Indonesian-to-Javanese Machine Translation

Aji P. Wibawa, Andrew Nafalski, A. Effendi Kadarisman, and Wayan F. Mahmudy

***Abstract*—Javanese is a multi-level language; it comprises four speech levels used to convey local politeness. However, the negative tendency is detected regarding the use of Javanese speech levels among teenagers. They prefer to use the Indonesian national language (bahasa Indonesia, BI) because of the Javanese speech level complexity. A combination of statistical and memory-based machine translation is designed by the authors to help Javanese youths in translating between both languages. The evaluation shows that translating speech levels of Javanese to Indonesian is more accurate than translating in the opposite direction, as revealed by the average accuracy (A) of 0.83 for Javanese-Indonesian translation and 0.68 for the other direction.**

***Index Terms*—Indonesian, Javanese speech levels, machine translation.**

I. Introduction

Javanese is a compulsory subject in most primary and secondary schools in Central and East Java, Indonesia. Students must learn the language, including its speech levels [1] and the traditional alphabet (*hanacaraka*), for six years in primary school and another three years in secondary school [2]. Nowadays, the teaching of Javanese speech levels is focused on translating the basic speech level (*ngoko*) into the high speech level (*krama*) as well as memorising low-frequency words used in speech and writing [2], [3].

Students also have to learn two other languages, Indonesian as the instructional language in schools and English as the primary foreign language [3]. As a result, Javanese is getting more rarely used outside the class. While speaking at home and in the community, Javanese students in urban areas prefer to use Indonesian, the mono-level national language, which they can handle more easily and they believe to be more reliable to use in the global era [2], [4] and [5]. Furthermore, in verbal communication among speakers of equal social status, the use of *ngoko* or *bahasa Indonesia* is preferable, with the former indicating intimacy. However, among speakers of different age and/or social status, the use of the polite *krama* variant is called for. Realizing that they cannot handle this polite form, younger speakers usually switch into Indonesian. If this linguistic attitude and communicative act keep going on, the *krama* form—a unique characteristic Javanese—is in the danger of diminishing.

To prevent Javanese speech levels from fading away, a serious effort, such as creating a machine translation, is badly needed. A machine translation is a software-based system that translates one language to another. Fundamentally, the approach can be rule-based machine translation (RBMT) or corpus-based machine translation (CBMT). The RBMT is based on linguistic information—semantic, morphological, and syntactic. On the other hand, CBMT is derived from large bilingual texts and is divided into example-based machine translation (EBMT) and statistical machine translation (SMT) [6], [7]. This paper is focused on developing a multilingual machine translation which combines the EBMT and SMT approaches. The developed-machine translation is expected to help students in translating Indonesian to one of the four Javanese speech levels or *vice versa*.

II. Language Modelling and Database Design

In Tata Bahasa Baku Bahasa Jawa (Standard Grammar of Javanese, 1992), the speech levels have been simplified to four categories: ngoko (Ng), ngoko alus (NgA), krama (Kr) and krama alus (KrA) [3], [8]. Translating Javanese word by word into Indonesian (BI) often results in failures due to unequal quantity of words constituting the source and target sentences [9]. For instance, as shown in Table I, the Indonesian sentence *Sapi saya dibeli isterimu* (BI), 'My cow was bought by your wife', may be translated into four different versions of Javanese, according to its speech levels.

TABLE I: Example of Indonesian to Javanese Translation

Sapi saya	*dibeli*	*isteri-mu*	(BI)
Sapi-ku	*dituku*	*bojo-mu*	(Ng)
Sapi-ku	*dipundhut*	*garwa panjenengan*	(NgA)
Lembu kula	*ditumbas*	*bojo sampeyan*	(Kr)
Lembu kula	*dipunpundhut*	*garwa panjenengan*	(KrA)
Cow-my	**buy-PSV**	**wife-your**	

'My cow was bought by your wife'

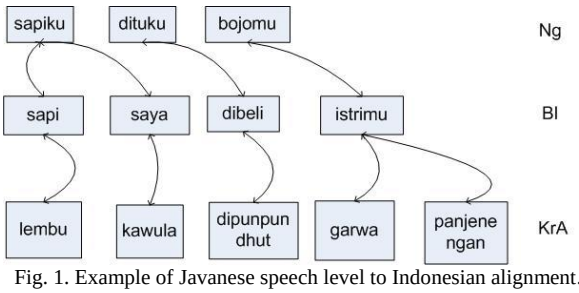

Fig. 1. Example of Javanese speech level to Indonesian alignment.

