Parsing Indonesian Syntactic with Recursive Neural Network

Karisma Trinanda Putra[1], Djoko Purwanto[1], and Ronny Mardiyanto[1]

Abstract - Sentence is a form of human communication which is closely related to language system. Sentence is one of the recursive structures that are often found in daily conversation. Learning syntactic structure is useful to explore the meaning of the sentence contained on it or translated it into another language such as machine language. The problem is meaning, ambiguity, and the language that is not according to the rules of syntax, causing the command translation become more complex. This research is about parsing Indonesian syntax based on natural language rules for applications in the field of human-machine interaction. Each word that is a part of the sentence, is mapped into vector-space model. To estimate the potential connection of two words, we use the recursive neural network. The potential connection of two words translated into a higher structure to obtain a complete sentence structure. We obtain 93% accuracy, with 50 data-set are given in the learning process to represent a hundred vocabularies.

Term Index - Natural language processing, vector-space model, recursive neural network.

INTRODUCTION

Basically, communication is one of the important things needed by humans as social beings. Humans can share information with each other with communication. Language is one way to communicate between individuals in society. With language, someone can express what he was thinking to others.

Natural Language Processing (NLP) is one branch of AI which focuses on solving problems that arise in natural language processing. Natural language is the language generally used by humans to communicate with each other. At present, natural language began to be implemented on a computer so that the computer can understand commands given by user. The problems in NLP include meaning, ambiguity, and the language that is not according to the rules of syntax.

Vector-space model (VSM) is a set of modeling languages which include semantic and syntactical features in natural language processing. VSM has been shown to improve performance in NLP tasks such as syntactic parsing [1]. VSM based on neural network can outperform n-gram models in statistical modeling language standard [2]. Recursive structure is commonly found in natural language syntax rules. Recursive neural network (RNN) can predict the hierarchical tree structure [3] [4]. RcNN has succeeded in representing the sentence based on the vector-space model [5].

The goal of this research is to study the natural language processing in Indonesian. Learning the language syntax tree structure will be beneficial to understand the meaning for more natural human-machine interaction.

PROPOSED METHOD

The parsing system consists of two processes, namely mapping into VSM and parsing words with RNN. Neural network (NN) is used to support the vocabulary mapping process. In the process of syntax parsing, the words will be transformed into a vector by matching them with a VSM database. RNN will calculate the value of the connection for each pair of words in a sentence.

A. Mapping Indonesian Word into Vector-space Models

VSM mapping based on class divisions words and phrases. Each word will be represented in the one-hot representation [x1, x2, ..., xR] as NN input. The more recognizable vocabulary would certainly increase the dimensional number of one-hot representation. NN output is a vector value of the distributed representation [y1, y2, ..., yS] with a predetermined number of dimensions.

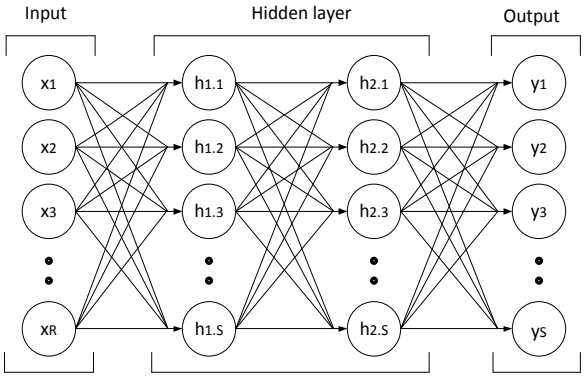

Figure 1. VSM-NN Topology.

B. Parsing Syntax with Recursive Neural Network

RNN has two outputs specially a potential value between words and phrases label. Potential value is the value that describes the strong-correlation between two words. Our RNN is a modification of the SU-RNN without using the different weighting values. In Figure 2, RNN output is syntactic features between two words. Output syntactic features include all phrases contained in Indonesian. Furthermore, the output value of the greatest of all connections word assessed by finding the highest value. Nodes that generate the highest value will unite generate vectors and label the

[1]Nadiar Pratiwi, and Purwanita Setijanti are with Department of Architecture, Faculty of Civil Engineering and Planning, Institut Teknologi Sepuluh Nopember, Surabaya. Email : nadiarrp@gmail.com; p.setijanti@gmail.com

[2]Christiono Utomo is with Department of Civil Engineering, Faculty of Civil Engineering and Planning, Institut Teknologi Sepuluh Nopember, Surabaya. Email : christiono@ce.its.ac.id

new structure. The highest RNN output will be translated into a vector by using the VSM. This process is repeated to obtain the complete structure of the sentence.

