Pengembangan Tolok Ukur Translasi Standar untuk Bahasa Daerah dengan Sumber Data Terbatas di Indonesia = Developing a Standardized Translation Benchmark for Low Resource Local Languages in Indonesia

Lucky Susanto, author

Abstrak

Neural machine translation (NMT) untuk bahasa daerah yang low resource di Indonesia menghadapi tantangan yang signifikan, meliputi kurangnya tolok ukur dasar yang representatif dan ketersediaan data yang terbatas. Penelitian ini mengatasi masalah tersebut dengan cara mengembangkan sebuah tolok ukur dasar yang bersifat replicable untuk empat bahasa daerah di Indonesia yang sering digunakan menggunakan sumber daya komputasi terbatas pada dataset FLORES-200. Penelitian ini mengadakan penyelidikan sistematis dan pemeriksaan menyeluruh terhadap berbagai pendekatan dan paradigma untuk melatih model NMT pada konteks sumber daya komputasi terbatas yang pertama. Tolok ukur ini, dilatih menggunakan sumber daya komputasi dan data pelatihan terbatas, mencapai performa yang kompetitif serta mampu melewati performa GPT-3.5-turbo yang telah di zero-shot untuk berbagai arah translasi dari bahasa Indonesia ke bahasa daerah yang low resource. Penelitian ini berkontribusi kepada kemajuan bidang NMT untuk bahasa-bahasa low resource di Indonesia dan membuka jalan untuk penelitian kedepannya sekaligus mengeksplorasi limitasi GPT-3.5-turbo dalam melakukan translasi bahasa daerah yang low resource. Akhirnya, penelitian ini menunjukkan bahwa melatih model XLM menggunakan data sintetis hasil code-switch memiliki performa translasi diatas pendekatan pelatihan penuh dan pelatihan model XLM dengan data monolingual saja.

.....Neural machine translation (NMT) for low-resource local languages in Indonesia faces significant challenges, including the lack of a representative benchmark and limited data availability. This study addresses these challenges by establishing a replicable benchmark for four frequently spoken Indonesian local languages using limited computing resources on the FLORES-200 dataset. This study conduct the first systematic and thorough examination of various approaches and paradigms for NMT models in low-resource language settings. The benchmark, trained with limited computing power and training data, achieves competitive performance and surpass zero-shot GPT-3.5-turbo in multiple translation directions from Indonesian to low-resource local languages. This work contributes to the advancement of NMT for low-resource Indonesian languages and pave ways for future studies while exploring the limit of GPT-3.5-turbo in translating low-resource local languages. This study shows that training XLM models using synthetic data through code-switching increases translation performance of NMT models down the line compared to just training NMT models from scratch or training XLM models with only monolingual data.