Universitas Indonesia Library >> UI - Tesis Membership

Strategi Strategi Fine-Tuning Dan Augmentasi Data Lintas Bahasa Untuk Meningkatkan Kinerja Model Bert Pada Tugas Machine Reading Comprehensive Dalam Bahasa Sumber Daya Rendah = Fine-Tuning And Crosslingual Data Augmentation Strategies To Improve BERT Model Performance On Machine Reading Comprehension Task In Low Resource Languages

Ryan Pramana, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920520299&lokasi=lokal

---

Abstrak

Machine Reading Comprehension (MRC) merupakan salah satu task di bidang natural language processing (NLP) dimana mesin memiliki tugas untuk membaca secara komprehensif dari sebuah bacaan (passage) yang diberikan agar dapat menjawab pertanyaan terkait. Metode terkini untuk mengautomasi MRC menggunakan deep learning dengan memanfaatkan pretrained language models (PLMs) berbasis BERT. Dalam menangani kasus MRC sumber daya rendah, digunakan PLM multilingual seperti XLM-R. Namun PLM multilingual memiliki masalah untuk bahasa sumber daya rendah yaitu: bahasa sumber daya rendah yang tidak terepresentasi dengan baik, imperfect cross-lingual embeddings alignment dan instabilitas ketika di fine-tuning pada data berukuran kecil. Penelitian ini mengusulkan beberapa strategi fine-tuning dan metode pembentukan data augmentasi untuk meningkatkan kinerja MRC dibahasa sumber daya rendah. Strategi fine-tuning yang diusulkan adalah 2-step fine-tuning dan mixed fine-tuning. Untuk metode pembentukan data augmentasi yaitu dengan penggunaan data asli, pengaplikasian model machine translation dan perturbasi code-switching. Hasil eksperimen menunjukkan, untuk dataset FacQA (Bahasa Indonesia) dan UIT-ViQuAD (Bahasa Vietnam) diperoleh strategi terbaik dengan kombinasi strategi penggunaan data asli dan metode 2-step finetuning dimana menghasilkan peningkatan kinerja sebesar 3.858%, 2.13% secara berurutan. Untuk dataset FQuAD (Bahasa Prancis), strategi terbaik diperoleh de- ngan kombinasi strategi pembentukan data perturbasi code-switching dan metode mixed fine-tuning dimana menghasilkan peningkatan kinerja sebesar 1.493%.

.....Machine Reading Comprehension (MRC) is one of the tasks in the field of natural language processing (NLP) where the machine has the task of reading comprehensively from a given passage in order to answer related questions. The latest method for automating MRC uses deep learning by utilizing pretrained language models (PLMs) based on BERT. For handling low-resource MRC, multilingual PLMs such as XLM-R are used. However, multilingual PLM has problems for low resource languages: low resource languages that are underrepresented, imperfect cross-lingual embeddings alignment and instability when finetuned on small data.This study proposes several fine-tuning strategies and data augmentation generation methods to improve lowresource languages MRC performance. The proposed fine-tuning strategies are 2-step fine-tuning and mixed fine-tuning. For the method of form- ing augmented data, namely by using data original model, application of machine translation and code-switching pertubation to optimize cross-lingual embeddings alignment in multilingual PLM. The experimental results show that for the FacQA (Indonesian) and UIT-ViQuAD (Vietnamese) datasets, the best strategy is obtained by combining the strategy of using original data and the 2-step fine-tuning method which results in an performance improvement of 3.858%, 2.13%, respectively. For the FQuAD dataset (French), the best strategy was obtained by a combination of

code-switching perturbation strategy and mixed fine-tuning method which resulted in an performance improvement of 1.493%.