Perbandingan Metode Grammatical Error Correction antara T5 dan GECToR = Comparison of Grammatical Error Correction Methods between T5 and GECToR

Napitupulu, Jeremy Victor Andre, author

Abstrak

Grammatical Error Correction (GEC) adalah salah satu task Natural Language Processing (NLP) yang mendeteksi dan mengoreksi kesalahan tata bahasa dalam sebuah teks. Task ini terus berkembang sampai saat ini dan telah diterapkan menggunakan berbagai metode, seperti rule-based, machine learning-based, dan sebagainya. Tugas akhir ini bertujuan membandingkan dua metode state-of-the-art Grammatical Error Correction yaitu metode T5 dan GECToR menggunakan dataset bahasa Inggris dan bahasa Indonesia. Untuk metode T5, akan dibandingkan model Flan-T5 dan mT5 dengan variasi ukuran base dan large. Adapun model yang dibandingkan untuk metode GECToR adalah model RoBERTa dan XLNet dengan variasi ukuran base dan large. Untuk dataset bahasa Inggris, akan digunakan dataset FCE untuk training dan dataset CoNLL-14 untuk testing. Sedangkan untuk dataset bahasa Indonesia, akan digunakan dataset Gramatika. Kemudian, untuk evaluasi digunakan metrik F0.5. Berdasarkan hasil uji coba, didapatkan bahwa untuk dataset bahasa Inggris FCE+CoNLL-14, metode T5 dengan varian model Flan-T5 unggul dari kedua varian metode GECToR dengan skor F0.5 sebesar 52,85%. Varian Flan-T5 ini unggul dengan margin sebesar 15,83% dari varian terbaik metode GECToR, yaitu RoBERTa. Sedangkan, metode GECToR dengan varian RoBERTa lebih unggul dengan margin 10,12% dari metode T5 dengan varian model mT5. Untuk dataset bahasa Indonesia Gramatika, kedua varian metode T5 lebih unggul dari metode GECToR. Varian terbaik metode T5 dengan skor F0.5 sebesar 45,38% dengan margin 31,05% dari varian terbaik metode GECToR, yaitu RoBERTa.

.....Grammatical Error Correction (GEC) is one of the Natural Language Processing (NLP) tasks that detect and correct grammatical errors in a text. This task continues to grow today and has been implemented using various methods, such as rule-based, machine learning-based, and so on. This final project aims to compare two state-of-the-art Grammatical Error Correction methods, namely the T5 and GECToR methods using English and Indonesian datasets. For the T5 method, Flan-T5 and mT5 models will be compared with base and large size variations. As for the GECToR method, RoBERTa and XLNet models will be compared with base and large size variations. For the English dataset, the FCE dataset will be used for training and the CoNLL-14 dataset for testing. As for the Indonesian dataset, the Grammatical dataset will be used. Then, the F0.5 metric is used for evaluation. Based on the experimental results, it is found that for the FCE+CoNLL-14 English dataset, the T5 method with the Flan-T5 model variant is superior to both variants of the GECToR method with an F0.5 score of 52.85%. The Flan-T5 variant is superior by a margin of 15.83% to the best variant of the GECToR method, RoBERTa. Meanwhile, the GECToR method with the RoBERTa variant is superior by a margin of 10.12% to the T5 method with the mT5 model variant. For the Indonesian Grammatical dataset, both variants of the T5 method are superior to the GECToR method. The best variant of the T5 method with an F0.5 score of 45.38% with a margin of 31.05% from the best variant of the GECToR method, which is RoBERTa.