Literally, the number of words in every sentence is not equal: BI = 4; Ng = 3; NgA, Kr, KrA = 5. Therefore, it is impossible to align a single word with its equivalent in another language. The correct translation is produced by aligning both word and phrase combinations. Fig. 1

Manuscript received November 4, 2012; revised February 17, 2013.

Aji P. Wibawa is with the School of Engineering, University of South Australia, Adelaide Australia and Department of Electrical Engineering, State University of Malang, Malang, Indonesia (e-mail: aji.wibawa@mymail.unisa.edu.au; ajipw@um.ac.id).

Andrew Nafalski is with the School of Engineering, University of South Australia, Adelaide Australia (e-mail: Andrew.nafalski@unisa.edu.au).

A. Effendi Kadarisman is with the English Department, State University of Malang, Malang, Indonesia (e-mail: effendi_kadarisman@gmail.com).

Wayan F. Mahmudy is with the Department of Computer Science, Brawijaya University, Malang, Indonesia (e-mail: wayanfm @ub.ac.id).

illustrates the proper alignment of Javanese speech levels to Indonesian.

The basic sentence structure of Javanese and Indonesian is the same: SVO (subject, verb, object), as illustrated by the Indonesian and Javanese sentences in Table II. However, the meaning of the Javanese sentences is pragmatically different, relative to the subject-verb agreement that derives from non-linguistic factors (i.e. social status, age, and personal relationship) [10]. For example, sentences (1) and (2) in Table II illustrate the use of a specific verb (eating) based on the subjects' or agents' difference in social status.

TABLE II: EXAMPLE OF THE USE OF VERB BASED ON THE DIFFERENCE OF SOCIAL STATUS

Murid-murid	sedang	makan (BI)
Murid-murid	saweg	nedha (Kr)
Student-PL	**PROG**	**eat**
'The students are eating'		
Guru-guru	sedang	makan (BI)
Guru-guru	nembe	dhahar (KrA)
teacher-PL	**PROG**	**eat**
'The teachers are eating'		

The Indonesian word *makan* 'eat' in examples (1) and (2) is translated differently into Javanese: *nedha* (Kr) and *dhahar* (KrA). This is due to the fact that that agent of *nedha* (Kr) is *murid-murid* 'students' whereas that of *dhahar* (KrA) is *guru-guru* 'teachers'. Obviously, the latter has a higher social status than the former, and hence the contrast between KrA and Kr. On the other hand, Indonesian applies an identical verb in both examples, suggesting that it does not require a pragmatic subject-verb agreement. Therefore, the bilingual text alignment should be modelled based on this consideration.

Pair combination of Javanese text (1) is modelled to accommodate the complicated characteristics of the speech levels and is then divided into two categories: lexical and pragmatic combinations [9]. The lexical combination consists of 3 pairs: one word to one word (1:1), one word to two words (1:2), and two words to one word (2:1). The pragmatic combination is achieved by aligning through (2:2) combinations that refer to Javanese subject-verb agreement [9].

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

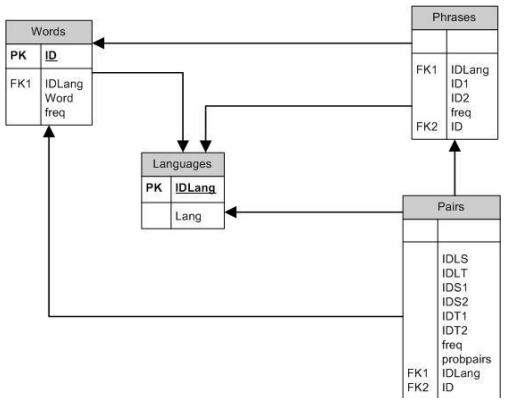

Fig. 2. The database of five languages.

The multilingual machine translation uses a database to record both linguistic and statistical data. The relational database consists of four tables: tables of Languages, Words, Phrases, and Pairs (Fig. 2), which record linguistic and statistical data of inputted parallel texts. Since it is a relational database, any editing process applied to the tables may change the records in other Tables.

