Augmented Javanese Speech Levels Machine Translation

Aji P. Wibawa[1], **Andrew Nafalski**[2], and **Wayan F. Mahmudy**[3], Non-members

ABSTRACT

This paper presents the development of the hybrid corpus-based machine translation for Javanese language. The system is designed to deal with the complexity of politeness expression and speech levels of Javanese that is considered as a local language with the biggest number of users in Indonesia. Statistical features are embedded to increase the performance of the system. The edit shifting distance is applied due to increase the alignment efficiency. However, improper alignment contributed by recorded impossible pair and insufficient data training is still detected. This paper proposes a new improvement of the developed alignment algorithm based on the impossible pair restriction. Based on experimental results, the new developed algorithm is more accurate (A=93.8%) even though the number of training data is less than the old one (A=87.9%).

Keywords: Javanese Speech Levels, Corpus, Machine Translation, Impossible Pair Limitation

1. INTRODUCTION

Respecting others properly by using paralinguistic forms and proper speech is part of the politeness in Javanese culture. The appropriate degree of politeness is often articulated in the form of deference in communication. Speech levels [1, 2], speech styles [3] or Javanese language politeness [4] are frequently used to name the degree of deference.

The levels of speech and associated politeness forms are being neglected with fewer speakers being conversant in them even though Javanese is considered as the most widely used regional language in Indonesia [5, 6].Negative tendency is detected concerning the use of Javanese speech levels among teenagers. They may select an incorrect speech level to address a high-status person since they are unable to transform the local politeness value into its equivalent refined language [1, 7, 8]. Furthermore, the selection of incorrect vocabularies [4] indicates that they lack mastery of speech levels and do not know how to use them appropriately in verbal communication. In fact, the acquisition of speech levels among teenagers can be classified as very poor: 36.45 out of 100. This finding was revealed in a research on the use of speech levels by youngsters in Solo [8] and the result was based on written vocabulary translation tests.

Realizing that they cannot handle this polite form, younger speakers usually switch into Indonesian language(*bahasa Indonesia*), which they can handle more easily and they believe to be more reliable to use in the global era [7-9]. If this continues, the *krama* form- a unique characteristic Javanese-is in the danger of diminishing. In addition, unskilled educators and the lack of speech levels' guides may be exacerbating the problem [10, 11]. While teachers are expected to serve as language models at school [1], some of them use inappropriate words and levels [4] - a fact which further suggests that Javanese speech levels dying out. Hence, a machine translation should be developed to protect the Javanese speech levels from being extinct.

A machine translator has been developed in order to protect the existence of the speech level [12, 13]. The translator provides bilingual pragmatic translation between speech levels [12] and also Indonesian [13]. The translation knowledge is based on a bi-text alignment that reinforced by edit shifting distance algorithm [14]. The results are impressive; however, word repetition [13] and pragmatic translation mistakes [12] are detected during the translation.

This paper is an extended version of [12], focused on modifying the bilingual text alignment due to avoid the unwanted mistakes as well as increasing the translation accuracy. The novel algorithm will be compared to the previous model with extra training data.

2. THEJAVANESE SPEECH LEVELS

Javanese linguists [2, 15, 16] divide the speech levels into three classes: *krama*, *madya* and *ngoko*. The classes can be further classified into nine sub-levels: *mudha-krama* (MK), *kramantara* (KA), *wredha-krama* (WK), *madya-krama* (MdK), *madyantara* (Md A), *madya-ngoko* (Md Ng), *basa-antya* (BA), *antya-basa* (AB), *ngoko-lugu* (Ng L). The sublevels has been simplified into four categories; *ngoko* (Ng), *ngoko alus* (NgA), *krama* (Kr) and *krama alus*

Manuscript received on March 31, 2013 ; revised on June 26, 2013.

Final manuscript received August 10, 2013.

[1] The author is with Department of Electrical Engineering, State University of Malang, Indonesia., E-mail: ajipw@um.ac.id

[2] The author is with School of Engineering, University of South Australia, Australia., E-mail: Andrew.Nafalski@unisa.edu.au

[3] The author is with Department of Computer Science, Brawijaya University, Indonesia ., E-mail: wayanfm@ub.ac.id

(KrA) in the first Javanese Congress in 1991[4]. The example of simplified of Javanese speech levels and its lexical characteristics is shown in Table 1.

Table 1: *Example of Simplified Speech Levels and Lexical Features.*

	Speech Levels				
	Ng	*NgA*	*Kr*	*KrA*	
Lexical Features	***Ng***	***Ng&KI***	***Kr***	***Kr &KI***	**Meaning**
Word List	*ana*	*ana&w onten*	*wonten*	*wonten*	be (indica tingexiste nce)
	bapak	*bapak*	*bapak*	*bapak*	father
	celuk	*celuk&t imbali*	*timbali*	*timbali*	call
	guru	*guru*	*guru*	*guru*	teacher
	ibu	*ibu*	*ibu*	*ibu*	mother
	lagi	*lagi&ne mbe*	*saweg*	*nembe*	"progressi ve" marker
	mangan	*mangan &dhaha r*	*nedha*	*dhahar*	eat
	murid	*murid*	*murid*	*murid*	student
	omah	*omah& dalem*	*griya*	*dalem*	house
Prono uns					
1st person SG	*aku*	*aku*	*kula*	*kula, kawula, dalem*	I
2nd person	*kowe*	*sliramu (young er),panj enenga n (older)*	*sampe yan*	*panjenen gan*	You
1st person PL	*awake dhewe*	*awake dhewe*	*kita*	*kita*	We
Affixes					
-ku	*-ku*	*-ku*	*kula*	*kula, kawula, dalem*	My
-mu	*-mu*	*panjene ngan*	*sampey an*	*panjenen gan*	Your
di-	*di-*	*di-*	*dipun-*	*dipun-*	"passive marker"

