Abstract

Informal writing style and misspelled words often have a negative impact on several research about Natural Language Processing (NLP), for this reason a normalization process is needed, which play a role to changing the form of unstructured data into a more structured, normalization is considered as a process that can reduce the impact of informal writing style and increase NLP performance.

In this research, a normalization process will be build based on Translation Machine, using Sequence to Sequence architecture with Long Short Term Memory (LSTM) network and attention mechanisms. Datasets used in this research was collected from other translation machine research which using Indonesian Language as targeted, fragments of social media upload comments, and through daily used sentences from questionnaire. The expected output from the system is a structured from of words that can be processed properly by all NLP models that used the similar datasets.

In this research obtained the average values which calculated using BLEU score algorithm, the obtained results 27.66% accuracy value for test datasets which recognized by tokenizer dictionary, and 18.37% accuracy value for test datasets which containing out-of-vocabulary words. Based on this value, the system considered not capable yet to be implemented directly on NPL research and still need a lot of further improvement.

Keywords: *normalization, unstructured sentences, indonesian language, sequence to sequence.*