English – Indonesian Phrase Translation
Using Recurrent Neural Network and ADJ Technique

Wenty Octoviani, Muhammad Fachrurrozi*, Novi Yusliani, Mukhlis Febriady, Afriyan Firdaus

Computer Science, Universitas Sriwijaya, Indonesia

mfachrz@unsri.ac.id

Abstract. Recurrent Neural Network (RNN) and annotated disjunction are discussed to develop English – Indonesian phrase based. Phrase – based English – Indonesian machine translation becomes important because there are differences between two languages and it can be a sub system for English – Indonesian machine translation. Automatic translation can be done using RNN and ADJ Technique. The main process in this research are preprocessing, determination of phrase type, and translation the phrase. In the preprocessing has process such as case folding, tokenizing, pos tagging, and stemming. In the determination of phrase type, system counts weight of input phrase using RNN. The translation phrase is entered to target language using ADJ Technique. The software experiment is tested by using 70 English phrases. The result shows accuracy of RNN and ADJ Technique for English – Indonesian phrase translation is 88.57 %.

1 INTRODUCTION

Natural language is used as tool to communicate to others. The most primary language used to communicate is English. English is also one of our subjects since elementary. But elementary students still can’t understand English [1]. Phrase is being important to machine translation because subject and object are usually a phrase [2]. Differences at phrase structures in English to Indonesian make machine translation should know the rule of target language [3].

Some methods are used to phrase translation are statistical phrase – based [4], syntax – based phrase translation [5], the joint-probability model [6] and phrase – based transfer rule [7]. Comparison from those methods known that statistical phrase based has better result but weakness from this method is the accuracy based on more and variant from training data [8]. Rule based translation has good translation too. So, this research use ADJ Technique which uses transfer rule concept.

Implementation of ADJ Technique at [8] research has another weakness because has a lot of rules to identify the phrases. So, there are some phrases can’t be translated. Based on this reason, need a technique to identify a phrase without use a lot of rules. Research about identify phrase has been done [9] using Recurrent Neural Network.

This research is translated English – Indonesian phrase uses combination Recurrent Neural Network and ADJ Technique. So, it won’t use a lot of rules to identify the phrase. And this machine will identify and translate noun phrase, adjective phrase, adverb phrase, propositional phrase, and verb phrase.

2 METHODOLOGY

500 training phrases are used in this research by each phrase group has 100 phrases. In this research, those phrases were saved at the .txt file based on phrase groups. Based on phrase groups, this research has 5 phrase groups, those are noun phrase, adjective phrase, adverb phrase, propositional phrase, and verb phrase. Generally, the process of this system shows at figure 1.
Figure 1. Process of English – Indonesia Phrase Translation using RNN and ADJ Technique

2.1 Phrase

Noun phrase has some pattern, those are determiner – noun, adjective – noun, and determiner – adjective – noun. Adjective phrase has an adjective, sometimes with an adverb of degree in front of it. Adverb phrase has an adverb, sometimes with an adverb of degree in front of it. Prepositional phrase is a preposition + noun phrase. And verb phrase has an ordinary verb [10]. There can also be one or more auxiliaries. Table 1 shows phrase example

| Phrase                  | Example                                    |
|-------------------------|--------------------------------------------|
| Noun phrase             | Global information, new cars, the red car  |
| Adjective phrase        | Most cool, so happy, very angry            |
| Adverb phrase           | Every morning, really unhappily, really quickly |
| Prepositional phrase    | in the new car, on the book, inside the bottle |
| Verb phrase             | Is going home, have gone, had taught, could be friends |

Noun phrase has some pattern such as determiner-noun, adjective-noun, and determiner-adjective-noun. Determiner noun has 3 parts, those are article, possessive, and demonstrative. Noun phrase classification with determiner-noun pattern can be seen at Table 2.

Adjective – noun is a phrase which has pattern adjective and followed by plural and uncountable noun. Determiner-adjective-noun pattern is a combination from determiner-noun and adjective-noun. Example “my red car”. Adjective phrase is a phrase which is formed by adjective and modifier or determiner. Example “brown hair”, “very beautiful”. Adverb phrase is a phrase which is formed by preposition, verb, noun, or modifier which combine together and be an explanation at sentence. Example “good manner”, “loud voice”. Prepositional phrase is a phrase which is formed by preposition, noun, or pronoun and
modifier. Example “on the table”, “near a wall”, and verb phrase is combination from verb and another word.