III. LEARNING AND TRANSLATION PROCESS

The unsupervised learning process is divided into two stages: the monolingual text parsing process and the bilingual alignment process. At the monolingual stage, every sentence is split into set of words. The database (Fig. 2) may record and automatically index the result as well as the frequency of related words. The next stage restructures the sentence into a monolingual array which consists of unique numbers representing the words' indexes in order to create a faster alignment process.

The bilingual text alignment procedure basically aligns the array of sentences. The method aligns every possible combination of S with T that starts from the first word through the end of the bilingual sentence. Similar to the monolingual process, the combination pairs and their frequencies are subsequently sent to the database (Fig. 2). The shifting distance algorithm (D) is implemented in order to reduce the number of irrelevant pairing iterations and to increase the efficiency of data-storage consumption [9]. Overall, the procedure of the optimised bi-text alignment is summarized in the form of pseudo code in Fig. 3.

```
for each sentence in source and target language
    for c:= 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
            end if
    end for
end for
```

Fig. 3. Procedure of the optimised bi-text alignment.

Sentences are used as the input in the source language and then parsed into words and phrases. The next stage is retrieving all possible pairs from the database based on the parsed source language. The similarity between source (S) and target (T) languages is measured by Dice Coefficient [11]-[13] that is modified based on conditional probability of Javanese speech levels. The best translation (BT) is then obtained by selecting the maximum value (ArgMax) of all corresponding pairs. Finally, the selected BT is recombined into the translated sentence.

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

$$BT=ArgMaxP(S,T) \quad (3)$$

P(S|L) and P(T|L) refer to the conditional probability of source and target texts in a specific language. Fig. 4 shows the described translation process in the form of an algorithm.

```
//Function Dice coefficient
 begin
  P(S|L):= freq of source pair/number of specific language
  P(T|L):= freq of target pair/number of specific language
  Result: = 2( P(S|L)* P(T|L))/ 2( P(S|L)+ P(T|L))
 end
-------------------------------------------------------------------------
//Procedure BT calculation
  begin
   BT:=0;
    for idL:=1 to 5 do //each language
     begin
       execute Dice Coefficient ()
      if result>BT then
        begin
         BT:=result
        end if
     end for
  end
-------------------------------------------------------------------------
// Procedure Translation
   begin
          Parse source language
          Procedure BT calculation
          Arrange target sentence based on BT
   end
```

Fig. 4. The translation algorithm.

IV. Testing and Results

Both training and testing data is created based on literature review since the most of available Javanese corpora on the web [14]-[18] are insufficient and not in accord with the Standard Grammar of Javanese (1992). The accuracy (*A*) measures the efficiency of the speech level translation which compares the number of correct translations with all testing data.

$$A = \frac{\sum correct\_translations}{\sum testing\_data} \quad (4)$$

As illustrated in Fig. 5, the lowest *A* (0.55) occurs when translating BI into KrA. Both *dhahar* (KrA) and *nedha* (Kr), 'eating', are translated into *makan* in BI; but ambiguity might happen when translating into reverse direction. As a result, translating various speech levels to BI is always better than translating into the other direction because of the mono-level characteristic of the language. Training more parallel texts with various patterns and complexities may increase the intelligence of the machine translation.

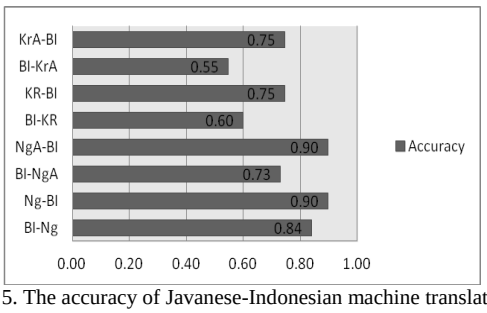

Fig. 5. The accuracy of Javanese-Indonesian machine translation.

The translation from both Ng and NgA into BI are the best among others. However, a word- repetition mistake, as shown in Table III, occurs in translation because of improper alignment. An algorithm to reduce the duplication should be developed in order to address this problem.

TABLE III: Example of Word Repetition Mistake

Ayah	sedang	makan (BI)
Bapak	lagi	dhahar (correct NgA)
teacher-PL	**PROG**	**eat**
'The teachers are eating'		
Bapak	lagi	lagi dhahar (duplication error)
teacher-PL	**PROG**	**PROG eat**

V. Conclusions

The developed system can be used as a Javanese translation tool, even though some mistakes are identified. The average accuracy of overall translation results (*A*=0.75) indicates that the developed translation is reasonably efficient. The accuracy of the developed corpus-based machine translation can be increased by increasing the amount of training data.

