Cross-lingual Transfer Learning untuk Dependency Parsing Bahasa Jawa = Cross-lingual Transfer Learning for Javanese Dependency Parsing

Fadli Aulawi Al Ghiffari, author

Abstrak

Penelitian ini bertujuan untuk membangun model dependency parser untuk bahasa Jawa menggunakan pendekatan cross-lingual transfer learning. Metode transfer learning dipilih untuk mengatasi kurangnya dataset yang tersedia untuk proses training model pada bahasa Jawa yang merupakan low-resource language. Model dibangun menggunakan arsitektur encoder-decoder, tepatnya menggunakan gabungan dari self-attention encoder dan deep biaffine decoder. Terdapat tiga skenario yang diuji yaitu model tanpa transfer learning, model dengan transfer learning, dan model dengan hierarchical transfer learning. Metode transfer learning menggunakan bahasa Indonesia, bahasa Korea, bahasa Kroasia, dan bahasa Inggris sebagai source language. Sementara metode hierarchical transfer learning menggunakan bahasa Prancis, bahasa Italia, dan bahasa Inggris sebagai source language tahap satu, serta bahasa Indonesia sebagai source language tahap dua (intermediary language). Penelitian ini juga mengujikan empat word embedding yaitu fastText, BERT Jawa, RoBERTa Jawa, dan multilingual BERT. Hasilnya metode transfer learning secara efektif mampu menaikkan performa model sebesar 10%, di mana model tanpa transfer learning yang memiliki performa awal unlabeled attachment score (UAS) sebesar 75.87% dan labeled attachment score (LAS) sebesar 69.04% mampu ditingkatkan performanya hingga mencapai 85.84% pada UAS dan 79.22% pada LAS. Skenario hierarchical transfer learning mendapatkan hasil yang lebih baik daripada transfer learning biasa, namun perbedaannya tidak cukup signifikan.

.....This research aims to develop a Javanese dependency parser model using a cross-lingual transfer learning approach. The transfer learning method was chosen to overcome the lack of available datasets for the model training process in Javanese, a low-resource language. The model uses an encoder-decoder architecture, precisely combining a self-attention encoder and a deep biaffine decoder. Three scenarios are experimented with: a model without transfer learning, a model with transfer learning, and a model with hierarchical transfer learning. The transfer learning process uses Indonesian, Korean, Croatian, and English as source languages. In contrast, the hierarchical transfer learning process uses French, Italian, and English as the first-stage source languages and Indonesian as the second-stage source language (intermediary language). This research also experimented with four word embedding types: fastText, Javanese BERT, Javanese RoBERTa, and multilingual BERT. The results show that the transfer learning method effectively improves the model’s performance by 10%, where the model without transfer learning has an initial unlabeled attachment score (UAS) performance of 75.87% and labeled attachment score (LAS) of 69.04% can be increased to 85.84% in UAS and 79.22% in LAS. Hierarchical transfer learning has a slightly better result than standard transfer learning, but the difference is insignificant.