English–Indonesia Machine Translation Using Statistical Approach

Y. Astuti[*], T.B. Adji[#], S.S. Kusumawardani[#]
EE&IT Departement, Gadjah Mada University
Jalan Grafika 2, 55281, Yogyakarta, INDONESIA
[*]yennistut_s209@mail.te.ugm.ac.id
[#]{adji, suning}@mti.ugm.ac.id

Abstract— **Most of the digital information is available in English language. However, Indonesian people do not use English as the daily conversation. This makes the English proficiency of most Indonesian becomes very low. To overcome this situation, the development of Machine Translation (MT) is needed which maps English words into Indonesian words in one-to-many, many-to-one, or many-to-many. Thus, a method should be provided to handle these words mapping. This paper proposed an MT technique using statistical approach to solve the problem. By using the technique, the English–Indonesian translation of a source word becomes more adaptable to the word context within a sentence.**

Keywords-component; machine translation; statistical

I. INTRODUCTION

Digital information is available in many languages, which most of them use English language. Indonesia is one of the countries that do not use English as a daily language. It makes the English' ability of Indonesian people becomes very low. This fact triggers a need of a Machine Translation (MT). The MT is defined in Reference [1] as the use of computer in automating some or all the translating process from a language to others.

For many decades, researcher had used many methods to make a robust and flexible MT. There are five methods that are already known in the MT development field. The first three methods are called the classical approaches namely direct approach, rule-based/transfer approach, and Interlingua approach [2]. The other two methods are a data-driven approach, which are the example-based approach and the statistical approach.

A direct approach MT research was developed by a research group from Gadjah Mada University, Indonesia [3]. This MT research could handle many tenses; such as present, present continuous, present perfect, past, past perfect, and future tenses. However, the precision of this MT system had not been examined yet. An Interlingua MT research was made by a group project namely Multilingual Machine Translation System (MMTS) [4]. This is a multi-national project research among China, Indonesia, Malaysia, Thailand, and Japan as the project leader. The MMTS includes an analysis component for Indonesian language part that called Bahasa Indonesia Analyzer System (BIAS). BIAS uses Indonesian texts as the input and abstract meaning as the output. Unfortunately BIAS accuracy was not presented. A rule-based MT research was done by two researchers from Petra University, Indonesia [5]. This research translates many sentences from English to Indonesian language but the precision had not been evaluated yet. This rule-based MT is able to translate daily conversation sentences. However, the system could not handle a word that has more than one meaning.

The data-driven approaches are also known as corpus-based approaches. This approach uses bilingual corpora to automate the information of the translation learning. Bilingual corpora are defined as texts that are available in parallel in two different languages. The use of bilingual corpora can minimize the human involvement. Consequently, this approach can achieve a rapid development of MT systems only in a couple of months. As well as the rapid achievement, this approach can also overcome the bottlenecks of rule-based approach [6]. A research that uses example-based method was done by Brown [7]. This research uses the example-based method in translating Spanish to English language. An MT application that uses statistical approach is Google-Translate. This application can translate many languages (includes English) into Indonesian. This MT system uses a statistical approach based on phrase translation [8]. However, this MT system was not provided in open-source. Another MT activity that uses statistical approach was conducted by Agency for the Assessment and Application of Technology (BPPT) and National News Agency (ANTARA). This MT system was developed using Pharaoh as the decoder in 500K training pair sentences [9].

In this paper, the MT system is developed in the availability of the bilingual corpora, in which English is the source language and Indonesian is the target language. The technique of English-Indonesian MT using statistical approach will be explained in several sections. Section II provides the explanation about statistical method in MT. Section III and IV give details about other techniques that are needed in the statistical MT; such as alignment and decoder. Section V explains the technique of English – Indonesian MT using statistical approach. Section VI describes the implementation and analysis. Finally, Section VII gives the discussion about the work.

II. STATISTICAL METHOD

Statistical method for MT is developed by applying Bayes Rule as we can see in (1) [10].

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)} \qquad (1)$$

The denominator, *P*(*S*), can be discarded because the right equation does not depend on *S*. This *S* symbol is equal for all the target sentence possibilities. Thus, (1) can be written as (2), where *T* refers to the target sentence (Indonesian) and *S* refers to the source sentence (English).

$$P(T|S) = P(S|T)P(T) \qquad (2)$$

The *P*(*S*|*T*) factor is called Translation Model (TM) that represents the MT faithfulness. The *P*(*T*) factor is named as Language Model (LM), which represents the MT fluency [11].