One-to-one word translation in Javanese may produce poor, inaccurate and inappropriate results [14] since one word may be literally translated into two words at another speech level, and vice versa.As a result, the source and target sentences may consist of an unequal number of words, known as asymmetric formation phenomenon. For example, the sentence ***awake dhewe ditimbali ibumu*** (Ng) is translated as ***kita dipuntimbali ibu panjenengan*** (Kr). Both sentences have four words; however, translating the words one-to-one based only their order in the sentence, produces an inaccurate translation. In fact, the sentences are composed of three asymmetric pairs (i.e. *awake dhewe*↔*kita; ditimbali?dipuntimbali; ibumu*↔*ibu panjenengan*). An example of alignment when translating ***bapakku mangan ana omahku*** into ***bapak kula dhahar wonten griya kula*** is shown in Figure 1. The top part shows direct alignmentthat produces inaccurate translation whereas the bottom part shows the correct alignment.

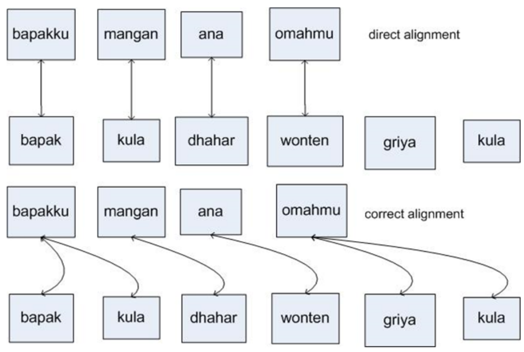


Fig.1: *Example of Direct and Correct Javanese Speech Level Alignment.*

The Javanese sentence has a basic structure of SVO (subject, verb, and object). The meaning of sentences is pragmatically diverse and related to subject and verb agreement (SVA). The SVA derives from non-linguistic factors (i.e. social status, ages and relationships) [17]. For instance, sentences (1) and (2) show the differences of SVA based onthe social status of the subjects where both are operating at krama level.

(1) ***Murid-murid****saweg<u>nedha</u>* (students are eating).

(2) ***Guru-guru*** *saweg<u>dhahar</u>* (teachers are eating).

The underlined words in both sentences have a same meaning (eating). However, the first sentence must use the word *nedha* because the subject of the first sentence is students who obviously have lower social status than their teachers in sentence (2). In consequence, the word *nedha* is used instead of *dhahar*, which is only suitable for honoured people.

3. REVIEW OF MACHINE TRANSLATION

Machine translation (MT), a branch of computational linguistics, is simply defined as the automatic computer-assisted translation of bilingual or multilingual natural languages[18]. Based on its knowledge base, MT is classified into two categories: Rule-based Machine Translation (RBMT) and Corpus-based Machine Translation (CBMT).

RBMT uses linguistic information such as semantic, morphological, and syntactic information as its knowledge base. The involvement of linguists in knowledge base development is unavoidable. They also build transfer rules between languages that are relatively complex and time consuming, especially for those with very different structures. As a result, the development costs increases in an effort to achieve the required quality threshold. The participation of language experts is desperately needed when more adaptive RBMT is desired. The linguists have to retrain the developed RBMT by adding new rules, vocabulary and other linguistic information, and this contributes further to extra development time and costs.

On the other hand, the knowledge of CBMT is based on large sets of bilingual text that are recognised as corpora[19, 20]. The two kinds of CBMT are example-based machine translation (EBMT) and statistical machine translation (SMT). The availability of corpora is a key factor in reducing development costs. Once available, the development time is reduced in parallel with the translation development cost. In contrast with RBMT, the role of linguists in the development of CBMT is purely optional. When a new instance of translation arises and unknown words occur, the corpus may be updated automatically. While RBMT is highly efficient for more general translation, CBMT may work better for a specific domain. Although some inconsistency may be found, for example, in pure SMT technique [21, 22] RBMT generates more natural translation than CBMT.

Both SMT and EBMT derive mainly from large bilingual texts; that is known as a corpus (i.e. corpora in the plural form) [19, 20]. While SMT focuses more on word combinations and their occurrence, EBMT focuses on text segmentation [20], phrase memorisation [22]and analogical sentence recombination [19]. Moreover, SMT has a more obvious definition than EBMT in terms of selecting the most appropriate translation. SMT selects the translation from the target with highest probability [20]while EBMT focuses on string matching of the user's input with the recorded source language [19, 20]. Since the quantity, quality, and domain of the data are crucial factors that determine its accuracy [23], SMT definitely outperform EBMT by increasing the training data [24].

Several studies of combination of machine translations reveals that the hybrid system (RBMT and EBMT [22], RBMT-SMT [25-28]) is better than stand-alone systems.Most of these approaches are used to translate English into another language such as Chinese [29-31], Portuguese [32], Persian [33], Swedish [34] and Japanese [35, 36]; however, the existence of Javanese translation does not exist.

4. THE DESIGN OF JAVANESE TRANSLATOR

The hybrid Javanese machine translationis a memory-based machine translation [37] with statistical features to retrieve and recombine the correct translation. The system works using a corpus (Fig. 2) and mainly consists of training and translation process.

