Enriching English Into Sundanese and Javanese Translation List Using Pivot Language

Arie A. Suryani[1], Isye Arieshanti[2], Banu W. Yohanes[3]
SoC[1], Teknik Informatika[2], FTEK[3]
Telkom University[1], ITS[2], Univ. Kristen Satya Wacana[3]
Bandung[1], Surabaya[2], Salatiga[3]

M. Subair[4], Sari D. Budiwati[5], Bagus S. Rintyarna[6]
Pulse Lab Jakarta[4], SoAS[5], Informatics Dept[6]
Global Pulse[4], Telkom University[5], UNMUH[6]
Jakarta[4], Bandung[5], Jember[6]

***Abstract*—This paper discusses the problem of sparse translation of English into Sundanese and Javanese that were found in Translator-Gator. Translator-Gator is a language game created by the United Nation Global Pulse, to support the research initiatives in Indonesia. Thousands of keyword were generated and translated from English into some Indonesian local languages using the crowd resource. Unfortunately, many English words are still has no translation in Javanese as well as Sundanese. To overcome this problem we propose a technique to fill the un-translated English words in Javanese and Sundanese using Indonesian translation as a pivot language. Evaluation was made by manually investigated whether each phrase results a proper translation. Experiment shows that our technique results relatively low translation accuracy. Limited coverage of phrase translation list and ambiguous words are identified as causes of translations errors in our technique.**

Keywords—pivot language; translation weight; phrase translation.

I. INTRODUCTION

Parallel corpus is a collection of text in one language and their equivalent translation to other language. In machine translation research area, some language pairs contain a large number of parallel corpus are easy to obtain and ready to use. Conversely, for many languages pairs with a low resources language, there only a few of parallel corpus in small scale or even not found at all. The sparse of parallel corpus directly will result to a poor translation.

Similar problem faced by the Translator-Gator. Translator-Gator is an online language game created by Pulse Lab Jakarta. It was built to collect a large number of keyword related to some social, cultural, educational, and environmental problems. These keywords were firstly defined and translated into Indonesian by using the Google-Translate. These keywords were then translated into some Indonesian local language, such as Sundanese, Javanese, Bugenese, and Minangnese language by the crowd. To attract many people to translate actively, the Translator-Gator was represented as an online game, thus there was a reward and penalty. People will get some points when their translation was agreed by other (vote-up), otherwise they will lose their points (vote-down) as well as can be banned from continue playing at certain limit. A certain number of accumulative points were then can be redeem with a phone cell credit. In further, these translated keywords will be used to disseminate crucial information of food resilience, global warming, public health, as well as to be used by researches in computational linguistics and some related areas.

To date, the Translator-Gator collected more than one thousand and six hundreds of keywords. These keywords either can be a single word or a phrase contains more than one word. All of the keywords translated into Indonesian completely. Unfortunately, only 80% and 20% of these keywords were translated into the Javanese and the Sundanese respectively (Riyadi and Amin 2016). Therefore, we proposed a technique to enrich the translation list of English into Javanese as well as Sundanese using Indonesian as a pivot language. A pivot language was being chosen as a solution because English and Indonesian local language pair are low resources, which contains a rare language resources such as the parallel corpus, dictionaries, and other language tools. We hypothesized that using the existing Translator-Gator data is a reasonable solution that can implemented immediately.

Our technique comprises of three sequential steps. Firstly, three pairs of translation terms are chosen. Those pairs are English-Indonesian translation, English-Javanese Translation, and English-Sundanese translation. Since one English term can be translated into several terms in Indonesian, Javanese and Sundanese, we choose only one translated term according to the weight. The weight calculation involve vote-up, vote-down and frequency of translated result. Secondly, Indonesian-Javanese and Indonesian-Sundanese dictionaries are generated using Moses Translation System (Koehn 2015). At last, a rule-based technique is employed to fill the un-translated English terms into Javanese and Sundanese terms. In the rule-based model, the process requires keyword label whether the word is a borrowed word or not. Thus for this step we employ a borrowed word list collected from the online resources. The translation result is then evaluated manually to observe how well our technique produces the translation.

