AC-IQuAD: Automatically Constructed Indonesia Question Dataset by Leveraging Wikidata = AC-IQuAD: Dataset QA Indonesia yang Dibuat Secara Otomatis

Kerenza Doxolodeo, author

Abstrak

Konstruksi dataset QA membutuhkan akses ke sumber daya dan finansial yang tidak kecil, sehingga dataset untuk bahasa-bahasa yang kurang dipelajari seperti Ba- hasa Indonesia minim. Studi ini mengkonstruksi dataset QA Indonesia yang dibuat secara otomatis dari awal hingga akhir. Proses dimulai dengan mengambil tripel dari Wikidata dan mengkonversikan tripel tersebut menjadi pertanyaan menggu- nakan CFG. Teks konteks dicari dari korpus Wikipedia Bahasa Indonesia dengan heuristik untuk mencari teks yang sesuai. Pertanyaan-pertanyaan tersebut dival- idasi dengan model M-BERT yang fungsinya sebagai proxy model yang menilai kelayakan pertanyaan. Dataset terdiri dari 134 ribu baris pertanyaan simpel dan 60 ribu pertanyaan kompleks yang menggandung dua buah fakta dalam satu per- tanyaan. Untuk pertanyaan simpel dataset mendapatkan evaluasi yang mirip oleh manusia (72% AC-IQuAD vs 67% SQuAD terjemahan) dan model QA Indonesia yang terbaik adalah yang menggabungkan dataset SQuAD Inggris dan AC-IQuAD (F1 57.03 terhadap dataset TydiQA).

.....Construction of QA datasets requires access to considerable resources and fi- nance, so datasets for less-learned languages such as Indonesian are scarce. This study constructs an Indonesian QA dataset that is generated automatically end- to-end. The process begins by taking triples from Wikidata and converting those triples into questions using CFG. The context text is searched from the Indonesian Wikipedia corpus with heuristics to find the appropriate text. These questions were validated with the M-BERT model which functions as a proxy model that assesses the feasibility of questions. The dataset consists of 134 thousand lines of simple questions and 60 thousand complex questions containing two facts in one ques- tion. For simple queries the datasets received similar evaluations by humans (72% AC-IQuAD vs 67% translated SQuAD) and the best Indonesian QA model was the one combining English SQuAD and AC-IQuAD datasets (F1 57.03 against TydiQA dataset).