4.1 Language Modeling

This study uses multilingual texts that are based on 1991 categories of Javanese speech levels. The

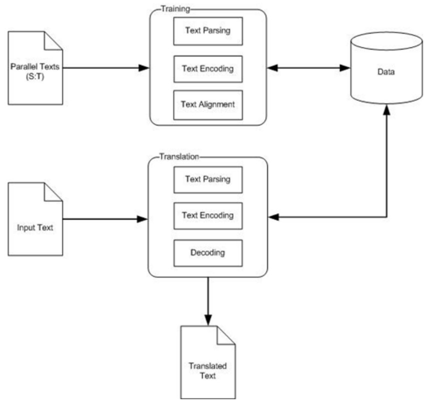


Fig.2: *The Design of Javanese Machine Translation.*

translation process uses the texts in both source language (S) and target language (T). Each text is divided into sentences (s_i and t_j) and can be formulated as in (1) and (2). Here, n_i and n_j represent the number of sentences in the source and target text respectively.

$$S = \{s_i : S|0 \leq i < n_i\} \quad (1)$$
$$T = \{t_j : T|0 \leq i < n_j\} \quad (2)$$

Afterwards, all correspondence sentences are parsed into a set of words (w_s and w_t) from the first word of both sentences (w_{s1} and w_{t1}) to the end of source sentence (w_{sj} where $j = n_{ws}$) and target sentence (w_{tj} where $j = n_{wt}$).

$$s_i = \{ws_i : s_i|1 \leq i < n_i\} \quad (3)$$
$$t_j = \{wt_i : s_t|1 \leq j < n_j\} \quad (4)$$

The bilingual translation is modelled as the translational equivalence models that records all feasible structural paired texts [23]. There are two probabilistic models of speech levels' translation, joint model and conditional model. The first modelin (5) considers that translation is a product of a joint probabilitybetween source (S) and target (T) language. The second multilingual translation model, (6) and (7), is based on the rule of conditional probability where both probabilities of S and T must meet the terms that they are part of a particular level of speech (L).

$$P(S,T) \in [0,1] \quad (5)$$
$$P(S|L) \in [0,1] \quad (6)$$
$$P(T|L) \in [0,1] \quad (7)$$

Firstly, the pair combination of Javanese text is modelled to accommodate the characteristics of the relevant speech levels. The pair combination (C) is then divided into two categories: lexical and pragmatic combinations. The lexical combination is based on the characteristic of Javanese words' translation and consists of (1:1), (1:2) and (2:1) word pair combinations. The pragmatic combination refers to Javanese subject-verb agreement (SVA) [9] is modelled by aligning (2:2) pairs. This pair captures the relationship of subject and verb in the parallel sentences as well as reinforces the one word to one word (1:1) alignment.

$$C=(S,T)\begin{cases} ((ws_i,0),(wt_j,0):(1:1)) \\ ((ws_i,0),(wt_j,wt_{j+1}):(1:2)) \\ ((ws_i,ws_{i+1}),(wt_j,0):(2:1)) \\ ((ws_i,ws_{i+1}),(wt_j,wt_{j+1}):(2:2)) \end{cases} \quad (8)$$

4.2 Database

A database is developed to store both language and statistical records. The database consists of four tables, table of Languages, Words, Phrases and Pairs. The relationship among table is presented in Fig. 3. In consequence, any editing process in a table may change the records in other tables.

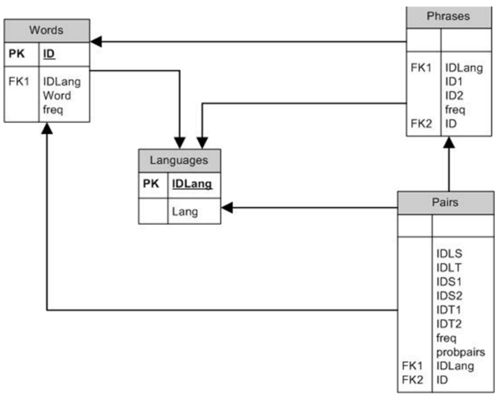


Fig.3: *The Database of Javanese Speech Level Translation.*

The *Languages* table has two columns, *IDLang* and *Lang*, that store the language identification number and the type of languages respectively. Four languages [4],*ngoko* (Ng),*ngoko alus* (NgA), *krama* (Kr) and*krama alus*(KrA) are recorded in the database. Others languages can be added anytime based on the users' needs.

The list of trained words is stored in the *Word* table and indexed by unique numbers (ID). The frequency of word is also stored as a base for the probabilistic calculation. The *IDLang* in this table is taken from the parent table (Languages) in order to differentiate the language from others. The *Phrases* table is used to record a pair words or phrases. The phrases are obtained from two words (ID1 an ID2) which refer to the word's identification number in the *Word* table. The *Pairs* table records any pair combinations of bilingual text, both in words or phrases. A dice coefficient based on the frequency (*freq*) of the word and phrase in the parallel text are used to calculate the probability (*probpairs*).

4.3 Text Parsing and Alignment

The learning process consists of two stages, text parsing and alignment process as shown in Fig.4. The parsing stage is employed to split every sentence into a set of discrete words. The result as well as the frequency of related words is automatically indexed and recorded into the database. The alignment process restructures the sentence into a monolingual array which consists of a unique number representing the word's index. The process is employed to speed up the alignment process. .The parallel text alignment process pairs the array of sentences based on the pair combination models of the Javanese. The stage involves aligning every possible word combination in S and T. Similarly to a monolingual process, the combination pairs and their frequency are then sent tothe database.

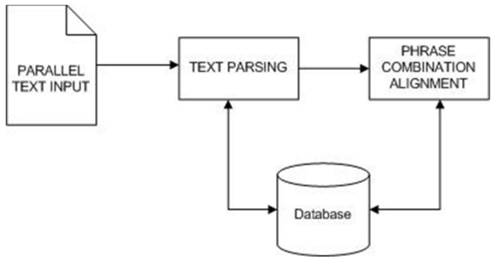


Fig.4: *Learning Process.*

The example in Fig.5 is used to analyze the parallel text alignment algorithm. As detailed in Table 2, total of 35 possible pairs have generated to represent three targeted pairs: (S2,T3), (S3,T4) and (S1,T1T2). The probability of the targeted pair is equal to 0.086 (3/35).

