Eksplorasi Keefektifan Cross-Lingual Transfer Learning untuk Constituency Parsing Bahasa Indonesia = Exploring the Efficacy of Cross-Lingual Transfer Learning for Indonesian Constituency Parsing

Muhammad Faisal Adi Soesatyo, author

Abstrak

Pendekatan transfer learning telah digunakan di beragam permasalahan, khususnya low-resource language untuk meningkatkan performa model di masing-masing permasalahan tersebut. Fokus pada penelitian ini ingin menyelidiki apakah pendekatan cross-lingual transfer learning mampu meningkatkan performa pada model constituency parsing bahasa Indonesia. Constituency parsing adalah proses penguraian kalimat berdasarkan konstituen penyusunnya. Terdapat dua jenis label yang disematkan pada konstituen penyusun tersebut, yakni POS tag dan syntactic tag. Parser model yang digunakan di penelitian ini berbasis encoder-decoder bernama Berkeley Neural Parser. Terdapat sebelas macam bahasa yang digunakan sebagai source language pada penelitian ini, di antaranya bahasa Inggris, Jerman, Prancis, Arab, Ibrani, Polandia, Swedia, Basque, Mandarin, Korea, dan Hungaria. Terdapat dua macam dataset bahasa Indonesia berformat Penn Treebank yang digunakan, yakni Kethu dan ICON. Penelitian ini merancang tiga jenis skenario uji coba, di antaranya learning from scratch (LS), zero-shot transfer learning (ZS), dan transfer learning dengan fine-tune (FT). Pada dataset Kethu terdapat peningkatan F1 score dari 82.75 (LS) menjadi 84.53 (FT) atau sebesar 2.15%. Sementara itu, pada dataset ICON terjadi penurunan F1 score dari 88.57 (LS) menjadi 84.93 (FT) atau sebesar 4.11%. Terdapat kesamaan hasil akhir di antara kedua dataset tersebut, di mana masing-masing dataset menyajikan bahwa bahasa dari famili Semitic memiliki skor yang lebih tinggi dari famili bahasa lainnya.

.....The transfer learning approach has been used in various problems, especially the low-resource languages, to improve the model performance in each of these problems. This research investigates whether the cross-lingual transfer learning approach manages to enhance the performance of the Indonesian constituency parsing model. Constituency parsing analyzes a sentence by breaking it down by its constituents. Two labels are attached to these constituents: POS tags and syntactic tags. The parser model used in this study is based on the encoder-decoder named the Berkeley Neural Parser. Eleven languages are used as the source languages in this research, including English, German, French, Arabic, Hebrew, Polish, Swedish, Basque, Chinese, Korean, and Hungarian. Two Indonesian PTB treebank datasets are used, i.e., the Kethu and the ICON. This study designed three types of experiment scenarios, including learning from scratch (LS), zero-shot transfer learning (ZS), and transfer learning with fine-tune (FT). There is an increase in the F1 score on the Kethu from 82.75 (LS) to 84.53 (FT) or 2.15%. Meanwhile, the ICON suffers a decrease in F1 score from 88.57 (LS) to 84.93 (FT) or 4.11%. There are similarities in the final results between the two datasets, where each dataset presents that the languages from the Semitic family have a higher score than the other language families.