The best translation of the source language is found by applying (3).

$$\arg\max P(T|S) \qquad (3)$$

A. Translation Model (TM)

The TM or the *P*(*S*|*T*) shows the probability of the source sentence given the target sentence. TM gives information of the relation strength between the candidates target sentence and the source sentence. The bigger the probability, the stronger relation it must be. Mathematically, TM is presented in (4) [12], where s_n refers to the source word in the *n*-th position that is translated into t_n.

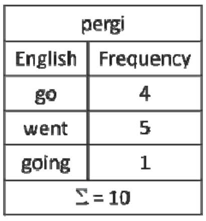
$$P(S|T) = P(s_1|t_1)P(s_2|t_2)\ldots P(s_n|t_n) \qquad (4)$$

Given a source sentence “I go to the market by bicycle” then the relation with the target sentence *Saya pergi ke pasar naik sepeda* can be calculated as shown in (5).

P(I go to the market by bicycle|Saya pergi ke pasar naik sepeda) =

P(I|Saya) *P*(go|pergi) ... *P*(bicycle|sepeda) (5)

The probability of *P*(go|pergi) can be obtained from the training corpora. If we have a translation table of the word *Saya* as in Fig.1, then the relation between the word *pergi* and the word “go” is equal to 0.4. Besides the word “go”, the word *pergi* can be the tranlation of the words “went” and “going”.



pergi	
English	Frequency
go	4
went	5
going	1
Σ = 10	

Figure 1. The word pergi and its translations

B. Language Model (LM)

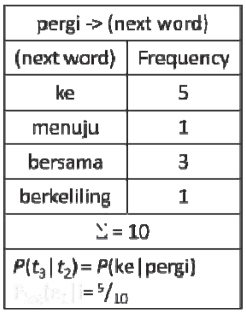
P(*T*) shows the relation of the word orders of the target language. Mathematically, the LM can be computed using chain-rule as shown in (6)

$$P(t_1, t_2, \ldots t_n) = P(t_1)P(t_2|t_1)\ldots P(t_n|t_1 t_2 \ldots t_{n-1}) \qquad (6)$$

It is explained in (6) that the occurrence of a target word is affected by (*N*-1) previous target words. Equation 6 is known as *N*-gram model [13]. If the occurrence of a word is affected by one previous word, then it is called bigram model. If the occurrence of a word is affected by two previous words, then it is trigram model. Therefore, the chain-rule for the bigram model is expressed as in (7) [10].

$$P(t_n|t_{n-1}) = \frac{count(t_{n-1}t_n)}{count(t_{n-1})} \qquad (7)$$

The bigram for *P*(t_3|t_2) of the sentence *Saya pergi ke pasar naik sepeda* is shown in Fig. 2.



pergi -> (next word)	
(next word)	Frequency
ke	5
menuju	1
bersama	3
berkeliling	1
Σ = 10	
$P(t_3\,|\,t_2) = P(\text{ke}\,|\,\text{pergi})$ $= {}^{5}/_{10}$	

Figure 2. Bigram probability for P(t_3|t_2) in Saya pergi ke pasar naik sepeda

There are occasion where word order does not appear, i.e. $P(t_i|t_x) = 0$. In this case, a smoothing technique should be applied. It is stated in [10] that there are two smoothing techniques; one of them is *Witten-Bell discounting*. The Witten-Bell discounting smoothing technique is applied using (8), where $T_y(t_x)$ refers to the number of t_x bigram type.

$$\sum_{i:count(t_x t_i)=0} P_{WB}(t_i|t_x) = \frac{T_y(t_x)}{count(t_x) + T_y(t_x)} \qquad (8)$$

III. Alignment

Alignment is a process of translation training from the parallel corpora. This process must produce the word translations to be used in the TM. However, before getting the word alignments, we must solve the sentence alignment. Sentence alignment finds the parallel sentences from the parallel paragraphs. Afterwards, these parallel sentences will be worked out to obtain the parallel words. In [14], there are two approaches to solve the word alignment. The first method is lexical alignment and the second one is EM (Expectation Maximization) algorithm. This work will use the lexical alignment because of its simplicity.