The satisfactory results of testing this Indonesian-Javanese machine translation are meant to be a preliminary attempt to help preserve Javanese all along with its speech levels and pull it away from its current status as an endangered local language. Moreover, since there are other multi-level local languages in Indonesia with the same endangered status, it is possible—with respect to a similar vein for future research—to further develop a bilingual machine translation with the purpose of saving any other local language.

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, no. 1, pp. 111-121, 2006.
[2] N. J. Smith-Hefner, "Language shift, gender, and ideologies of modernity in central java, Indonesia," *Journal of Linguistic Anthropology*, vol. 19, no. 1, pp. 57-77, 2009.
[3] S. Wibawa, "Efforts to maintain and develop Javanese language politeness," in *Proc. International Seminar of Javanese Language*, Paramaribo,Suriname, 2005, pp. 1-10.
[4] S. Suwadji, "Javanese language today," *Lokakarya Pengajaran Bahasa dan Sastra Jawa*, Yogyakarta, 1996, pp. 55-61.
[5] D. E. Subroto *et al.*, "Endangered krama and krama Inggil varieties of the Javanese language," *Linguistik Indonesia*, vol. 26, no. 1, pp. 89-96, 2008.
[6] J. Hutchins, "Example-based machine translation: a review and commentary," *Machine Translation*, vol. 19, no. 3, pp. 197-211, 2005.
[7] H. Somers, "Review article: example-based machine translation," *Machine Translation*, vol. 14, no. 2, pp. 113-157, 1999.
[8] S. Sudaryanto, Ed., *Tata bahasa baku bahasa Jawa (standard grammar of Javanese)*, Surakarta: Duta Wacana University Press, 1992.
[9] A. P. Wibawa, A. Nafalski, N. Murray, and W. F. Mahmudy, "Edit distance algorithm to increase storage efficiency of javanese corpora," presented at the International Conference on Computer, Electrical, and Systems Sciences, and Engineering (ICCESSE), Singapore, 12-13 September, 2012.
[10] Sukarno, "The reflection of the javanese cultural concepts in the politeness of javanese," *k@ta*, vol. 12, no. 1, pp. 59-71, 2010.
[11] N. Anuar and A. B. M. Sultan, "Validate conference paper using dice coefficient, " *Computer and Information Science*, vol. 3, no. 3, pp. 139-145, 2010.
[12] 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, no. 9, pp. 4466-4472, 2012.
[13] L. Egghe, "Good properties of similarity measures and their complementarity," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 10, pp. 2151-2160, 2010.
[14] Tembi.org. (1 September, 2000). *Pasinaon Basa Jawa*. Available: http://www.tembi.org/bjawa/index.htm
[15] Mylanguages.org. (1 September, 2010). *Learn Javanese*. Available: http://mylanguages.org/learn_javanese.php

[16] Kamus Jawa Online. (1 September, 2009). [Online]. Available: http://kamusjowo.com/
[17] S. Karti. (1 September, 2009). *Javanese-Indonesian-English Dictionary*. [Online]. Available: http://kamusjawa.info/
[18] Wikimedia. (1 September, 2010). *Kamus Indonesia–Jawa*. [Online]. Available: http://id.wiktionary.org/wiki/Kamus_Indonesia_%E2%80%93_Jawa

Aji Prasetya Wibawa was born in Malang, East Java, Indonesia in 1979. He 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. Mr. Wibawa 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 electromagnetism, electrical power engineering, 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.

A. Effendi Kadarisman earned his M.A. and Ph.D. degrees in Linguistics from the University of Hawaii at Manoa. His research interests include linguistic universality, linguistic relativity, ethnopoetics, and the application of the first two in the areas of TEFL (Teaching English as a Foreign Language) and SLA (Second Language Acquisition). He has presented papers in the regional, national, and international seminars and conferences. Most of his papers have been published in two professional journals in Indonesia: Linguistik Indonesia and TEFLIN (Teachers of English as a Foreign Language in Indonesia), and then compiled into a book, Mengurai Bahasa, Menyibak Budaya (Analysing Languages, Peering into the Cultures). He is currently a Senior Lecturer of Linguistics at the English Department and School of Graduate Studies, State University of Malang, Indonesia.

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.