（別紙）

専攻分野及び研究計画
Field of Study and Study Program

Full name in your native language HADING MUHAIMIN
（姓名（自国語））　(Family name/Surname)　(First name)　(Middle name)

Nationality INDONESIA
（国　籍）

Proposed study program in Japan (State the outline of your major field of study on this side and the concrete details of your study program on the back side of this sheet. This section will be used as one of the most important references for selection. The statement must be typewritten or written in block letters. Additional sheets of paper may be attached, if necessary.)
（日本での研究計画；この研究計画は，選考及び大学配置の重要な参考となるので，表面に専攻分野の概要を，裏面に研究計画の詳細を具体に記入すること。記入はタイプ又は楷書によるものとし，必要な場合は別紙を追加してもよい。）

If you have Japanese language ability, write in Japanese.
（相当の日本語能力を有する者は，日本語により記入すること。）

1　Present field of study（現在の専攻分野）

My present field of study is Information Engineering. Here, I focused myself on Natural Language Processing (Machine Learning). For my final Project, I took one of semantic approach of document by using Text mining and vector Spaces Model.

2　Your research theme after arrival in Japan: Clearly explain the research you wish to carry out in Japan.（渡日後の研究テーマ：日本においてどういった研究がしたいかを明確に記入すること）

Attached in separate paper

3　Study program in Japan: (Describe this in detail and concretely—particularly about the ultimate goal of your research in Japan)（研究計画：詳細かつ具体に記入し、特に研究の最終目標について具体的に記入すること。）

(Study program on the back side of this sheet)

2. My study Plan in Japan

Described below is my brief study plan once I am leaving for Japan.

I am planning on finishing my study in Japan for three years, which will be started from my arrival in Japan on April 2015. During the first six months, from April 2015 to October 2015, I will focus to study Japanese since it's very important during my research, so that I will pass the JLPT N2 test. I believe I can pass the test, since I learned Japanese before at Osaka University for three months, and consistently learn and teach the language to my peers after I went back to my hometown. Learning Japanese is also a compulsory to pass the scholarship program.

Besides learning Japanese, I want to focus on literatures that have strong relevance with my research. I also hope I can visit a laboratory where I will conduct my research. Thus, I expect the institution where I learn Japanese is the same university where I will do my research.

For the next six months, which is from October 2015 to April 2016, I will develop and future construction my previous research at University of Hasanuddin relating to Natural Language Processing. The title itself is "*An Analysis of Plagiarism Elements in Scientific Writing Using a Text Mining Algorithm through a Semantic Approach". I collaborated with Ms. Pratiwi Hamdhana, Ms. Mukarramah Yusuf and Mr. Rhiza Sadja. At first, we were trying to label the word by post tagger and then use TF-IDF and Vector Space Model with combination token through the rule of Indonesia Language. I will continue my research to more specific paraphrase identification for semantic analysis using compositionality Neural Network Language Model. To realize this, I want to focus on the introducing Natural Language Processing by Speech and Language Processing handbook, District Mathematics handbook. I also want to focus in word representation in vector and Neural Language model. Then, I will take my master degree at the similar university. If I pass, I will continue my research on a different degree level, and if I don't I still continue my research until October 2016.*

As a master student, I obviously concentrate on the Semantics task special in paraphrase identification, Neural Network Language Model that can sustain the success of my research. I target to finish my master degree within two year time-frame. The first year I will study about the relevant theories, and the next year I expect for results from my research.

After being a student for two years, I want to get involved with various communities on campus to upgrade skills that I have, such as AIESEC, muslim community, and Indonesian Student Community in Osaka-Nara (PPI-ON). I prefer AIESEC, because I believe AIESEC will help me learn more to become a professional and expand my networks. I choose muslim community, because I am a muslim, and being part of PPI-ON can help me create healthy relationship with other muslims and other Indonesians in Japan. I believe those activities will not disturb my study in Japan, because I will manage my time well, just like what I was expected at University of Hasanuddin where I was active at six different organizations but capable of balancing my academics. I also want to join some conference related with my research to get more knowledge and experience.

After April 2018, I will looking for a job in Japan related with my major to implement my knowledge that I get while I study. I have an experience working in Japanese company for 6 weeks in Hitachi Solution Nexus, at the time I enjoyed my experience, and I studied a lot of things. I think work in Japanese company will make me more study and creative. I also have a plan to write the books for Indonesian student to motivate them for better Indonesia in the future, to inspire the Indonesian children to fight for their dream. Japan has always been my primary plan for my study, and that is my concept on what I will do once I arrive in Japan.

