INTISARI

POS TAGGING KALIMAT BAHASA INDONESIA DENGAN MEMANFAATKAN FITUR MORFOLOGIS KATA PADA ARSITEKTUR *BIDIRECTIONAL LSTM*

Oleh
I Nyoman Prayana Trisna
18/433778/PPA/05593

Part of speech (POS) memiliki peran penting dalam pemrosesan bahasa natural, karena mampu mengklasifikasikan kata sesuai fungsi dalam kalimat. Beberapa pendekatan dalam pembangunan *POS Tagger* telah dilakukan, baik dengan model probabilitas maupun *neural network*, yang mana pemodelan dengan *Long Short Term Memory* (LSTM) menghasilkan performa yang lebih baik dibandingkan model probabilitas. Dalam Bahasa Indonesia, pemodelan POS *Tagger* dengan *Bidirectional* LSTM (BiLSTM) dan *word embedding* belum memanfaatkan fitur morfologis dari kata, padahal dalam bahasa lain pemanfaatan ini telah dilakukan sebelumnya.

Penelitian ini berfokus pada tiga fitur morfologis kata yaitu *prefix*, *suffix*, dan kapitalisasi setiap kata. Tiga fitur ini kemudian ditambahkan ke dalam vektor kata yang dibangun dengan *word embedding* sebagai input. Model dibangun dengan arsitektur dua *layer* BiLSTM. Penelitian ini mencoba untuk memodifikasi *weight size* pada masing-masing *layer, batch size* sekaligus modifikasi terhadap vektor input dari kata.

Performa model dievaluasi berdasarkan nilai *accuracy* dari keseluruhan *tag*, dan nilai *precision*, *recall*, dan *f-measure* pada masing-masing *tag*. Model terbaik diperoleh ketika fitur morfologis *lightstemmer* dimanfaatkan sebagai tambahan pada input *layer* dengan kombinasi ukuran *weight size* 256-128 dan *batch size* 32. Walaupun *lighstemmer* sebagai tambahan fitur lebih baik dibandingkan *fixed character affix*, namun *fixed character affix* lebih baik ketika *word embedding* dihilangkan dalam input *layer*.

Kata kunci: *part of speech tagging*, *bidirectional* LSTM, fitur morfologis, *word embedding*

ABSTRACT

POS TAGGING FOR BAHASA INDONESIA SENTENCE UTILIZING MORPHOLOGICAL FEATURE OF WORD IN BIDIRECTIONAL LSTM

Oleh
I Nyoman Prayana Trisna
18/433778/PPA/05593

Part of speech (POS) procures important role on natural language processing, since it can label words based on their function in sentence. Several approaches to build proper POS Tagger has been done either with probabilistic model or neural network, which shows that modelling with Long Short Term Memory (LSTM) yields better result than probabilistic model. In Bahasa Indonesia, POS Tagger modelling with Bidirectional LSTM (BiLSTM) and word embedding is yet to use morphological feature of word, even though it has been done in other languages.

The conducted research focuses on three morphological features of word, such as prefix, suffix and capitalization of each word. These features are added into word vector that is represented with word embedding as input. Model is constructed with two layers of BiLSTM. This research also tries to modify the weight size in each BiLSTM layer and batch size, as well as input vector of words.

Performance of each model is evaluated based on the accuracy of overall tags, as well as precision, recall, and f-measure of each tag. The best model is obtained when lightstemmer morphological features is used as addition in input layer with combination of 256-128 as the weight size and 32 as batch size. However, even though lightstemmer as additional features is better than fixed character affix, the fixed character affix is better if the word embedding is absent in the input layer.

Key words: part of speech tagging, bidirectional LSTM, morphological feature, word embedding