Table 2 Classification Noun Phrase with determiner-noun

| Determiner   | Words      | Explanation            | Example              |
|--------------|------------|------------------------|----------------------|
| Article      | “the”      | Followed by noun       | “the cars”, “the fox”|
| Possessive   | “my”, “your”, “our”, “their”, “his”, “her”. | Followed by noun | “my car”, “his water” |
| Demonstrative| “this”, “that” | Followed by singular noun | “this book” |
| Adjective    | “red”, “smart” | Followed by noun | “red cars”, “smart girl” |
| Another noun | “storm”, ”clouds” | Followed by noun | “storm cloud” |

2.2 Pre Processing

Preprocessing is one of steps in this system. Preprocessing in this system are: case folding, tokenizing, pos-tagging, and stemming. Figure 2 will show Preprocessing process in this system.

Figure 2 Pre-Processing in English – Indonesia Phrase Translation using RNN and ADJ Technique.

Case folding is a process to generalize all of character into the same form. In this system, case folding is changing upper case into lower case. Tokenizing is process split sentence into words. Pos-tagging is labeling every word in input. And stemming is a process to find the root word. In this case, stemming is used when the translation can’t be found. Table 3 shows the preprocessing.

Table 3. Preprocessing

| Preprocess   | Before       | After            |
|--------------|--------------|------------------|
| Case Folding | are going home | are going home  |
| Tokenizing   | are going home | are, going, home |
| Pos-tagging  | are, going, home | are/vb, going/vbg, home/nn |
| Stemming     | are, going, home | are, go, home    |

2.3 Recurrent Neural Network

Recurrent neural network (RNN) is a neural network which can remember every process that it done before, because it will be used to next process. RNN has 3 layers are input, hidden, and output [11]. Figure 3 shows RNN process.
Fig 3. RNN Process [11]

X is input symbol, context is hidden layer or will be symbol as S and output will be symbol as Y. RNN has two processes, encoder and decoder. Encoder is read input sequential. In this research, encoder transforms pos-tagging result being binary. After read until the last symbol, hidden layer which be symbol as S will be counted with

\[ S(t) = f\left(\sum_{i}^{n} X_i + S(t-1)\right) \]  

\( S(t) \) = hidden value \( t \)
\( i \) = 0
\( n \) = max value of input
\( X_i \) = value of input layer \( i \) when hidden layer at \( t \)
\( S(t-1) \) = value of hidden layer \((t-1)\)

If we count \( S(1) \) so that \( S(0) \) can be initialized with 0.1. Sigmoid function is

\[ f(z) = \frac{1}{1 + e^{-z}} \]  

after got hidden layer value, next process is decoder. Decoder is training to get output which be symbol as Y.

\[ y_t = f\left(\sum_{j}^{n} S_j(t) V_k(j)\right) \]  

\( Y(t) \) = output layer \( t \)
\( j \) = 0
\( n \) = max index at hidden layer
\( S_j \) = hidden layer \( j \) at output layer \( t \)
\( V_k \) = weight between hidden layer and output layer

2.4 Annotated Disjunct

Annotated disjunct (ADJ) is combination of word from source language, word from target language, and disjunct. Disjunct is a unique letter which represents source language and used to transfer rule at translation. Every phrase which input at software will be mapped to target language. Phrase at target language will be arranged based on the language rule. Fig 4 will show the transfer rule process.
Adj technique can solve phrase problem with difficult case like “the big red car” which has more than one adjective. Adj set for this phrase is \{(the,itu,( ) (D)), (big,besar,( ) (A)), (red,merah,( ) (A)), (car,mobil,((O,D,A)( )))\}.

The phrase “the big red car” can form two phrases. First phrase is adjective – noun “big red car”. The adjective “big” and “red” explain noun “car”. This valid form at Indonesian because adjective (“besar” and “merah”) will follow noun “car” (“mobil”). So, the result of arrangement is “mobil merah besar”.

The second phrase is formed at “the big red car” is determiner – noun phrase which consist determiner “the” and followed by noun phrase “big red car”. Determiner “the” explains noun phrase. But, in Indonesian, determiner “the” (“itu”) follow noun “car” (“mobil”). This case is solved by change the position of determiner at target language and noun. And the result of this translation is “mobil merah besar itu”.

