatan lawar (BLEU = 100.00 %)*

Note : BL = Baseline

Some examples of translation results show that the system of translators with EWSB managed to fix translation errors from the baseline system. In the first sentence, the word "saya" were not successfully translated as "kamék" by the baseline system can be decoded correctly by the EWSB system. In the second sentence, the word "that" were not successfully translated as "nang" by the baseline system can be decoded correctly by the EWSB system. In the third sentence, the phrase "terdakang juga memang agak sulit" translates to "tekadang ugak iye payah gak" tby the baseline system, while the system EWSB translates as "tekadang ugak iye payah gak" as the reference sentence. The phrase "tidak ada biaya" in the fourth sentence should be translated as "tadak ade duet" translated into "tadak ade kalok biaye" by the baseline system, these errors can be corrected by the EWSB system. In the fifth sentence, the phrase "kelihatan ganteng" translates to "kelihatan lawar ke" by the system baseline, the error is corrected by the EWSB system with translation "kelihatan lawar". From the above experimental results, it is evident that the addition of the word class information enclosed as linguistic information can improve the accuracy of statistical machine translation translator. Because there are no other variables that differ between the two systems except the word clustering algorithm used in the training process, it can be concluded that the improvements of the translation results caused by the use of algorithms EWSB.

V. Conclusion

From the experiments, performed on statistical machine translation algorithms using EWSB, the use of the algorithm can improve the accuracy of translations of 2.48% compared to the baseline system, so it can be concluded that EWSB algorithm can be recommended for use in machine translation using cognates, especially in languages that use rules "Diterangkan-Menerangkan" (DM) as Indonesian and Malay.

In the future, there should be further studies to find a new clustering algorithm word that can improve the quality of translations

References

[1] P. Koehn, dan H. Hoang, "Factored translation models", Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning , Prague, 2007.

[2] I. Youssef, M. Sakr, dan M. Kouta, "Linguistic factors in statistical machine translation involving arabic language*", IJCSNS International Journal of Computer Science and Network Security*, Vol.9 No.11, 1999.

[3] Nedjo1 A.T. dan Degen, H., "Augmenting Performance of SMT Models by Deploying FineTokenization of the Text and Part-of-Speech Tag", *Computer and Information Science,* Vol. 8, No. 1, 2015.

[4] Razavian, N. Sharif, dan S. Vogel, "Fixed length word suffix for factored statistical machine translation*", Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, 2010.

[5] H. Sujaini, Kuspriyanto, A.A. Arman, and A. Purwarianti, "Extended Word Similarity Based Clustering on Unsupervised PoS Induction to Improve English-Indonesian Statistical Machine Translation", *16th ORIENTAL COCOSDA/CASLRE-2013*, Gurgaon, India, 2013.

[6] Brown, P. F., Pietra, V.J.D., Pietra, S.A.D., dan Mercer, R. L., "The mathematics of statistical machine translation", *Computational Linguistics*, 19(2), 263–313, 1993.

[7] Zens, R., Och, F. J., dan Ney, H, "Phrase-based statistical machine translation", *Proceedings of the German Conference on Artificial Intelligence (KI 2002),* Heidelberg, 48-54, 2002.

[8] Koehn, P., Och, F. J., dan Marcu, D., "Statistical phrase based translation". Proceedings of the Joint Conference on Human Language Technologies and the An-nual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL), Edmont, 2003.

[9] Och, F. J. danWeber, H., "Improving Statistical Natural Language Translation With Categories And Rules". *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL),* Montreal, 1998.

[10] Och, F. J., Tillmann, C., and Ney, H., "Improved Alignment Models For Statistical Machine Translation". Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC), 20–28, 1999.

[11] Och, F. J. dan Ney, H., "The Alignment Template Approach to Statistical Machine Translation", *Journal Computational Linguistics*, 30(4), 417–449, 2004.

[12] Marcu, D., "Towards A Unified Approach To Memory And Statistical-Based Machine Translation", *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL),* Toulouse, 378-385, 2001.

[13] Och, F. J. dan Ney, H., "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation", *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL),* Philadelphia, 295-302, 2002.

[14] P. F. Brown, V.J. Della, V. Peter, Desouza, J.C.. Lai, dan R.L. Mercer, "Class-based n-gram models of natural language", Computational Linguistics, Vol. 18, No. 4., pp. 467-479, 1992.

[15] A. Clark, "Combining distributional and morphological information for part of speech induction*", In Proceedings of EACL 2003*, pages 59–66, Morristown, NJ, USA, 2003.

[16] C. Biemann, "Unsupervised part-of-speech tagging employing efficient graph clustering*", In Proceedings of COLING ACL 2006*, pages 7–12, Morristown, NJ, USA, 2006.

[17] S. Goldwater and G. Tom, "A fully bayesian approach to unsupervised part-of-speech tagging*", In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, 2007.

[18] J. Graca, G. Kuzman, T. Ben, and P.Fernando, "Posterior vs parameter sparsity in latent variable models", *Advances in Neural Information Processing Systems* 22, pages 664–672, 2009.

[19] T. Berg, Kirkpatrick, B.C. Alexandre, D. John, and K. Dan, "Painless unsupervised learning with features", *In Proceedings of NAACL 2010*, pages 582–590, Los Angeles, California, June, 2010.