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`

Abstract

. The code and pretrained Word2vec word embedding will be made publicly available[1].

1 Introduction

Jelasin tentang essay scoring dan ukara challenge?

2 Indonesian Essay Scoring

Jelasin ngapain (Gusfield, 1997)

2.1 Dataset

dataset gmn isinya dan preprocessing yang dilakukan

2.2 Word embedding

gmn cara training bikin word embedidngnya pake Gensim (Řehůřek and Sojka, 2010) dan data Opensubs bahasa Indonesia (Lison and Tiedemann, 2016) kita se

2.3 Bi-LSTM

gmn detail modelnya

2.4 Experiment

gmn hasil experimentnya

3 Conclusion

Hehe selesai

References

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

P. Lison and J. Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

A Hyperparameter Detail

A.1 Gensim Hyperparameter

We use `gensim.models.word2vec.Word2Vec` default parameter as of version 3.8.1.

A.2 Model Hyperparameter

We use the default parameter as of Keras version 2.3.0 and Tensorflow version 1.14.0 as the backend with the exception of the following:

Bi-LSTM:

- `units`: 50
- `return_sequences`: True
- `return_dropout`: 0.1
- `return_recurrent_dropout`: 0.1

EarlyStopping:

- `monitor`: ’val_f1’
- `min_delta`: 0.0001
- `patience`: 8
- `mode`: ’max’
- `baseline`: None
- `restore_best_weights`: True

[1] `https://github.com/ilhamfp/ukara-1.0-challenge`

ReduceLROnPlateau:

- `monitor`: 'val_f1'
- `factor`: 0.5
- `patience`: 3
- `mode`: 'max'
- `min_lr`: 1e-6