Cross-lingual Transfer Learning untuk Part-of-speech Tagging Bahasa Jawa = Cross-lingual Transfer Learning for Javanese Part-of-speech Tagging

Gabriel Enrique, author

Abstrak

Part-of-speech tagging, adalah task di bidang Natural Language Processing di mana setiap kata di dalam suatu kalimat dikategorisasi ke dalam kategori parts-of-speech (kelas kata) yang sesuai. Pengembangan model POS tagger menggunakan pendekatan machine learning membutuhkan dataset dengan ukuran yang besar. Namun, dataset POS tagging tidak selalu tersedia dalam jumlah banyak, seperti dataset POS tagging untuk bahasa Jawa. Dengan jumlah data yang sedikit, model POS tagger yang dilatih kemungkinan tidak akan memiliki performa yang optimal. Salah satu solusinya adalah dengan menggunakan pendekatan cross-lingual transfer learning, di mana model dilatih menggunakan suatu source language pada suatu task agar dapat menyelesaikan task yang sama pada suatu target language. Penelitian ini bertujuan untuk menguji performa pre-trained language model (mBERT, XLM-RoBERTa, IndoBERT) dan melihat pengaruh cross-lingual transfer learning terhadap performa pre-trained language model untuk POS tagging bahasa Jawa. Percobaan yang dilakukan menggunakan lima source language, yaitu bahasa Indonesia, bahasa Inggris, bahasa Uighur, bahasa Latin, dan bahasa Hungaria, serta lima jenis model, yaitu fastText + LSTM, fastText + BiLSTM, mBERT, XLM-RoBERTa, dan IndoBERT; sehingga secara keseluruhan ada total 35 jenis model POS tagger. Model terbaik yang dilatih tanpa pendekatan cross-lingual transfer learning dibangun menggunakan IndoBERT, dengan akurasi sebesar 86.22%. Sedangkan, model terbaik yang dilatih menggunakan pendekatan cross-lingual transfer learning dalam bentuk dua kali fine-tuning, pertama menggunakan source language dan kedua menggunakan bahasa Jawa, sekaligus model terbaik secara keseluruhan dibangun menggunakan XLM-RoBERTa dan bahasa Indonesia sebagai source language, dengan akurasi sebesar 87.65%. Penelitian ini menunjukkan bahwa pendektan cross-lingual transfer learning dalam bentuk dua kali fine-tuning dapat meningkatkan performa model POS tagger bahasa Jawa, dengan peningkatan akurasi sebesar 0.21%–3.95%.

.....

Part-of-speech tagging is a task in the Natural Language Processing field where each word in a sentence is categorized into its respective parts-of-speech categories. The development of POS tagger models using machine learning approaches requires a large dataset. However, POS tagging datasets are not always available in large quantities, such as the POS tagging dataset for Javanese. With a low amount of data, the trained POS tagger model may not have optimal performance. One of the solution to this problem is using the cross-lingual transfer learning approach, where a model is trained using a source language for a task so that it can complete the same task on a target language. This research aims to test the performance of pre-trained language models (mBERT, XLM-RoBERTa, IndoBERT) and to see the effects of cross-lingual transfer learning on the performance of pre-trained language models for Javanese POS tagging. The experiment uses five source languages, which are Indonesian, English, Uyghur, Latin, and Hungarian, as well as five models, which are fastText + LSTM, fastText + BiLSTM, mBERT, XLM-RoBERTa, and IndoBERT; hence there are 35 POS tagger models in total. The best model that was trained without cross-

lingual transfer learning approach uses IndoBERT, with an accuracy of 86.22%. While the best model that was trained using a cross-lingual transfer learning approach, implemented using a two fine-tuning process, first using the source language and second using Javanese, as well as the best model overall uses XLM-RoBERTa and Indonesian as the source language, with an accuracy of 87.65%. This research shows that the cross-lingual transfer learning approach, implemented using the two fine-tuning process, can increase the performance of Javanese POS tagger models, with a 0.21%–3.95% increase in accuracy.