Pengembangan Neural Language Model Untuk Bahasa Singlish Dengan ELECTRA = Developing a Singlish Neural Language Model using ELECTRA

Galangkangin Gotera, author

Abstrak

Singlish adalah sebuah bahasa informal yang sering digunakan warga Singapura. Karena informal, bahasa Singlish jarang ditemukan di media umum seperti majalah, koran, dan artikel internet. Meski demikian, bahasa ini sangat sering digunakan oleh warga Singapu- ra pada percakapan sehari-hari, baik daring maupun luring. Banyak campuran bahasa lain (code-mixing) merupakan tantangan lain dari Singlish. Keterbatasan GPU juga menjadi tantangan dalam mendapatkan model yang baik. Mempertimbangkan semua tantangan ini, penulis telah melatih sebuah model Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) pada data berbahasa Singlish. ELECTRA merupakan sebuah model baru yang menawarkan waktu training lebih cepat sehingga menjadi pilihan baik jika memiliki keterbatasan GPU. Data Singlish didapatkan melalui web scraping pada reddit dan hardwarezone. Penulis membuat sebuah dataset benchmark pada dua buah permasalahan yaitu sentiment analysis dan singlish identification dengan anotasi manual sebagai metode untuk mengukur kemampuan model dalam Singlish. Penulis melakukan benchmarking pada model yang dilatih dengan beberapa model yang tersedia secara terbuka dan menemukan bahwa model ELECTRA yang dilatih memiliki perbedaan akurasi paling besar 2% dari model SINGBERT yang dilatih lebih lama dengan data yang lebih banyak.

.....Singlish is an informal language frequently used by citizens of Singapore (Singaporeans). Due to the informal nature, Singlish is rarely found on mainstream media such as magazines, news paper, or internet articles. However, the language is commonly used on daily conversation, whether it be online or offline. The frequent code-mixing occuring in the language is another tough challenge of Singlish. Considering all of these challenges, we trained an Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) model on a Singlish corpus. Getting Singlish data is hard, so we have built our own Singlish data for pre-training and fine-tuning by web scraping reddit and hardwarezone. We also created a human-annotated Singlish benchmarking dataset of two downstream tasks, sentiment analysis and singlish identification. We tested our models on these benchmarks and found out that the accuracy of our ELECTRA model which is trained for a short time differ at most 2% from SINGBERT, an open source pre-trained model on Singlish which is trained with much more data.