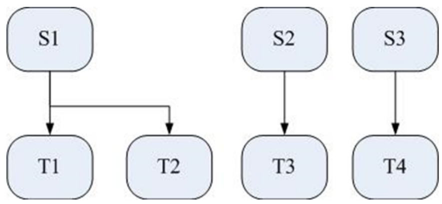


Fig.5: *Alignment Example.*

The knowledge base training relies on the phrase based alignment that is recognised as a flexible and accurate approach [31, 34]. All potential pairs are recorded [14]; however, the training algorithm may become less effective due to the bulk corpora generated [36]. The shifting distance algorithm is developed in order to trim down the number of irrelevant

Table 2: *Generated Pairs by the Original Alignment.*

Pair	Combination	Total
1:1	(S1,T1),(S1,T2),(S1,T3),(S1,T4) (S2 ,T1),(S2,T2),(S2,T3),(S2,T4) (S3 ,T1),(S3,T2),(S3,T3),(S3,T4)	12
1:2	(S1,T1T2),(S1,T2T3), (S1,T3T4) (S2,T1T2),(S2,T2T3), (S2,T3T4) (S3,T1T2),(S3,T2T3), (S3,T3T4)	9
2:1	(S1S2,T1),(S1S2,T2), (S1S2,T3), (S1S2,T4) (S2S3,T1),(S2S3,T2), (S2S3,T3), (S2S3,T4)	8
2:2	(S1S2,T1T2),(S1S2,T2T3), (S1S2,T3T4) (S2S3,T1T2),(S2S3,T2T3), (S2S3,T3T4)	6
		35

pair by limiting the alignment iteration. While the original alignment pairs the reference chunk (i) with all target words ($j - 1$ to $j = n_j$), setting a specific shifting distance (D) initiates the iteration in range of $j = i - D$ to $j = i + D$. Fig.5, illustrates the application of shifting distance coefficient in the bilingual alignment in form of pseudo code

for each sentence in source and target language
for j:= i - D to i + D do
train all possible pair combination
check the database
if the combination is unavailable in database then
record the pair combination with its frequency
else update the frequency of the pair

Fig.6: *Procedure of theParallelAlignment Adjusted byEdit Shifting Distance Coefficient.*

The larger coefficient selected, the more iteration produced that may prolonged the learning process and increase the data space consumption. In contrast, adjusting D to zero provides the quickest learning period and the most efficient data storage. However, the suggested adjustment of the shifting distance coefficient is one due to the Javanese lexical rule that one word may be translated into two words in another speech level [14].Table 3, clarifies the result of applying the shifting distance coefficient to the parallel text alignment algorithm. In this example, the coefficient is altered to its optimum value: $D = 1$[14]. As a result, the aligned pair is reduced from 35 (Table.2) to 25 pairs. The iteration is limited without erasing the targeted pairs in bold. The reduction of the iteration causes the probability of the aimed pairs rise up to 0.12 (3/25). Therefore, edit shifting distance coefficient is applicable for limiting the number of potential pairs without reducing the alignment efficiency. As a result, the training process is faster and the data-storage consumption lower than before applying the algorithm [14].

Table 3: *Generated Pairs after Applying Edit Shifting Distance Algorithm.*

Pair	Combination	Total
1:1	(S1,T1),(S1,T2) (S2,T1),(S2,T2),**(S2,T3)** (S3,T2),(S3,T3),**(S3,T4)**	8
1:2	**(S1,T1T2)**,(S1,T2T3) (S2,T1T2),(S2,T2T3), (S2,T3T4) (S3,T2T3),(S3,T3T4)	7
2:1	(S1S2,T1),(S1S2,T2) (S2S3,T1),(S2S3,T2), (S2S3,T3)	5
2:2	(S1S2,T1T2),(S1S2,T2T3) (S2S3,T1T2),(S2S3,T2T3), (S2S3,T3T4)	5
		25

4.4 TranslationProcess

Sentences in the source language are used as an input. The input is parsed into smaller units such as words and phrases and then the system retrieves all possible pairs from the database.The Dice Coefficient (P) is used to measure the similarity between the source and target languages and is usually employed to detect similarity between vectors [38-40].$P(S,T)$ is a modified form of the Dice Coefficient based on probabilistic constraints and Javanese speech levels' modelling. The coefficient of each pair is compared to obtain the best translation (BT) by selecting the maximum value ($ArgMax$) of $P(S,T)$. Finally, the results are recombined into the translated sentence. $P(S,T)$ and BT is formulated in (9) and (10) respectively.

$$P(S,T) = \frac{2(P(S|L) \cap P(T|L))}{P(S|L) + P(T|L)} \quad (9)$$

$$BT = ArgMaxP(S,T) \quad (10)$$

4.5 Evaluation

The data are parallel texts of 1991 Javanese speech level classifications[4, 41]and used in training and evaluation process. They are created based on the review of literature as the availability of Javanese corpora on the web are insufficient[42, 43], mixed and not follow the standard [44-46]. Table 4provides several examples of created parallel texts with various lengths and complexity.

Table 4: *Examples of Javanese Parallel Texts.*

Level	1	2	3
Ng	*adikkum angan.*	*suketedipanga nsapimu.*	*telungdinaeng kas, kabehmangan ingomahmu.*
NgA	*adikkum angan.*	*suketedipanga nsapimu.*	*telungdinaeng kas, kabehdhahari ngdalemmu.*
Kr	*adikkula nedha.*	*rumputipundi puntedhalemb usampeyan.*	*tigangdintenm alih, sedayadhahar wontengriyas ampeyan.*
KrA	*adikkaw ulanedha.*	*rumputipundi puntedhalemb upanjenengan.*	*tigangdintenm alih, sedayadhahar wontendalem panjenengan.*
English	my brother is eating	the grass is eat by your cow	inthe next three days, all of us will eat at your home.

