Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 579–586

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

Recurrent Neural Network to Deep Learn Conversation in Indonesian

Andry Chowanda[a,*], Alan Darmasaputra Chowanda[a,b]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. KH. Syahdan No. 9, Jakarta 11480, Indonesia*
[b]*GDP Labs, Jl. Aipda K.S. Tubun II C No.8, Jakarta 11410, Indonesia*

Abstract

Natural Language Processing (NLP) is still considered a daunting task to solve for us, researcher in this field. Specifically, there is not many research has been done in a local language like Indonesian Language. Nowdays, there are hundreds of systems that require NLP as their main functions. This could be a good opportunity for us to explore this opportunity. This paper contributes models from deep learning training in Indonesian conversation using dual encoder LSTM as well as vector representation models trained with three corpora using Skip-gram method. The results show that the models are able to make a good correlation, synonym from a particular word in the words representation of vector models. In addition, the conversation models resulted in 1.07 of perplexity in the Combined model in the 14000th steps.

Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: LTSM; Conversation; Indonesian; Word2Vec; Deep Learning

1. Introduction

The Famous quote of "Can machines think?" [1] has driven researchers in computer science to build not only computer that can think but can communicate naturally with us, human. However, having a natural conversation is currently considered as a daunting task to the machines. Some researchers have successfully built a conversational agent that can communicate with humans [2,3,4,5]. However, they are still limited in a specific context (e.g. making a reservation or asking information about a museum). This is because, conversation is complex to a logical and un-social machine. There are immense possibilities to string words and sentences in a conversation. Every word and sentence might has different meaning of interpretation depending on the situation. The meaning of a word also dependent on the culture of the interlocutor. The Ethnologue catalogue of world languages recorded that there are almost 7000 languages spoken

* Corresponding author. Tel.: +62-21-534-5830
E-mail address: achowanda@binus.edu

1877-0509
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.078

in the world[1]. Yet, mostly resources available for NLP training (i.e. datasets and corpora) are in English. Only view accessible in local languages.

This paper aims to deep learn conversation in a local language, Indonesian by using one of Recurrent Neural Network (RNN) architecture, Long Short-Term Memory (LSTM). We collected three corpora to deep learn conversation in Indonesian. Two corpora were used to train Indonesian words representation in a vector space and the other one was collected to train the conversation with LSTM architecture. In this case, word representations are a substantial component. There are several techniques to represent words[6,7,8], this paper implements vector-based models[6] to represent words in a vocabulary. Unfortunately, there are not a lot of vector-based words representation for Indonesian languages available publicly. Hence, this paper aims to build one and validate it with an application in natural conversation system by using deep learning. Understanding natural conversation in Indonesian is a daunting task, similarly to vector-based models, a model for natural conversation system in Indonesian is not prevalently available for public[9]. This research aims to also present a conversation model in Indonesian language publicly.

The results indicated that the models of words representation in vector shows good results to correlate a particular word to the others. For example: the word Galaxy is the closest word of Samsung according to the model trained with Kompas corpus, and the word LG according to the models trained with Wiki and Combined Corpora. Moreover, the results show that the model trained with news corpus provides more updated data, while the model trained with Wiki dumps results in a more accurate data for general facts and knowledge. The models also capable to pick up the correlation of words such as if "Jokowi" (The name of current President of Republic Indonesia) is a "Presiden" (President), then Einstein is a Physicist ("Fisikawan"). Finally, the conversation models provide a promising results of perplexity around 1.07 - 1.22. This paper contributes the pre-trained models for words representation in a vector space as well as conversation model in Indonesian. To obtain the models, please kindly contact the corresponding author.

2. Related Work

Natural Language Processing (NLP) is still considered a daunting task to solve for us, researcher in this field. Some NLP application has been successfully developed[2,10,3,4,5]. However, they are still very limited in a specific context. For example, we can ask Google about facts, but we can not have a social conversation with it. With a large number of data, Google can easily answer "Who is the president of Indonesia?" precisely. With more than 278.000 search results in less than a second (0.68 second), Google is able to return the answer correctly. There are a number of techniques in NLP, this paper focuses on Word Representations in Vector and Deep Learning for NLP techniques. In order to train the data adequately, corpora or datasets are required.

2.1. Corpora/Datasets for NLP

A conversation system (or commonly called as dialogue system[5]) is generally categorised into two groups: retrieval-based and generative model[11]. A retrieval-based model generally pick the best answer from a pre-scripted list of possible answers prepared beforehand given a specific context. This model always provides a good answer or reply to the user as the answers were prepared and loaded before the system was started. This method, however, requires a huge amount of resources (e.g. time and annotators). Moreover, this method prone to provide a repeated answer, where it can lead to un-natural feeling to the user when interacting with such system.

Meanwhile, a generative model generates a response, word by word based on the input from user. The model statistically observes the probability of the word given the previous words, in a particular context. This model is not easy to train and prone to grammatical errors. The model highly depends on a large number of data. The larger the data, the better it performs. The challenge is that how to find the best datasets or corpora for this model. Table 1 illustrates some of datasets or corpora publicly available for NLP.

Ubuntu Dialogue Corpus[12] offers the biggest number of dialogues and words. It provides human and human chat corpus as well as The Twitter Corpus[13] and Chat Datasets[9]. Only Chat Datasets[9] offers chat in Indonesian language. There are not many corpora or datasets those publicly available in Indonesian Language. The SEMAINE Corpus[14]

[1] http://www.education.rec.ri.cmu.edu/fire/naclo/pages/Ling/Fact/num-languages.html