Pembangunan Korpus dan Model Relasi Semantik Hiponim-Hipernim Bahasa Indonesia dengan Pendekatan Pattern-Based, Crowdsourcing, dan Machine Learning = Building Indonesian Hyponym-Hypernym Semantic Relations Corpus and Model Using Pattern-Based, Crowdsourcing, and Machine Learning Approach.

Yudhistira Erlandinata, author

Abstrak

Korpus relasi semantik dapat menunjang berbagai penelitian di bidang pengolahan bahasa manusia. Untuk Bahasa Indonesia, korpus relasi semantik yang berukuran besar dan berkualitas baik masih belum tersedia. Korpus relasi semantik dapat dibuat secara manual dengan melibatkan anotator dan juga dapat dihasilkan secara otomatis menggunakan algoritma rule-based atau machine learning. Penelitian ini bertujuan untuk mengevaluasiseberapa baik kualitas korpus relasi semantik Bahasa Indonesia, khususnya relasi hiponim-hipernim, apabila dibangun dengan pendekatan machine learning dan metode crowdsourcing yang menerapkan gamifikasi. Algoritma pattern-based yang sebelumnya pernah diteliti untuk Bahasa Indonesia akan digunakan untuk menghasilkan data training algoritma machine learning dan kandidat entri korpus untuk dianotasi dengan metode crowdsourcing. Kualitas korpus hasil metode crowdsourcing diukur berdasarkan tingkat persetujuan antar anotator dan diperoleh hasil yang cukup baik walaupun belum sempurna. Untuk pendekatan machine learning, beberapa model
machine learning yang diterapkan masih belum memberikan hasil optimal karena keterbatasan resource.
Kata kunci: relasi semantik, hiponim-hipernim, crowdsourcing, gamifikasi, machine learning, pattern-based
.....Semantic relations corpus is vital to support research in the field of Natural Language Processing. Currently, there is no existing corpus of semantic relations in Indonesian language which is enormous and high-quality. The corpus can be constructed manually by employing human annotators or built automatically using rule-based or machine learning algorithms. This research aims to evaluate the quality of Indonesian hyponym-hypernym semantic relations corpus that is produced by crowdsourcing mechanism with gamification, and to test the model for semantic relations prediction using machine learning algorithms. The pattern-based method is applied to obtain the training data for machine learning experiments and corpus entry candidates to be annotated using the crowdsourcing method. The quality of the crowdsourced corpus is measured using inter-annotator agreement. The experimental result shows that the gamification-based crowdsourcing method is promising to produce the corpus. On the other hand, machine learning models tested in this research have not given optimal results yet due to the limitations of the lexical resources in Indonesian language.