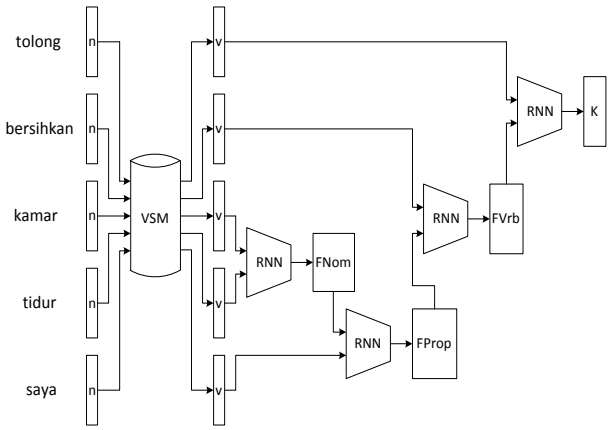

Figure 2. System design.

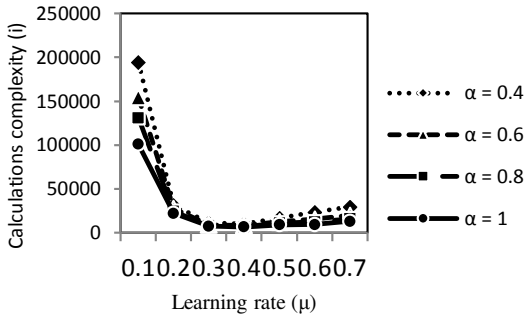

Figure 3. Variation of α vs μ and *i*.

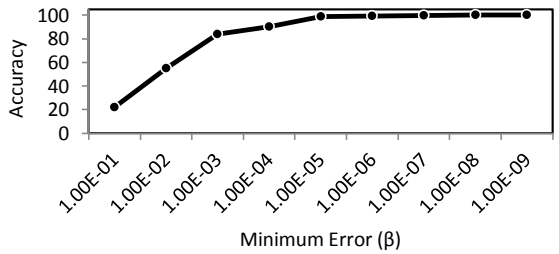

Figure 4. Testing Accuracy with 20-100 pairs of words.

EXPERIMENTS

A. *VSM-NN Characteristic*

The tests use variation of α, β, μ. We design NN which has 100 input neurons representing the word vector in one-hot representation, two hidden layers each consisting of 6 and 3 neurons and 3 output neurons representing the distributed representation in 3 dimension (6 → 6 → 3 → 3). Based on figure 3, the variation of α and μ will affect the complexity of the calculations. The more complex computer calculations cause the longer it takes to complete the process. For this reason, it takes an optimum value which produces the smallest computation complexity. The smallest value is obtained when using α = 1 and μ = 0.4.

B. *Structure Parsing Accuracy*

The tests use a 3-dimensional vector to represent 100 vocabulary in Indonesian. We design NN which has 6 input neurons representing the word vector sequence, two hidden layers each consisting of 6 neurons and 11 output representing the class of phrase (6 → 6 → 6 → 11). The learning process using 30, 40 and 50 combinations of common words and corresponding to Indonesian syntax. Testing is performed by giving 20 to 100 vector combination of two words following each other.

Accuracy will decrease with growing number of data learning. Decreased accuracy is also affected by the amount of data that is tested on RNN. To improve accuracy, we can multiply the learning data sets. The consequence is a growing number of learning data sets, the longer the process and there is a possibility that learning process does not reach the limit value of β. In figure 4, the average accuracy reaches 93%.

CONCLUSION

Indonesian natural language processing system can be performed by considering the semantic and syntactical rules. With 50 learning-set are given in the learning process to represent 100 vocabulary, we obtain about 93% accuracy. Accuracy decreases with increasing number of data tested. Accuracy can be improved by increasing the number of learning data sets but it sacrificed the learning time.

REFERENCES

[1] S. Richard, B. John, M. Christopher, N. Andrew, "Parsing with compositional vector grammars", Proceedings of the ACL conference. 2013.

[2] Yoshua, D. Réjean, V. Pascal, J. Christian, "A Neural Probabilistic Language Model" Journal of Machine Learning Research 3 1137–1155. 2003.

[3] S. Richard, C. Chiung-Yu Lin, Y. Andrew, M. Christopher, "Parsing Natural Scenes and Natural Language with Recursive Neural Networks", Proceedings of the 26th International Conference on Machine Learning (ICML). 2011.

[4] Ronan, W. Jason, B. L´ eon, K. Michael, K. Koray, K. Pavel "Natural Language Processing (Almost) from Scratch" Journal of Machine Learning Research 12. 2011.

[5] Danqi and M. Christopher "A Fast and Accurate Dependency Parser using Neural Networks" Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.