The texts have 126 sentences for each language. However, as shown in Table 5, the difference in the percentages of words (%W) and phrases (%Ph) shows that the quantity of words (#word) and phrase (#ph) that form the sentence is unequal despite the balanced training sentences.

Table 5: *Statistics of Words and Phrases.*

Level	#word	diffW	%W	#ph	diffph	%Ph
Ng	1704	98	23.8	1172	180	23.1
NgA	1708	101	23.8	1176	190	23.1
Kr	1876	83	26.2	1344	177	26.5
KrA	1880	91	26.2	1388	186	27.3
Total	7168	373	100.0	5080	733	100.0

The excellence of the speech levels' translation is measured by using two formulas. Firstly, the accuracy (A) indicates only the number of perfect translations that are lexically and pragmatically correct within the testing data as show in (11). The second indicator is related to the quality of translation as shown in (12).The indicator is obtained by classifying the similarity between expected translation and the result of the translation into several categories, as shown in Table 6. The last stage is linking the calculated indicators with the category of teenagers' competence in understanding and using Javanese: 81 to 100 (very good), 71 to 80 (good), 61 to 70 (fair), 51 to 60 (poor), 0 to 50 (very poor) [8].

$$A = \frac{\#perfect\_translation}{\#testing\_data} \times 100 \quad (11)$$

$$Q = \frac{((100 \times A) + (75 \times B) + (50 \times C) + (25 \times D))}{0.01(\#testing_{d}ata)} \quad (12)$$

5. TRANSLATION PERFORMANCE

Javanese are expected to use higher speech levels to address those of higher social status, and apply lower level language to those of lower social status. To capture the pragmatic conditions that pertain, the results of translation are detailed into translations from lower to higher speech levels (Table 7) and vice versa (Table 8).

Table 6: *The Classification of the Quality of Translation.*

	Category	score	Expected	Result
A	Pragmatically and lexically correct	100	*Mbah putri dhahar.*	*Mbah putri dhahar.*
B	Pragmatically correct with 25% lexically mistake (e.g. incomplete sentence).	75	*Mbah putrid kula dhahar.*	*Mbah putri kula.*
C	Pragmatically correct with 50% lexically mistake (e.g. incorrect alignment).	50	*Mangga dipundhahar pisang menika.*	*Mangga dipundha har dipundha har pisang.*
D	Pragmatically correct with 75% lexically mistake (e.g. incorrect alignment).	25	*Iku panganen.*	*Iku panganen dipangan.*
E	Pragmatically incorrect	0	*Mbah putri dhahar.*	*Mbah putrid nedha.*

Table 7: *Translation from Lower to Higher Speech Level.*

No	**Source**	**Target**	**A(%)**	**Q**
1	Ng	KrA	69.0	78.6
2	Ng	Kr	69.8	79.0
3	NgA	KrA	70.6	86.7
4	NgA	Kr	88.1	94.6
5	Ng	NgA	89.7	89.7
6	Kr	KrA	96.8	98.6

Table 7 shows that the lowest A (69%) and Q

(78.6) occurred when translating Ng into KrA. Despite an acceptable result, the incorrect pragmatic translation contributes to the reduction of the accuracy; KrA use mixture of *krama* and *kramainggil* vocabularies.

Table 8: *Translation from Higher to Lower Speech Level.*

No	Source	Target	A(%)	Q
1	Kr	NgA	77.0	89.5
2	KrA	NgA	76.2	87.9
3	Kr	Ng	77.8	89.5
4	KrA	Ng	77.8	89.3
5	KrA	Kr	99.2	99.6
6	NgA	Ng	100.0	100.0

Both accuracy and quality of translation detailed in Table 8 is always better than its opposite direction. While translating Kr to NgA, a word- repetition mistake, as shown in Fig.7, occurs in translation because of improper alignment. An algorithm to reduce the duplication should be developed in order to address this problem.However, there is no mistake in NgA-Ng translation since pragmatic vocabulary items are translated into common words. For example, both *dhaharandnedhaare* translated into *mangan* in *ngoko*.

Bapak saweg dhahar(KR)
Bapak lagi dhahar(NgA)
father-SL PROG eat
'Fatheris eating'
Bapak lagi lagidhahar(duplication error)
father-SL PROG PROG eat

Fig.7: *Example of Word Repetition Mistake.*

6. EFFORT TO IMPROVE THE TRANSLATION PERFORMANCE

6.1 Increasing the Training Data

The developed system can be used as a Javanese translation tool, even though some mistakes are identified. The accuracy of the developed corpus-based machine translation can be increased by increasing the amount of training data [12]. By increasing the amount of training data the probability of the correct pairs is expected to be higher. Thus, it will improve the accuracy.Therefore, an experiment should be conducted to indicate the influence of additional training data the translation accuracy (A) and (Q). The shifting distance is adjusted to its optimum value, D=1 [14] and the total of testing data is 504 sentences [12] to keep the experiment consistency. The number of training data is different in every experiment scenario; they are 504 (TR1), 824 (TR2) and 1144 (TR3) sentences.

The experimental result is detailed in Table 9 and Table 10 that clearly reveals the more data trained the better results obtained. The most performance is increased because of the training data used in TR2 and TR3. Those sentences have similar structure with TR1 sentences. Each of scenario used specific verb to differentiate another part of sentences implicitly. The probability of the corresponded pairs may increase during the training process since new pairs generated. As a result, the translation errors can be reduced since the correct targeted pair becomes more probable.