Detail of the Proposed Research

1. Research theme

Semantic Analysis thought paraphrase detection using compositionality Neural Network Language Model

2. Classification

Based on Informatics Engineering subject classification, this research will be under these subjects:

131D442 Discrete Mathematics
231D443 Numerical Computation Method
234D443 Artificial Intelligence
306D443 Language Theory and Automata
475D442 Machine Learning

3. Introduction

Phrases and sentence are composed of words. Words can be represented as vectors. It is logical to assume that phrases and sentence can be represented by the composition of word vectors. One of the key ideas in use is to represent the meaning of phrases and sentences by mapping the result of the composition of multiple vectors into the same (word) vector space. It is important that the meaning of a phrase or sentence is determined by the meaning of the words and the rules that combine them [2]. Paraphrase detection is understood of sentence to evaluate two different sentences with different arbitrary lengths and form and to defect whether they capture the same mining or not. The following two sentences are paraphrases:

a. *Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.*
b. *Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.*

Semantic analysis has been actively studied in natural language processing. For the study of semantic analysis, corpora with semantic annotations are essential [4]. Modeling of semantic compositionality in vector space model has emerged as another important line of research. The purpose of language model is to assign a probability to a sequence of words. A statistical language model can be represented by the probability of the next word given all the previous ones, since

$$P(w_1^T) = \prod_{i=1}^{T} P(w_t|w_1^{t-1})$$

An approximation is made by using the previous n words as context instead of all previous words in the sequence

$$P(w_t|w_1^{t-1}) \approx P(w_t|w_1^{t-n+1})$$

This process takes advantage of the fact that temporally closer words are more related than distant words, the Markov assumption. Language models are usually evaluated by using either perplexity or word error rate. Perplexity is the inverse probability of the test set, normalized by the number of words. The perplexity of a sequence of words W and length N is defined in

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{1\ldots.i-1})}}$$

The neural probabilistic language model learns simultaneously a distributed word representation and the probability function for word sequence. The model consists of input projection, hidden and output layers. There is mapping from an element of the vocabulary V to the distrusted feature vectors of dimension m, associated with it (i.e. a|V| x m matrix) [1] [2] [3]

4. Motivation

The main problem of semantic Analysis of Natural Language Processing is how to know the ambiguities of the sentence. There are many semantic tasks such as polysemy identification, word sense induction, paraphrase identification, word mining in the context, etc. For paraphrase identification the purpose is to know that two sentences have the same meaning or not.

Some research has been actively studied about this. The research of Masashi Tsubaki, Kevin Duh, Masashi Shimbo and Yuji Matsumoto is "Modeling and learning semantic Co-Compositionality through prototype projection and Neural Networks". They implement co-compositionality using prototype projections on predicates/arguments and show that this is effective in adapting their word representations. Their model as neural network and propose an unsupervised algorithm to jointly train word representations with co-compositionality. Their model achieves the best result to date $\rho = 0.47$ on the semantic similarity task of transitive verb [3].

On other hand semantic analysis also can use Semantic Role Labeling (SRL) [6]. The SRL is one of the important tasks since its benefits a wide range of Natural Language Processing. Given a sentence then SRL will identify predicated and assign semantically meaningful labels of them.

The goal of this research is to investigate other semantic task (paraphrase detection) using the compositionality using Neural Network Language Model from these point view.

5. Objective

- Objective 1 : Identification paraphrase for understand about the mining of the sentence in semantic analysis. I propose this method for future construction of my previous research. In my previous research I worked in text mining and vector spaces model by semantic approach thought the phrase.
- Objective 2 : I plan to investigate other semantic task such as polysemy, ambiguities of sentence, word sense induction, word meaning in the context, etc. This may lead to other method or word representation in semantic.
- Objective 3 : I plan to investigate the Indonesia-Japanese translation or Japanese error correction by phrase or paraphrase identification. Since I need that system based on my experience when study Japanese.

6. People and Places

This research is expected to be conducted at Graduate School of Information Science, Nara Instituted of Science and Technology, with the supervisor of Professor Yuji Matsumoto in his lab "Computational Linguistic". Early work done by the supervisor which is relevant of this research includes [3], [5] and [6]. I meet him and visited his lab on 8th April 2014. I got the explanation about the classification of his Lab, the tools, the Japanese corpus and semantic analysis. The lab environment is very support this research development.

7. Reserach Planning

Objective 1 is going to be carried out first. Technically, they are easier. Conceptually, the Objective 1 and 2 are attempts to complete my previous research. This understanding will be a good start to tackle the last objectives which are the most ambitious ones.

An approximate schedule may be the following: Objective 1 and 2 seems to be within reach in the first year of the research if accepted. If possible, the preparation for Objective 3 may conduct in this first year. Since I predicted that the training of corpus and word representation by using compositionality Neural Network Language Model would take a greater effort, Objective 3 may be held in the second year, hopefully, in my future graduate school.

8. References

[1] Kevin Duh et.al, "Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation", Matsumoto Laboratory Annual Report 2013, 275-280

[2] Luis Riri and Dunja Miladenic, "Learning Semantic Representations of words and their compositionality" 2014, Jozef Stefan International Postgraduate School

[3] MasashiTsubaki. Kevin Duh, Masashi Shimbo, and Yuji Matsumoto, "Modeling and learning semantic Co-Compositionality through prototype projection and Neural Networks", Empirical Methods in Natural Language Processing (2013), 130-140

[4] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi, "Building a Diverse Document Leads Corpus Annotated with Semantic Relations" Pacific Asia Conference on Language, Information and Computation 2012, 535-544

[5] Sorami Hisamoto, Kevin Duh and Yuji Matsumoto, "An Empirical Investigation of Word Representations for Parsing the Web", the Association of Natural Language Processing (2013), 127-130

[6] Yanyan Luo, Kevin Duh and Yuji Matsuoto, "What information is helpful for dependency based Semantic Role Labelling", International Joint Conference on Natural Language Processing (2013), 781-787