Gramatika: Dataset Sintetik untuk Grammatical Error Correction Bahasa Indonesia = Gramatika: A Synthetic Dataset for Indonesian Grammatical Error Correction

Rico Tadjudin, author

Abstrak

Grammatical Error Correction (GEC) merupakan bagian dari Natural Language Processing yang membahas suatu task untuk mendeteksi dan setelahnya mengoreksi suatu teks. Pekerjaan tersebut mencakup pendeteksian dan pengoreksian kesalahan tata bahasa, kesalahan ortografi, dan semantik. Perkembangan GEC untuk bahasa Indonesia terkendala oleh sedikitnya dataset yang dapat digunakan untuk melatih model GEC. Penelitian ini mengusulkan pendekatan rule-based untuk membangun sebuah dataset sintetik yang mengandung kalimat salah secara tata bahasa baku bahasa Indonesia beserta koreksinya. Hal tersebut dapat dilakukan dengan memanfaatkan kamus tesaurus bahasa Indonesia dan alat bantuan NLP seperti tokenizer, part-of-speech tagger, morphological analyzer, dan dependency parser untuk mengekstrak informasi konteks dari kalimat. Kumpulan data sintetik dibangkitkan dengan menggunakan kalimat yang benar secara tata bahasa dari halaman0halaman situs Wikipedia sebagai kalimat input. Dataset ini menyediakan data dalam dua format yang berbeda, yaitu dalam format M2 dan dalam bentuk pasangan kalimat salah dan benar. Pembangkitan kesalahan tata bahasa akan memiliki 17 kemungkinan jenis kesalahan tata bahasa yang berbeda dengan total 16.898 kalimat salah yang dibentuk. Pengujian Gramatika dilakukan dengan melakukan evaluasi secara manual mengenai ketepatan pembangkitan tiap kesalahan pada kalimat. Pengujian manual dilakukan dengan melakukan stratified random sampling untuk mengambil sampel 100 kalimat. Sampel tersebut minimal memiliki 5 contoh untuk setiap jenis kesalahan tata bahasa. Dari pengevaluasian yang dilalukan oleh dua penguji, didapatkan nilai accuracy sebesar 91,1%. <hr /> Grammatical Error Correction (GEC) is a part of Natural Language Processing which deals with the task of detecting and correcting a text. This includes correcting grammatical errors, semantic errors, and orthographic errors. GEC development in Indonesian language has been hindered by the lack of suitable dataset that can be used to train GEC models. This research proposes a rule-based approach to develop a synthetic dataset that contains sentences in Indonesian with grammar errors and its corresponding corrections. It’s done with the help of dictionaries such as Indonesian thesaurus and NLP tools such as a tokenizer, part of speech tagger, morphological analyzer, and dependency parser to extract contextual information of sentences. The synthetic dataset is generated by using grammatically correct sentences from Wikipedia pages as the input. The resulting dataset is formatted to M2 format and pairs of correct and false sentences, containing 17 types of errors with a total of 16.898 sentences. The evaluation of Gramatika is done by manually assessing the accuracy of the sentence modifications. To do this, stratified random sampling is conducted to select 100 sentences with a minimum of 5 examples for each error type. From the manual evaluation by two evaluators, an average accuracy score of 91.1% is obtained.