Table 9: *Translation Accuracy Quality by Increasing the Amount of the Training Data.*

No	Translation	Accuracy		
		TR1	TR2	TR3
1	Ng-NgA	89.7	89.7	90.5
2	Ng-Kr	69.8	69.8	78.6
3	Ng-KrA	69.0	69.0	78.6
4	NgA-Ng	100.0	100.0	100.0
5	NgA-Kr	88.1	88.9	89.7
6	NgA-KrA	70.6	71.4	79.4
7	Kr-Ng	77.8	78.6	85.7
8	Kr-NgA	77.0	77.0	84.1
9	Kr-KrA	96.8	97.6	98.4
10	KrA-Ng	77.8	79.4	85.7
11	KrA-NgA	76.2	77.0	84.9
12	KrA-Kr	99.2	99.2	99.2
Average		82.7	83.1	87.9

Table 10: *Translation Quality by Increasing the Amount of the Training Data.*

No	Translation	Quality		
		TR1	TR2	TR3
1	Ng-NgA	89.7	89.7	90.5
2	Ng-Kr	79.0	79.0	82.7
3	Ng-KrA	78.6	78.6	83.3
4	NgA-Ng	100.0	100.0	100.0
5	NgA-Kr	94.6	94.8	95.2
6	NgA-KrA	86.7	87.1	89.7
7	Kr-Ng	89.5	89.1	92.9
8	Kr-NgA	89.5	88.9	92.5
9	Kr-KrA	98.6	98.8	99.0
10	KrA-Ng	89.3	90.1	93.5
11	KrA-NgA	87.9	88.3	93.1
12	KrA-Kr	99.6	99.6	99.6
Average		90.2	90.3	92.7

6.2 Extended Alignment Algorithm Based on Impossible Pair Limitation

Modifying the alignment algorithm with edit shifting distance coefficient is obviously more efficient than the fundamental procedure. However, some alignment mistakes are detected due to the inadequate data training. The optimized alignment still records the unwanted pairs, called impossible pairs. The system may select the error instead of the targeted pair if it is listed before or more probable than exact one.

There are impossible conditions in every pair combination. For example, it is unnecessary to align s_i with t_j when $i = 1$ and $j > 1$ in (1:1) combination since it will leave t1 unpaired. Therefore, a mechanism is needed to align the initial word in source language ($s1$) with only the first in the target language ($t1$). Fig.8, depicts four impossible situations should be considered for only (1:1) pair alignment. Similar impossible situation for other combinations is obtained then detailed in Table 11.

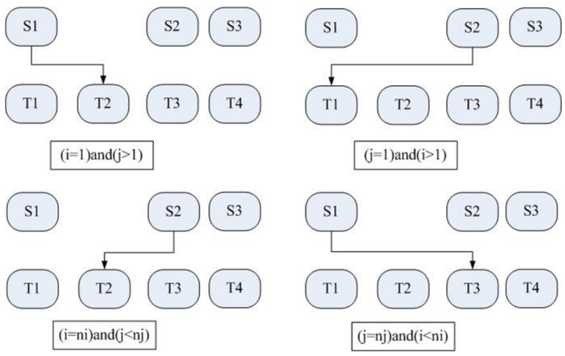


Fig.8: *Impossible Situation of (1:1) Pair Combination.*

Table 11: *Impossible Condition of Javanese Bi-text Alignment.*

Pair	Source(si)	Target(tj)
1:1	i=1	j>1
	i>1	j=1
	i=ni	j<nj
	i<ni	j>1
1:2	i=1	j>1
	j=1	j>1
	i=ni-1	j<nj
2:2	i>1	j=1
	i=1	j>1
	i<ni	j=nj-1
2:2	i=1	j>1
	i>1	j>1
	i=ni-1 i<ni-1	j<nj-1 j=nj-1

i=1:1st order in S; ni: numbers of words in S;
j=1:1st order in T; nj: numbers of words in T

The algorithm is not much different with the previous alignment, except the impossible pair consideration. As seen in Fig.9, the procedure after applying shifting distance coefficient is checking the alignment possibility. The word or phrase (chunk) will be aligned with its pair if they pass the impossible pair condition. The rest process checks the availability of the aligned chunk, as well as updates its frequency in the database.

for each sentence in source and target language
for j:= i - D to i + D do
check possibility
if possible then
train all possible pair combination
check the database
if the combination is unavailable in database then
record the pair combination with its frequency
else update the frequency of the pair

Fig.9: *Extended Parallel Tex Alignment based on Impossible Pair Limitation.*

The example in Fig.5 is again used to justify the performance of the extended algorithm. The generated pairs in Table 12 are reduced into 40% of the first model (Table 2) and 56 % of the second model (Table 3). This extended algorithm still captures the three targeted alignment as well as reduces more irrelevant pairs. As a result, the probability of this example case is 0.21(3/14) that overcomes all prior developed approaches.

Table 12: *Generated Pairs of the Extended Alignment Algorithm.*

Pair	Combination	Total
1:1	(S1,T1) (S2,T2),(**S2,T3**) (S3,T3),(**S3,T4**)	5
1:2	(**S1,T1T2**),(S1,T2T3) (S2,T2T3) (S3,T3T4)	4
2:1	(S1S2,T1) (S2S3,T2)	2
2:2	(S1S2,T1T2) (S2S3,T2T3), (S2S3,T3T4)	3
		14

The improved algorithm is then tested using the same parameter applied in previous evaluation. The number of training data (TR4) is equal to TR1: 504 sentences. Table 13 shows the translation performance which is better than all previous scenarios. Total five translation directions reach maximum accuracy (100%); they are NgA-Ng, NgA-Kr, NgA-KrA, Kr-KrA, KrA-Kr. This phenomenon illustrates that the reduction of impossible pairs successfully increases the translation accuracy and quality. Another efficiency indicator is that the training period using this method (t=5823s) is about three times faster than TR3 (t=15406s).

