Indonesian Essay Scoring using Bi-LSTM with Word Embedding Representation

Ilham Firdausi Putra
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
ilhamfputra31@gmail.com

INTRODUCTION

Automatic essay scoring as one of the topics in natural language processing has been greatly developed by the demand to make the assessment process faster. Despite the fast advancement in automatic essay scoring, research in this area for Bahasa Indonesia has been very limited and only recently emerged as a topic. The use of informal language and the diversity of local languages was the main challenge in developing automatic essay scoring for Bahasa Indonesia.

UKARA 1.0 Challenge aims to encourage more ideas and studies for developing automatic short-answer scoring specifically for Bahasa Indonesia. In this challenge, participants will be given access to datasets in the form of students short-answers in two phase. In the first phase, the participant developed their solution with the development set for 43 days from July 29 - September 10, 2019. Finally, the participant can submit their final solution on the second phase with the test set for 3 days from September 16 - September 19, 2019.

INDONESIAN ESSAY SCORING

In this solution, we cast the challenge as a binary classification problem. Given a short-answer to the stimulus, we build a model that tries to predict whether the answer was relevant to the stimulus or not. We process the inputs as a sequence of word. Each word represented as a low dimensional vector and processed sequentially by bidirectional LSTM (Hochreiter and Schmidhuber, 1997).

1. Dataset

Given a stimulus about the challenge of having to migrate because global warming, give one example of the challenges:

"intetraksi/beradaptasi terhadap lingkungan yang baru." (Label: 1)
"akan terjadinya perubahan tempat" (Label: 0)

The dataset is a short-answer from 2 different stimuli (For the size detail, see Table 1). The short-answer and stimulus consist of a total 36,930 word with 2,816 unique vocabularies. The label for each short-answer was a binary with 1 representing relevant answer and 0 representing non-relevant answer. The only text preprocessing done was converting character to lowercase and removing non-alphanumeric character.

Type of Set	Data A	Data B
Training	268	305
Development	215	244
Test	855	974

Table 1: The total of short-answer for each set type and data type.

2. Word Embedding

We pretrained Word2vec (Mikolov et al., 2013) 100 dimension word embedding using Gensim (Rehurek and Sojka, 2010) on Indonesian text from Wikipedia dump, Opensubs (Lison and Tiedemann, 2016), and the preprocessed UKARA. The addition of text from Opensubs and UKARA dataset helps in providing informal words that usually absent in Wikipedia article. With this dataset, we ended up with a total of 420,024 unique vocabularies.