On the next section we present the related work, short descriptions of Translator-Gator, enrichment techniques, detail our proposed technique, experiment results and finally enclosed with a conclusion.

II. RELATED WORK

Previous related researches are focus on the problem of text translation from one source language into a target

978-1-5090-1381-4/16/$31.00 ©2016 IEEE

language by using an intermediate (pivot) language. According to (Wu and Wang 2009), there are three different pivot translation techniques that are triangulation method, transfer method and synthetic method.

The first method trains the source-pivot and pivot-target translation model by using parallel corpus. Using these two model, the translation model of source-target is then induced. The triangular technique was used by (Cohn and Lapata 2007) to solve the small data size problem in English to French translation by using Dutch, Danish and Portuguese as an intermediate language. The Experiments show an improvement of translation results compare to the standard phrase-based translation. One of the problems raised in the triangular method is a very large resulted translation model and some phrase pair that might be not connected to each other because does not have the same pivot phrase (Cui, et al. 2015).

The second method is transfer method that translates a source into target text in two consecutive steps that are source to pivot and pivot to target translation. A sentence of a source language is firstly translated into N pivot sentences, and then each pivot sentences translated into M target sentences (Utiyama and Isahara 2007). The translation result selected by using a defined weighting mechanism. Experiment shows that the transfer method was inferior to the triangular method for an English-Germany translation by using French as intermediate (Utiyama and Isahara 2007).

The third method is synthetic method, which creates a new parallel corpus of source-target by translates pivot sentences into sources sentences using source-pivot translation as well as translates pivot sentences into target using pivot-target translation. This method applied by (Gispert and Mariño 2006) for Catalan-English using Spanish as pivot. The evaluation of this paper was performed by comparing the translation result with or without the synthetic method. The experiment shows the translation resulted were slightly inferior to the baseline. Other experiment was also employed by (Klementiev, et al. 2012) for English-Spanish translation. They use a large number of English and Spanish monolingual corpora and a small size of dictionary.

Generally, our proposed technique adopts the first idea. We attempt to create the pivot-target translation table by using the existing target-pivot translation list. The resulted translation table is then being used to translate source-target keywords that are still having empty translation.

III. PROPOSED TECHNIQUE

The goal of our technique is to fill the empty translation of Sundanese and Javanese for some English word in Translator Gator. The idea of our technique is to use English-Indonesian translation list as a pivot or as an intermediary to get the English-Sundanese and English-Javanese translations. The approach is basically a rule-based approach which is different with machine learning technique that rely on the training data (Sarno, et al. 2013). There are three sequential steps conducted to fill the empty translation of Sundanese or Javanese which are selecting English-Indonesian (EN-ID) translation pairs, creating ID-JW and ID-SU dictionary, and translating empty Javanese and Sundanese words. The Block Diagram of these steps shown in Figure 1.

The first step of our technique is selecting English-Indonesian (EN-ID) phrase pairs. A single English word in Translator Gator could be translated into many Indonesian words by some different users. Selecting English-Indonesian (EN-ID) phrase pairs was intended to choose only a good enough EN-ID translation pair. This step results a unique EN-ID translation list contains one-to-one English-Indonesian translation pairs. The selection was conducted based on the number of translation occurrence, the number of user that agreeing this translation (vote-up), and the number of disagree user (vote-down). For this purpose, we define a weighting formula to pick EN-ID translation pair as shown in (1).

$$\forall x : max(weight(y))$$
$$weight = \sum y + \sum voteUp_y - \sum voteDown_y \tag{1}$$

Given translation list consists of a number translation pair. For each translation pair x, we calculate the weight for each translation alternatives y. Selected translation pair was the one that has highest score among other translation pair. Whereas weight determined by frequency of each translation alternatives, then added by its number of vote-up and subtracted by its vote-down.