Table 13: *The Translation Performance Using Impossible Pair Limitation.*

No	Translation	TR4	
		A	Q
1	Ng-NgA	89.7	89.7
2	Ng-Kr	81.7	82.5
3	Ng-KrA	83.3	83.7
4	NgA-Ng	100.0	100.0
5	NgA-Kr	100.0	100.0
6	NgA-KrA	100.0	100.0
7	Kr-Ng	90.5	95.8
8	Kr-NgA	92.1	96.6
9	Kr-KrA	100.0	100.0
10	KrA-Ng	92.1	96.7
11	KrA-NgA	96.0	98.0
12	KrA-Kr	100.0	100.0
Average		93.8	95.3

7. CONCLUSION AND FUTURE DEVELOPMENT

The hybrid corpus-based machine translation for Javanese language is developed to deal with the complexity of politeness expression and speech levelsof Javanese. Based on [8], the performance of the machine translation (MT) can be categorized as very good translation. Translation mistakes are identified and corrected by increasing the quantity of training data.

Another approach to create more precise MT is by improving the training algorithm. The modification reduces the impossible pairs in means of increasing the correct targeted pair probability. The experimental results show that the proposed improvement totally more accurate and faster than the basic algorithm.

As youth do not know which forms of Javanese to use and under which circumstances [8], further developments will involve embedding pragmatic rules to govern the appropriate use of Javanese. Here, the system will guide the user choosing the proper language based on the interlocutor's social status, age and relationship with the speaker.

References

[1] G. Poedjosoedarmo, "The effect of Bahasa Indonesia as a lingua franca on the Javanese system of speech levels and their functions," *International Journal of the Sociology of Language*, vol. 177, pp. 111-121, 2006.

[2] S. Poedjosoedarmo, "Javanese Speech Levels," *Indonesia*, pp. 54-81, 1968.

[3] R. M. Kuncaraningrat and P. Southeast Asian Studies, *Javanese culture*. Singapore: Oxford University Press, 1989.

[4] S. Wibawa, "Efforts to maintain and develop Javanese language politeness," in *International Seminar of Javanese Language*, Paramaribo,Suriname, 2005, pp. 1-10.

[5] W. Wedhawati, W. E. S. Nurlina, E. Setiyanto, and R. Sukesti, *Latest structure of Javanese language*. Yogyakarta: Kanisius, 2006.

[6] G. Quinn, "Teaching Javanese Respect Usage to Foreign Learners," *Electronic Journal of Foreign Language Teaching*, vol. 8, pp. 362-370, 2011.

[7] N. J. Smith-Hefner, "Language Shift, Gender, and Ideologies of Modernity in Central Java, Indonesia," *Journal of Linguistic Anthropology*, vol. 19, pp. 57-77, 2009.

[8] D. E. Subroto, M. D. Rahardjo, and B. Setiawan, "Endangered krama and krama Inggil varieties of the Javanese language," *Linguistik Indonesia*, vol. 26, pp. 89-96, 2008.

[9] S. Suwadji, "Javanese language today," in *Lokakarya Pengajaran Bahasa dan Sastra Jawa*, Yogyakarta, 1996, pp. 55-61.

[10] S. Riyadi, "Development policy of Javanese language and literature and its application in junior high school," in *Lokakarya Pengajaran Bahasa dan Sastra Jawa*, Yogyakarta, 1996, pp. 31-39.

[11] F. Nugrahani, "Reactualisation of Javanese language and literature learning in multicultural era," *Varia Pendidikan*, vol. 20, pp. 70-80, 2008.

[12] A. P. Wibawa, A. Nafalski, N. Murray, A. E. Kadarisman, and J. Tweedale, "Hybrid machine translation for Javanese speech levels," in *5th International Conference on Knowledge and Smart Technology (KST)*, Burapha,Thailand, 2013, pp. 64-69.

[13] A. P. Wibawa, A. Nafalski, A. E. Kadarisman, and W. F. Mahmudy, "Indonesian-to-Javanese Machine Translation," *International Journal of Innovation, Management and Technology (IJIMT)*, vol. 4, pp. 451-454, August 2013.

[14] A. P. Wibawa, A. Nafalski, N. Murray, and W. F. Mahmudy, "Edit Distance Algorithm To Increase Storage Efficiency Of Javanese Corpora," in *International Conference on Computer, Electrical, and Systems Sciences, and Engineering (ICCESSE)*, Singapore, 2012, pp. 1056-1060.

[15] P. Purwadi, M. Mahmudi, and E. Setijaningrum, *Javanese language structure*. Yogyakarta: Media Abadi, 2005.

[16] A B. Setiyanto, Parama Satra: *Javanese Language*. Yogyakarta: Panji Pustaka, 2010.

[17] Sukarno, "The Reflection of the Javanese Cultural Concepts in the Politeness of Javanese," *kta*, vol. 12, pp. 59-71, 2010.

[18] W. J. Raynor, *The international dictionary of artificial intelligence*. Chicago, Ill. :: Fitzroy Dearborn :, 1999.

[19] H. Somers, "Review Article: Example-based

Machine Translation," *Machine Translation*, vol. 14, pp. 113-157, 1999.

[20] J. Hutchins, "Example-based machine translation: a review and commentary," *Machine Translation*, vol. 19, pp. 197-211, 2005.

[21] R. Jain, R. M. K. Sinha, and A. Jain, "Role of examples in translation," in *Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century., IEEE International Conference on*, 1995, pp. 1615-1620 vol.2.

[22] W. JianDe, C. ZhaoXiong, and H. HeYan, "Intelligent Case Based Machine Translation System Computational Linguistics and Intelligent Text Processing." vol. 2004, A. Gelbukh, Ed., ed Berlin/Heidelberg: Springer 2001, pp. 197-205.

[23] A. Lopez, "Statistical machine translation," *ACM Comput. Surv.*, vol. 40, pp. 1-49, 2008.

[24] A. Way and N. Gough, "Comparing example-based and statistical machine translation," *Natural Language Engineering*, vol. 11, pp. 295-309, 2005.

