Pengembangan universal part-of-speech tagger untuk bahasa Indonesia menggunakan bidirectional long short-term memory = Development of universal part-of-speech tagger for Indonesian language using bidirectional long short-term memory.

Yogi Lesmana Sulestio, author

Abstrak

Penelitian Part-of-Speech tagger (POS tagger) untuk bahasa Indonesia telah banyak dikembangkan. Sayangnya, sejauh ini baru Polyglot yang menggunakan POS tag menurut pedoman anotasi Universal Dependencies (UD). Namun, Polyglot sendiri masih mempunyai kekurangan karena belum dapat mengatasi klitik dan kata ulang yang terdapat dalam bahasa Indonesia. Tujuan penelitian ini adalah mengembangkan POS tagger untuk bahasa Indonesia yang tidak hanya sesuai dengan ketentuan anotasi UD, tapi juga sudah mengatasi kekurangan Polyglot. POS tagger ini akan dikembangkan dengan metode deep learning menggunakan arsitektur yang merupakan versi modifikasi dari Recurrent Neural Network (RNN), yaitu Bidirectional Long Short-Term Memory (Bi-LSTM). Dataset yang digunakan untuk mengembangkan POS tagger adalah sebuah dependency treebank bahasa Indonesia yang terdiri dari 1.000 kalimat dan 19.401 token. Hasil eksperimen dengan menggunakan Polyglot sebagai pembanding menunjukkan bahwa POS tagger yang dikembangkan lebih baik dengan tingkat akurasi POS tagging yang meningkat sebesar 6,69% dari 84,82% menjadi 91,51%.

.....There have been many studies that have developed Part-of-Speech tagger (POS tagger) for Indonesian language. Unfortunately, so far only Polyglot that has used POS tag according to Universal Dependencies (UD) annotation guidelines. However, Polyglot itself still has shortcomings since it has not been able to overcome clitics and reduplicated words in Indonesian language. The purpose of this study is to develop POS tagger for Indonesian language which is not only in accordance with UD annotation guidelines, but also has overcome Polyglot’s shortcomings. This POS tagger will be developed under deep learning method by using modified version of Recurrent Neural Network (RNN) architecture, Bidirectional Long Short-Term Memory (Bi-LSTM). The dataset used to develop POS tagger is an Indonesian dependency treebank consisting of 1.000 sentences and 19.401 tokens. Result of experiment using Polyglot as baseline shows that the developed POS tagger is better. This is indicated by increased accuracy POS tagging by 6,69% from 84,82% to 91,51%.