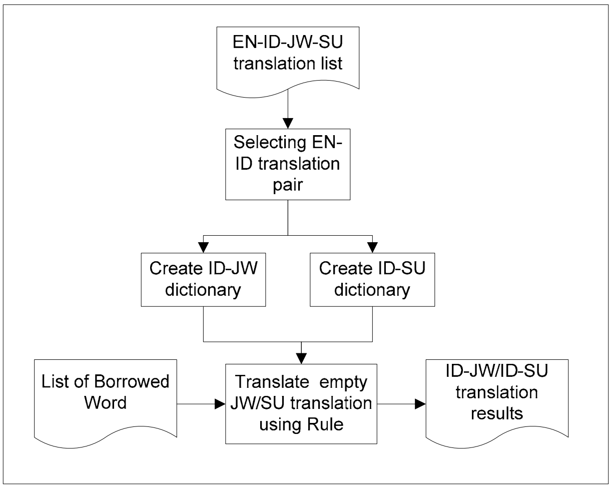
EN-ID-JW-SU translation list
Selecting EN-ID translation pair
Create ID-JW dictionary
Create ID-SU dictionary
List of Borrowed Word
Translate empty JW/SU translation using Rule
ID-JW/ID-SU translation results

Fig. 1. Our Proposed Technique Block Diagram

The second step is creating Indonesian-Javanese (ID-JW) and Indonesian-Sundanese (ID-SU) dictionary. To create this dictionary, previously we applied equation (1) to Translator Gator EN-JW and EN-SU translation list results the unique ID-JW and ID-SU translation list. After that, the new ID-JW and ID-SU translation list was created by joined the unique EN-ID translation pair resulted in step 1 with the unique ID-JW and ID-SU translation pair respectively. Unfortunately, the new ID-JW and ID-SU translation list were only contains phrase translation, whereas translation of either single word or combination of word that build the phrase are not covered by this list. Therefore, we do an enrichment of ID-JW and ID-SU translation list by assumed these lists as a parallel corpus and passed them into Moses translation system (Koehn 2015) to get its translation model as our final ID-JW and ID-SU dictionary.

In the last step, we fill the empty Javanese and Sundanese translation by using ID-JW and ID-SU dictionary resulted in step 2 and a defined translation rule. Generally, the translation rule covers two cases of translations, which are the case when the word to be translated was found in the dictionary and the other case is if it does not. In the first case, the corresponding Javanese or Sundanese word was simply given as the translation result. While in the second case, we need to check whether the word was a borrowed word or a phrase. If it was a borrowed word, then the Indonesian translation resulted as its translation. Otherwise, the phrase was break into N word and translate each word by using this translation rule iteratively.

TABLE I. TRANSLATION RULE

```
Rule #1:
if (keyword is found in phrase translation table) then
    return the translation result
else apply rule#2

Rule #2 :
if (keyword is a borrowed word) then
    Javanese or Sundanese =  Indonesian
else { keyword is not a borrowed word}
    if (keyword is a single word) then
        return "UNK" {UNK = unknown word}
    else {keyword is a phrase}
        split keyword into N words
        for each 1 until N word apply Rule 1-2
```

Translation result evaluation size was defined using Slovin formula (Almeda, T. Capistrano and G. Sarte 2010). The sample size was defined using the formula in (2), by using 5% error assumption e for a given N total population.

$$sample\ size = \frac{N}{1+Ne^2} \qquad (2)$$

IV. EXPERIMENT AND DISCUSSION

This part contains our experiment to observe the proposed technique performances, systematically started by description of experiment scenarios, dataset, experiment result and discussions.

Experiment Scenario and Dataset

There is only one experiment scenario that is to fill the Javanese or Sundanese translation for a number of keywords. The translation result was then evaluated manually to check their conformity. However, we also evaluate the ID-JW and ID-SU translation list resulted from applying equation 1, to ensure that this initial dictionary is quite good to be used as our initial dictionary.

In this experiment, we use 36.313 transactional data translation that translated by more than 100 Translator-Gator users. It consists of 1324 unique English keyword. By using these keywords we get 1340 pair of initial ID-JW dictionary and 460 pair initial ID-SU dictionary.