3 TESTING AND ANALYSIS

70 phrases are used in this research. And from these phrases there are 8 inputs which have wrong translation because of some reasons. Table 4 will show the testing result.

| No | Input               | Output           | Right translation | Description |
|----|---------------------|------------------|-------------------|-------------|
| 1  | Our right           | Kanan kita       | Hak kita          | Failed      |
| 2  | Rather brilliant    | Agak brilliant   | Agak cemerlang    | Failed      |
| 3  | Really classic      | Realli klasik    | Sangat klasik     | Failed      |
| 4  | Inside blue can     | -                | Didalam kaleng biru | Failed   |
| 5  | Very close          | Sangat tutup     | Sangat dekat      | Failed      |
| 6  | On the dot          | -                | Diatas titik      | Failed      |
| 7  | Fully harsh         | Fulli kasar      | Sangat kasar      | Failed      |
| 8  | Extremely black     | Extremeli hitam  | Sangat hitam      | Failed      |
At phrase “our right” result of pos-tagging process is our/prp, right/NN so that generalization result is p, n. The mistake happens at translation process at the word “right”. Input “rather brilliant” has mistaken translation because there’s no translation at “brilliant” word. Another mistake translation is “inside blue can” which cannot be translated because the weight didn’t match at corpus.

Inputs “really classic”, “fully harsh”, and “extremely black” each of these has mistaken at stemming process. Stemming process for “really” be “realli”, “fully” be “fulli”, and extremely be “extrem” which those words cannot be translated.

From those testing process, there are 8 phrases which cannot be translated. It means the translation from this English – Indonesia Phrase Translation using Recurrent Neural Network and ADJ Technique has 88.57%.

4 CONCLUSION AND FUTURE WORK

Conclusion from this research are recurrent neural network can be used to train identify English phrase pattern. And, from 70 phrase there are 8 phrase which are not translated well because there’s mistake at pos-tagging, stemming, and translation process. And the last, translation process depend on phrase pattern which based on word class. So that, software can not translate a phrase if there’s a mistake at definition of word class by package stanford pos-tag. The future work of this research are; Develop at package stanford pos-tag, word dictionary, and stemming library; Handle ambiguous meaning at phrase translation; Deep learning at phrase detection so it won’t need a lot of corpus. The result shows accuracy of RNN and ADJ Technique for English – Indonesian phrase translation is 88.57 %.

5 REFERENCES

[1] S. Aris, M Irena Y, “Pengembangan Media CAI untuk meningkatkan hasil belajar mata pelajaran bahasa inggris materi pengenalan kosa kata anggota tubuh manusia dalam bentuk teks deskriptif bergambar siswa kelas 1 SDN Menur Pumpingan IV Surabaya” 2004

[2] Suryadi, et all., “Pembangunan prototipe penerjemahan bahasa inggris ke dalam bahasa indonesia berbasis frasa” KOMMIT2000. Pp C-148. Jakarta: Gunadarma University. 2000

[3] M Novi S, G Susie C, “Translating english noun phrase into indonesian”. Translation strategies for non equivalence. Pp 25-32. 2014.

[4] F Och, H Ney, “The alignment template approach to statistical machine translation”. Vol 30. Computational Statistical. 2004.

[5] K Yamada, K Knight, “A syntax – based statistical translation model”. Proceedings of 39th ACL. 2001

[6] D Marcy, W Wang, “A phrase based joint probability model for statistical machine translation”. Proceedings of the conference on empirical methods in nature language processing. 2002

[7] Bharata, T. A., Baharudin, B., & Zamin, N. “The Development of Phrase-Based Transfer Rules for ADJ-Based Machine Translation”. LETTERS. 2009.

[8] Adji, T. B. “Hierarchical Phrase-Based English - Indonesian Machine Translation Using ADJ Technique”. International Seminar on Scientific Issues and Trends, (pp. A-79). 2011.

[9] Cho, K., Bahdanan, D., Bougares, F., & Bengio, Y. “Learning Phrase Representations using RNN Encoder - Decoder for Statistical Machine Translation”. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (pp. 1724-1734). Doha, Qatar.2014

[10] Eastwood, J. “Oxford Learner's Grammar”. Oxford University Press. New York.2005

[11] Mikolov, T. Recurrent Neural Network Based Language Model. 2005