Lexical alignment obtains the word translation by using a dictionary. For example a parallel sentence “I go to the market by bicycle” and *Saya pergi ke pasar naik sepeda*. From the English–Indonesian dictionary, we will find the words “I”, “go”, “to”, “market”, “by”, “bicycle” that correspond in parallel with the words *Saya, pergi, ke, pasar, naik, sepeda*. In this example, the word “the” does not align to any target word, so we add “NULL” (no-alignment) for the parallel of the word “the”.

IV. DECODER

The decoder is also called the searching algorithm. The function of the decoder is to find the best translation to fulfill Equation (3). The translation initiate with the initial translation. This initial translation will be then improved to obtain the best translation. In [15], there are three kinds of decoder methods that we can apply i.e. stack-based decoder, greedy decoder, and integer programming decoder. This research will use the greedy decoder because of its ease in making initial translation. In addition, the greedy decoder translation's quality is good enough compared to the other two decoders [15].

Let us consider to find the best translation of "I go to the market by bicycle" (see Fig. 3). For the initial translation, the greedy decoder will find the biggest probability among target words in the translation table. For example, the word "I" can be translated into the word *saya* with the probability of 0.6 or the word *aku* with the probability of 0.4. The decoder will choose the word *saya* because it has the biggest probability.

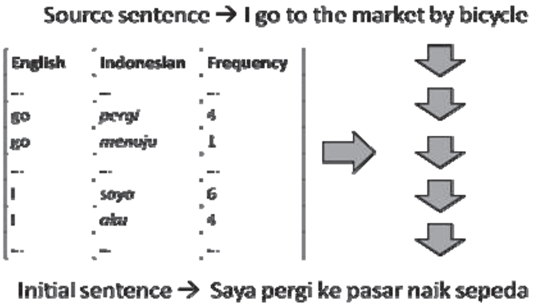

Source sentence → I go to the market by bicycle
English
Indonesian
Frequency
--
--
--
go
pergi
4
go
menuju
1
--
--
--
I
saya
6
I
aku
4
--
--
--
Initial sentence → Saya pergi ke pasar naik sepeda

Figure 3. The initial sentence for the source "I go to the market by bicycle"

V. ENGLISH–INDONESIA STATISTICAL MT

The English–Indonesian statistical MT technique can be illustrated as in Fig. 4 that is divided into two main blocks; Training block and Testing block.

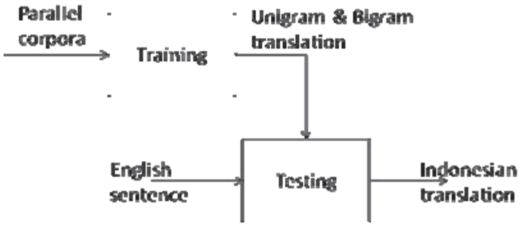


Figure 4. The main diagram of English-Indonesian Statistical MT

There are two kinds of input for this system. The first input is the parallel corpora that are the input for the Training block. The second one is the English sentences that are the input for the Testing step. The output of the Training step, which is Unigram & Bigram translation, will affect the Testing block's output.

The Training block can be seen in Fig. 5. In this work, we use 30 parallel sentences for the training. These parallel sentences are aligned in the Alignment component to obtain the word alignments. The word alignments are saved in the bigram and unigram models after being processed in the Bigram and Unigram components.

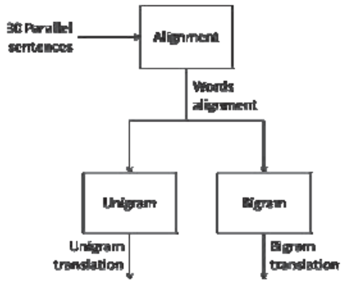


Figure 5. The Training block of English-Indonesian Statistical MT

The Testing block of the English – Indonesian Statistical MT consists of two components (see Fig. 6). The first component is called the Decoder. In this component, each English sentence is translated by using greedy decoder technique. This component provides the initial translation that will be improved by the next component, which is called the Bigram model. If the improved translation is the best translation then the system will consider this as the final translation. Nonetheless, if the translation can still be improved, the Bigram model will reprocess it until the best translation is obtained.

