Peninjauan Kembali Modul-Modul Pemrosesan Bahasa Indonesia dan Pemanfaatannya dalam Membangun Sistem Tanya Jawab = Review of Indonesian NLP Modules and Their Utilization in Question-Answering System

Ageng Anugrah Wardoyo Putra, author

Abstrak

<p>Walaupun belum semaju dan sekomprehensif bahasa-bahasa lainnya, penelitian NLP bahasa Indonesia telah mengalami perkembangan yang cukup signifikan. Penelitian NLP tersebut mencakup POS-Tagging, Named Entity Recognition, dependency parsing, coreference resolution, dan lain sebagainya. Dari penelitian-penelitian NLP bahasa Indonesia yang telah ada, perlu dilakukan validasi dan verifikasi apakah modul NLP pada penelitian tersebut masih relevan atau tidak. Hal tersebut perlu dilakukan karena mungkin saja terjadi kesalahan pada penelitian sebelumnya atau terdapat model yang lebih baik dari penelitian tersebut. Proses tersebut dapat dilakukan melalui evaluasi intrinsik maupun ekstrinsik. Evaluasi intrinsik dapat dilakukan dari reproduksi atau replikasi penelitian yang telah ada, sementara itu evaluasi ekstrinsik dilakukan dengan membangun sistem tanya jawab dari modul-modul NLP tersebut. Hasilnya, didapatkan beberapa modul seperti POS-Tagging dan NER masih cukup relevan dan memiliki dataset yang berkualitas. Namun, beberapa modul lain seperti coreference resolution, constituency parsing, dan dependency parsing masih perlu perkembangan lebih lanjut. Berdasarkan hasil evaluasi, sistem yang dibangun memiliki performa terbaik untuk metrik exact match dan F1 berturut-turut di angka 0,108 dan 0,151 untuk dataset SQuAD, 0,063 dan 0,191 untuk dataset TyDiQA, serta 0,127 dan 0,173 untuk dataset IDK-MRC. Dari evaluasi tersebut diketahui juga bahwa sistem tanya jawab yang dibangun menggunakan pipeline modul-modul NLP tidak sebaik model tanya jawab end-to-end menggunakan BERT yang telah di-finetuning. Meskipun begitu, dari hasil penelitian ini ditunjukkan bahwa kita dapat membangun suatu sistem tanya jawab berdasarkan modul-modul NLP bahasa Indonesia yang tersedia.
.....Although not as advanced and comprehensive as in other languages, research in Indonesian NLP has experienced significant development. This NLP research encompasses POS-Tagging, Named Entity Recognition, dependency parsing, coreference resolution, and other related areas. From the existing NLP studies conducted in the Indonesian language, it is essential to validate and verify whether the NLP modules used in the research are still relevant. This is important because there might have been errors in previous research or there might be better models available. This process can be accomplished through both intrinsic and extrinsic evaluations. Intrinsic evaluation can be conducted by reproducing or replicating existing research, while extrinsic evaluation involves building a question answering system using these NLP modules. The results show that some modules, such as POS-Tagging and NER, are still quite relevant and have high-quality datasets. However, other modules like coreference resolution, constituency parsing, and dependency parsing still require further development. Based on the evaluation results, the constructed system performs best in terms of exact match and F1 metrics, with scores of 0.108 and 0.151 for the SQuAD dataset, 0.063 and 0.191 for the TyDiQA dataset, and 0.127 and 0.173 for the IDK-MRC dataset, respectively. The evaluation also reveals that the question-answering system built using a pipeline of NLP modules does not perform as well as the end-to-end question-answering model using fine-tuned BERT.

Nevertheless, this research demonstrates the feasibility of building a question-answering system based on the available Indonesian NLP modules.</p>