Experiment Results & Discussions

Subjective evaluation to initial ID-JW and ID-SU dictionary were applied manually to all the ID-SU dictionary entry, and 40% of ID-JW. Result of this evaluation shown in Table 2.

TABLE II. PHRASE PAIR EVALUATION

Language Pair	Number of Evaluated Data	Translated Properly (%)
Indonesian – Sundanese (ID-SU)	460 of 460 (100%)	70%
Indonesian – Javanese (ID-JW)	540 of 1340 (40%)	68%

According to Table 2, the ID-SU and ID-JW were quiet good to be used as our initial dictionary to translate empty Sundanese or Javanese translation.

In the initial ID-JW and ID-SU dictionary, some error translation were found, one of them was caused by improper translation of a borrowed word, such as the word "*insiden obesitas*" (obesity incidence) and "*es dunia*" (global ice) that should be translated using Indonesian, rather than produces improper Javanese or Sundanese translation. The other error was caused by the translation that filled improperly by the user. In this paper, we overcome the first cause error by copying the translation if the keyword is a borrowed word. The resulted phrase pair was then used to translate empty Sundanese and Javanese translation.

TABLE III. TRANSLATION RESULT EVALUATION

Rule Applied to the translation	Number of Evaluated Samples	Number of words translated properly
Indonesian – Javanese (ID-JW)		
Rule #1 (using phrase translation / dictionary)	76	58 of 76 (76.3%)
Rule #2	193	42 of 193 (22.6%)
Total ID-JW evaluated sample	269	
Indonesian – Sundanese (ID-SU)		
Rule #1 (using phrase translation / dictionary)	68	60 of 68 (88%)
Rule #2	229	78 of 229 (34%)
Total ID-SU evaluated samples	297	

The translation produced 269 Javanese phrase translation, consist of 76 translations generated using rule#1 and the other 193 translations resulted using rule#2. Whereas for Sundanese keywords, there were 1149 translations, comprises of 261 translations produced using rule#1, and 888 translations created using rule#2. It means that both ID-JW and ID-SU are majorly translated using rule#2. There are 77% of ID-SU and 71% of ID-JW were translated using rule #2, while only small portions of them are translated by using rule #1. It shows that our phrase translation list that was built using existing translator gator translation list covers less than 30% of the whole translation. Interestingly, we found that rule#1 gives better translation than rule#2 as depicts in Table 3. The translation results of the rule#1 achieve more than fifty percent for both Indonesian-Javenese and Indonesian-Sundanese. Whereas using rule#2 both language pairs give translation result less than 35%. It means that although our phrase translation list covers only small portion of keywords it gives significant results in translation.

In this experiment, we evaluate all translation results of ID-JW. While for ID-SU translation we used only some translation sample determined using Slovin formula of ID-SU since a large number translation thus its relatively hard to evaluate all of these translation manually.

Besides, we also analyze some translation error produced in this translation. Since the rule #2 raises translation error majorly, our error analysis focus more on the translation error produced by rule#2. In addition, the translation errors generated in rule#1 commonly are caused by the low coverage of the phrase translation. Overall, translation error occurs in applying rule#2 was caused by two factors. The first factor is the incomplete or limited coverage of phrase translation list, whereas the second factor is the existence of ambiguous word.

We present some examples of translation errors arise in rule#2 in Table 4.

TABLE IV. ERROR EXAMPLES PRODUCED IN TRANSLATION

Source Phrase	Target Phrase (Translation Results)	Occurs in	Error Description
menyogok pemilihnya (bought his votes)	UNK UNK (resulted no translation)	ID-SU	words "*menyogok*" and "*pemilihnya*" does not exist in phrase translation list
energy ramah lingkungan (clean energy)	*Energy* UNK *lingkungan*	ID-JW	word "*ramah*" does not exist in phrase translation list
Sistem kesehatan yang layak (decent health systems)	*Sistem kasugengan ingkang layak*	ID-JW	not proper translation of the word "*kesehatan*" in phrase translation list
Sekolah yang buruk (bad school)	*Sakola ku awon*	ID-SU	not proper translation of the word "*ku*" in phrase translation list

The Source Phrase in the first column represents the Indonesian keywords that will be translated, while the target phrase refers to Javanese or Sundanese translation generated by our technique. Actually these two problems might be minimized when bigger and more variety of parallel text added to the phrase table thus the possibility of a word occurs in the phrase translation list is higher.