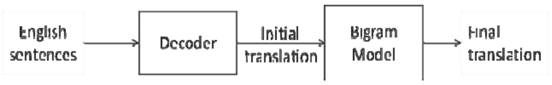


Figure 6. The Testing block of English-Indonesian Statistical MT

VI. THE RESULT AND ANALYSIS

The technique that is explained in Section V will be implemented to 30 English – Indonesian sentences for the Training block, and seven English sentences for the Testing block. The English sentences for both training and testing are provided in the Internet and taken from stories about myths, legends, and fables that we can find at the popularchildrenstories.com, eastoftheweb.com, and longlongtimeago.com. These sites provide stories only in the English language. Thus for the Training block, we have to translate them into Indonesian language, of which the grammar was given in details by Dwipayana [16], Keraf [17], Widyamartaya [18], and Wilujeng [19]. The testing results for the seven English sentences are given in Table I.

Now let us see the implementation in the Training block. Our first parallel sentence are "Otherwise the gods will be angry with you" and *Sebaliknya, para dewa akan marah padamu.* First, the Alignment component processes this parallel sentence. The result of this alignment can be seen in Table II. After all the word alignments are done the Unigram translation and the Bigram translations are implemented. The unigram table will save each word translation, as shown in Table III. Meanwhile, the bigram table will save every two words of the target words, as can be seen in Table IV.

TABLE I. TESTING RESULT

No	*English*	*Result*
1.	The cottage was surrounded with the angry people.	*Gubug mengelilingi dengan marah penduduk*
2.	They catch those fishes with nets.	*Mereka menangkap those fishes dengan jaring*
3.	The rich man was very pleased with the news.	*Man yang kaya itu sangat senang dengan kabar.*
4.	The crocodile did not agree with him	*Buaya itu did tidak setuju bersamanya.*
5.	I begged you to come with me to the party.	*I begged kau datang denganku ke pesta*
6.	She smiled at the Prince with joy.	*Dia tersenyum pada Prince dengan sukacita.*
7.	The two sisters married with the two rich gentlemen.	*Two sisters nikahi dengan two yang kaya pria.*

TABLE II. ALIGNMENT RESULT FOR THE 1^{ST} PARALELL SENTENCE

English	*Indonesian*	*Count*
the	*tersebut*	*1*
falcon	*elang*	*1*
flew	*terbang*	*1*
towards	*mendekati*	*1*
earth	*tanah*	*1*
with	*membawa*	*1*
the	*itu*	*1*
violin	*violin*	*1*

TABLE III. UNIGRAM TRANSLATION

English	*Indonesian*	*Count*
...	...	...
otherwise	*sebaliknya*	*1*
the	*para*	*1*
gods	*dewa*	*1*
will	*akan*	*1*
be	NULL	*1*
angry	*marah*	*1*
with	*pada*	*1*
you	*mu*	*1*
...	...	...

TABLE IV. TARGET LANGUAGE BIGRAM

Indonesian	*English*
...	...
sebaliknya	*Otherwise*
sebaliknya para	*otherwise the*
para dewa	*the gods*
dewa akan	*gods will*
akan (NULL)	*will be*
(NULL) *marah*	*be angry*
marah pada	*angry with*
padamu	*with you*
...	

Now, we can proceed to the Testing block. Given an English sentence: "The cottage was surrounded with the angry people." then this sentence will be translated into Indonesian language as explained in the following lines. The first step of the Testing is Decoder step. It means that we have to go to the Unigram Translation Table. By choosing the biggest probability of each word then we will obtain the initial translation: as (NULL) *gubug* (NULL) *mengelilingi dengan* (NULL) *marah penduduk*. This initial translation will be improved in the Bigram component. Unfortunately, this initial translation cannot be improved anymore because there are no bigram translations that can improve it. Thus, the translation of the sentence "The cottage was surrounded with the angry people" is the sentence *gubug mengelilingi dengan marah penduduk*.

The translation's accuracy of the system based on unigram and bigram evaluations are 58% and 35% respectively. These values are obtained by comparing the machine result and the human translation [20].

The performance of machine translation sentence is not good enough because of the small number of corpus that we used (30 parallel sentences only). If we add larger parallel sentences, then the translation coverage will increase. As a result, the translation will be better.

