INTISARI

PENANGANAN AMBIGUITAS PADA *POS TAGGING* BAHASA INDONESIA MENGGUNAKAN BERT

oleh

Ahmad Subhan Yazid
19/448685/PPA/05768

Ambiguitas merupakan masalah yang kerap muncul pada tugas-tugas pemrosesan bahasa alami, termasuk pada *POS tagging* (pelabelan kelas kata). Penelitian ini bertujuan menangani ambiguitas pada proses *POS tagging* bahasa Indonesia dengan pendekatan pembelajaran mendalam BERT (*bidirectional encoder representation from transformers*). Pendekatan ini dipilih juga untuk melengkapi kelemahan dari penelitian sebelumnya yang menerapkan metode berbasis aturan dan probabilistik.

Dataset yang digunakan dalam penelitian ini merupakan modifikasi dari korpus POS tagging IDN_Tagged_Korpus dengan jumlah total token sebanyak 267.054. Untuk mendapatkan model yang optimal dan dapat menyelesaikan ambiguitas, dilakukan 15 eksperimen dengan skenario *fine-tuning* BERT melalui penyesuaian parameter dan penambahan data terhadap dataset. Eksperimen tersebut dijalankan di atas model IndoBERT sebagai landasan dengan pustaka Simple Transformers.

Eksperimen yang dilakukan menghasilkan model terbaik dengan nilai *loss* 0,1113, *precision* 0,9635, *recall* 0,9658, dan *f1* 0,9647. Hasil tersebut didapatkan pada eksperimen dengan parameter *learning rate* 0,00004, ukuran *batch* 16, pada *epoch* ke-2. Hasil pengujian terhadap data uji juga menunjukkan bahwa model memiliki performa yang baik dan mampu menangani ambiguitas. Model yang dihasilkan dapat melabeli 96 dari 100 kata ambigu pada kalimat data uji dengan benar. Kendati demikian, masih terdapat kelemahan model untuk menangani kata ambigu dalam pola tertentu.

Kata Kunci: Ambiguitas, BERT, *Fine-tuning*

ABSTRACT

AMBIGUITY HANDLING IN INDONESIAN POS TAGGING USING BERT

by

Ahmad Subhan Yazid
19/448685/PPA/05768

Ambiguity is a recurring problem in natural language processing tasks, including POS tagging (word class labeling). This research aims to address ambiguity in Indonesian POS tagging with the BERT (bidirectional encoder representation from transformers) deep learning approach. This approach was also chosen to complement the weaknesses of previous studies that applied rule-based and probabilistic methods.

The dataset used in this research is a modification of the POS tagging corpus IDN_Tagged_Korpus with a total token count of 267,054. To get an optimal model that can resolve ambiguity, 15 experiments were conducted with BERT fine-tuning scenarios through parameter adjustments and adding data to the dataset. The experiments were run on top of the IndoBERT model as a foundation with the Simple Transformers library.

The experiments conducted resulted in the best model with a loss value of 0.1113, precision of 0.9635, recall of 0.9658, and f1 of 0.9647. These results were obtained in experiments with a learning rate parameter of 0.00004, batch size 16, at the 2nd epoch. Test results on test data also show that the model has good performance and is able to handle ambiguity. The resulting model can correctly label 96 out of 100 ambiguous words in the test data sentences. However, there are still weaknesses in the model to handle ambiguous words in certain patterns.

keywords: Ambiguity, BERT, Fine-tuning