Another solution of this problem is to use an Indonesian-Javanese and Indonesian-Sundanse dictionary to fill the unknown word resulted by this system. A morphological analyzer could also be added to smooth the translation result of an affixed word.

CONCLUSION

We proposed a technique to fill empty translation of English keyword into Javanese and Sundanese which were occurred in the phrase translation list of Translator-Gator System. We employed the existing phrase pair by consider bahasa Indonesia as a pivot to create English into Javanese or Sundanese translation.

Our experiment shows that as a whole our technique results relatively low translation accuracy. We found that in average our enrichment technique reaches only 37% correct translation result of Indonesian-Javanese and 46% of Indonesian-Sundanese translation. However, by using a weighting formula, at least we have create a quite good phrase translation pair from existing Translator-Gator data, which gives more than 65% proper phrase translation for both Indonesian-Javanese and Indonesian-Sundanese pair translation. In the future we will use a complete lexical dictionary to improve the translation.

ACKNOWLEDGMENT

We would like to thank to the Pulse Lab Jakarta for introduce us to the problem discussed in this paper, as well as for provides data used in this experiment. A great appreciation also addressed to Telkom University for supporting of this paper publication.

REFERENCES

[1] Almeda, J, T. Capistrano, and G. Sarte. *Elementary Statistics.* Quezon City: UP Press, 2010.

[2] Chahuneau, Victor, Eva Schlinger, Noah A. Smith, and Chris Dyer. "Translating into Morphologically Rich Languages with Synthetic Phrases." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1677–1687.* Seattle, Washington, USA: Association for Computational Linguistics, 2013. 1677-1687.

[3] Cohn, Trevor, and Mirella Lapata. "Machine Translation by Triangulation: Making Effective Use of Multi Parallel Corpora." *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 2007. 384-355.

[4] Cui, Yiming, Conghui Zhu, Xiaoning Zhu, and Tiejun Zhao. *Augmenting Phrase Table by Employing Lexicons for Pivot-based SMT.* Arxiv.org, 2015.

[5] Gispert, Adrià de, and José B. Mariño. "Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish." *In Processdings of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages.* 2006. 65-68.

[6] Klementiev, Alexandre, Ann Irvine, Chris Callison-Burch, and David Yarowsky. "Toward Statistical Machine Translation without Parallel Corpora." *EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Avignon, France: Association for Computational Linguistics, 2012. 130-140.

[7] Koehn, Philipp. *Moses-Statistical Machine Translation User Manual.* Edinburgh: University of Edinburgh, 2015.

[8] Riyadi, Yulistina, and Imaduddin Amin. "Translator Gator : Phase I Wrap Up." *United Nations Global Pulse.* June 30, 2016.

http://unglobalpulse.org/news/translator-gator-phase-I-wrap-up (accessed July 31, 2016).

[9] Sarno, Riyanarto, Putu Linda Indita Sari, Hari Ginardi, Dwi Sunaryono, and Imam Mukhlash. "Decision Mining for Multi Choice Workflow Patterns." *International Conference on Computer, Control, Informatics and Its Applications*. 2013. 337-342.

[10] Utiyama, Masao, and Hitoshi Isahara. "A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation." *Proceedings of NAACL HLT 2007*. Rochester, New York: Association for Computational Linguistics, 2007. 484-491.

[11] Wu, Hua, and Haifeng Wang. "Revisiting Pivot Language Approach for Machine Translation." *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*. Singapore: Association for Computational Linguistics, 2009. 154-162.