[25] M. Khalilov and J. A. R. Fonollosa, "Syntax-based reordering for statistical machine translation," *Computer Speech & amp; Language*, vol. 25, pp. 761-788, 2011.

[26] T. Xiao, J. Zhu, and M. Zhu, "Language Modeling for Syntax-Based Machine Translation Using Tree Substitution Grammars: A Case Study on Chinese-English Translation," vol. 10, pp. 1-29, 2011.

[27] R. Zbib, M. Kayser, S. Matsoukas, J. Makhoul, H. Nader, H. Soliman, et al., "Methods for integrating rule-based and statistical systems for Arabic to English machine translation," *Machine Translation*, vol. 26, pp. 67-83, 2012/03/01 2012.

[28] L. R. Nair and D. Peter, "Machine Translation Systems for Indian Languages," *International Journal of Computer Applications*, vol. 39, pp. 24 - 31, 2012.

[29] S. Le, Z. Yibo, Z. Junlin, and S. Yufang, "PECAT: a computer-aided translation tool based on bilingual corpora," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, 2001, pp. 927-932 vol.2.

[30] J. Li and B. Wang, "The automatic extraction of translation patterns and matching algorithm in an English-Chinese machine translation," in *Natural Language Processing and Knowledge Engineering*, Wuhan, 2005, pp. 839-843.

[31] J. Zhao, F. Liu, and D. Liu, "Two-phase base noun phrase alignment in Chinese-English parallel corpora," in Natural Language Processing and Knowledge Engineering, Wuhan, 2005, pp. 360-365.

[32] V. M. D. Bilbao, J. G. P. Lopes, and T. Ildefonso, "Measuring the impact of cognates in parallel text alignment," in *Artificial intelligence*, 2005. epia 2005. portuguese conference on, 2005, pp. 338-343.

[33] M. V. Yazdchi and H. Faili, "Generating english-persian parallel corpus using an automatic anchor finding sentence aligner," in *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*, 2010, pp. 1-6.

[34] L. Ahrenberg, M. Andersson, and M. Merkel, "A simple hybrid aligner for generating lexical correspondences in parallel text," in *36th Annual Meetingof the Association for Computational Linguistics Montreal*, Quebec, Canada., 1998, pp. 29-35.

[35] F. Bond and K. Ogura, "Reference in Japanese-English Machine Translation," *Machine Translation*, vol. 13, pp. 107-134, 1998.

[36] R. Terashima, H. Echizen-ya, and K. Araki, "Learning method for extraction of partial correspondence from parallel corpus," in *International Conference on Asian Language Processing*, Singapore, 2009, pp. 293-298.

[37] A. v. d. Bosch and P. Berck, "Memory-Based Machine Translation and Language Modeling," *The Prague Bulletin of Mathematical Linguistics*, vol. 91, pp. 17-26, 2009.

[38] N. Anuar and A. B. M. Sultan, "Validate Conference Paper Using Dice Coefficient " Computer and Information Science vol. 3, pp. 139-145, 2010.

[39] J. Ye, "Multicriteria decision-making method using the Dice similarity measure based on the reduct intuitionistic fuzzy sets of interval-valued intuitionistic fuzzy sets," *Applied Mathematical Modelling*, vol. 36, pp. 4466-4472, 2012.

[40] L. Egghe, "Good properties of similarity measures and their complementarity," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 2151-2160, 2010.

[41] S. Sudaryanto, Ed., *Tata bahasa baku bahasa Jawa (standard grammar of Javanese)*. Surakarta: Duta Wacana University Press, 1992, p.^pp. Pages.

[42] Tembi.org. (2000, 1 September). *Pasinaon Basa Jawa*. Available: http://www.tembi.org/bjawa/index.htm

[43] Mylanguages.org. (2010, 1 September). *Learn Javanese*. Available: http://mylanguages.org/learn_javanese.php

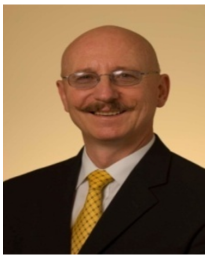
[44] (2009, 1 September). *Kamus Jawa Online*. Available: http://kamusjowo.com/

[45] S. Karti. (2009, 1 September). *Javanese-Indonesian-English Dictionary*. Available: http://kamusjawa.info/

[46] Wikimedia. (2010, 1 September). *Kamus Indonesia-Jawa*. Available: http://id.wiktionary.org/wiki/Kamus_Indonesia_%E2%80%93_Jawa

Aji P. Wibawa received his Bachelor of electrical engineering from Brawijaya University (2004) and Master of information technology and management from Institute Technology of Sepuluh Nopember (2007), both in Indonesia. He is currently a Lecturer at the Department of Electrical Engineering, State University of Malang (UM) and a PhD candidate at School of Engineering, University of South Australia. His current research interests are machine learning and computational linguistic.

He is one of Institute of Electrical and Electronics Engineers (IEEE) members since 2012.



Andrew Nafalski holds BEng(Hon), GradDipEd, MEng, PhD and DSc degrees. His career of several decades covers chronologically academic assignments in his native Poland, Austria, Slovak Republic, Japan, Germany, Wales, France, Australia, USA and Canada. His research interests include among others information technology, knowledge-based engineering, remote laboratories and innovative engineering education.

He has published some 32 books, monographs, book chapters and software sets, 100 journal papers and 215 conference papers. He is currently a Professor of Electrical Engineering at the University of South Australia in Adelaide.

Wayan F. Mahmudy obtained bachelor degree in Mathematics from Brawijaya University, Indonesia in 1995 and master degree in Information Technology from Institut Teknologi Sepuluh November (ITS), Indonesia in 1999. He is a Lecturer at Department of Computer Science, Brawijaya University (UB), Indonesia. Currently, he is a PhD candidate at School of Engineering, University of South Australia. His research interests include optimization of combinatorial problems and machine learning.