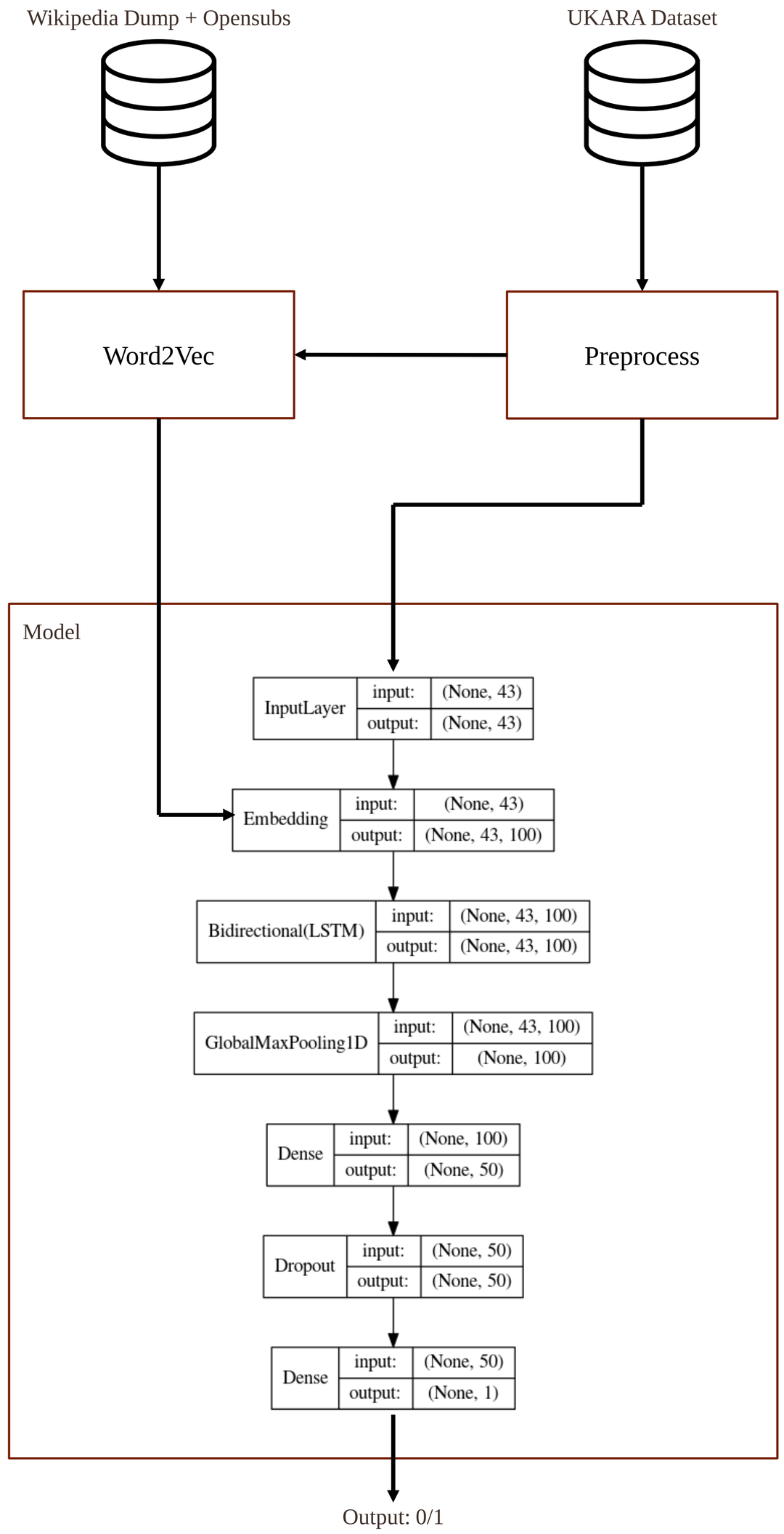
3. Model

We use Keras (Chollet et al., 2015) with Tensorflow (Abadi et al., 2015) as the backend to build the model. The text was tokenized and padded into maximum length of 43 (90th percentile of all short-answer length) before goes into the model. In order to build the embedding layer, we perform a multi-stage text processing using PySastrawi stemmer and a normalizer function (removing duplicate adjacent character) to minimize the amount of unknown vocabulary. This multistage process yields a total of 2.426 known vocabularies and 390 unknown vocabularies. We finally fit the model with an EarlyStopping and ReduceLROnPlateau callback.

4. Experiment

We run the experiment on RepeatedStratifiedKFold with 10 split and 10 repeats. For each split and repeat, we perform prediction to the validation and test set. We later normalize the result according to how many predictions made, essentially performing ensemble of 100 different model fit. We choose a threshold of 0.5 for Data A and 0.48 for Data B in predicting the label. See Table 2 for the result.

OVERVIEW



Type of Data	F1 Score	Precision	Recall
Data A CV	0.892	0.852	0.937
Data B CV	0.770	0.685	0. 880
Data Test	**0.81**	**0.75**	**0.89**

Table 2: The cross-validation and final test result.

CONCLUSION

In this work, we propose Bi-LSTM and pretrained Word2vec to solve Indonesian essay scoring problem. We try to maximize the amount of known vocabulary in building the word embedding by adding Opensubs & preprocessed UKARA data and performing multi-stage text processing. The result shows that the proposed solution was effective by placing 2nd in the UKARA 1.0 Challenge.

Full code & docs available: github.com/ilhamfp/ukara-1.0-challenge