If we have another parallel sentence that contains the words "was surrounded" and (NULL) *dikelilingi*, then the initial translation could be improved as *gubug dikelilingi dengan marah penduduk*. If we have another sentence containing the words "angry (NULL)" and "(NULL) people" that parallel with *penduduk yang* and *yang marah* then the translation could be improved as *gubug dikelilingi dengan penduduk yang marah*. If we have another parallel sentence that contains the words "the cottage" and *gubug itu*, then the translation would become *gubug itu dikelilingi dengan penduduk yang marah*. This last result would be much better than the initial sentence. However, this better result can be achieved if we have a huge corpus.

VII. DISCUSSION

As it is explained in Section VI, the number of parallel sentences is not sufficient to provide good translations. Thus, huge parallel sentences should be added in the Training block.

The decisions of the good or bad translation are based on the manual human perception since this work has not included any automatic evaluation method. There are many automatic evaluation methods in MT field and one of them is BLEU-metric [20]. This evaluation – that compares the output of the MT system with four human translations – will be the next possible future research.

REFERENCES

[1] R. M. Kaplan, "A general syntactic processor," in Natural Language Processing, Rustin, R., Ed. New York: Algorithmics Press, 1973, pp. 193-241.

[2] T. B. Adji, "Annotated disjunct for machine translation," Computer and Information Science Departement, Universiti Technology Petronas, Unpublished Dissertation Rep. Malaysia, 2010.

[3] F. Novento, "Perangkat Lunak Penerjemah Kalimat Inggris-Indonesia Menggunakan Metode Loading Data Sementara," Electrical Engineering Department, Gadjah Mada University, Final Rep., 2003.

[4] H. Yusuf, "An Analysis of Indonesian Language for Interlingual Machine-Translation System," Proc. 14th Conf. on Computational Linguistic (COLING), Nantes, France, Aug. 1992, vol. 4, pp. 1228-1232.

[5] E. Utami and S. Hartati, "Pendekatan metode rule based dalam mengalihbahasakan teks Bahasa Inggris ke teks Bahasa Indonesia," Jurnal Informatika, vol.8, no.1, 2007, pp: 42 – 53.

[6] K. Probst, "Learning transfer rules for machine translation with limited data," Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Ph.D. dissertation, Aug. 2005.

[7] R. D. Brown, “Example-based machine translation in the PANGLOSS system,” Proc. 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996, pp. 169-174.

[8] F. J. Och and H. Ney, “Improved statistical alignment models,” Proc. 38th Annual Meeting of the Association for Computational Linguistics (ACL), Hong Kong, 2000, pp. 440-447.

[9] H. Riza, “Resources report on languages of indonesia,” Proc. 6th Workshop on Asian Language Resources, Hyderabad, India, Jan. 2008, pp. 93-94.

[10] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, “A Statistical approach to machine translation,” in Journal of Computational Linguistics, vol.16 no.2, Jun. 1990, pp. 79-85.

[11] R. Francois and P. Lison, “Probabilistic language modeling with N-grams,” Artificial Intelligence Seminar, Universit´e Catholique de Louvain. Belgium, May 2005.

[12] C. Nusai, Y. Suzuki and H. Yamazaki, “Estimating word translation probabilities for Thai – English machine translation using EM Algorithm,” International Journal of Computational Intelligence 4, 2008.

[13] A. Ramanathan, P. Bhattacharyya and M. Sasikumar, “Statistical Machine Translation,” Mumbai, India, Dissertation Seminar Report, 2005.

[14] P. Koehn, “Empirical Methods in Natural Language Processing Lecture 15,” School of Informatics, California, 2008.

[15] U. Germann, M. Jahr, K. Knight, D. Marcu and K. Yamada, “Fast decoding and optimal decoding for machine translation”. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL), Toulouse, 2001, pp. 228–235.

[16] G. Dwipayana, “Sari Kata Bahasa Indonesia (The Essence of Indonesian Language)”, Surabaya: Terbit Terang, 2001.

[17] G. Keraf, “Tata bahasa rujukan Bahasa Indonesia”, Jakarta: PT Gramedia Widiasarana, 1999.

[18] A. Widyamartaya, “Seni menerjemahkan”, 13th ed. Yogyakarta: Kanisius, 2003.

[19] A. Wilujeng, “Inti sari kata Bahasa Indonesia lengkap”, Surabaya: Serba Jaya, 2002.

[20] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, Pennsylvania, 2002, pp. 311-318.