INTISARI

PENDEKATAN INSTRUKSIONAL LINGUISTIK UNTUK PELATIHAN MODEL TERJEMAHAN BERBASIS CONTINUAL INSTRUCTION TUNING DALAM BAHASA MELAYU KUPANG

Oleh

JOANITO AGILI LOPO
23/530909/PPA/06751

Model *Neural Machine Translation* (NMT) sangat bergantung pada data paralel, sehingga kurang efektif untuk terjemahan dalam bahasa-bahasa *low-resource*. Sementara itu, *Large Language Model* (LLM) menawarkan potensi baru dalam terjemahan, namun performanya cenderung menurun pada bahasa yang secara semantik berbeda dari bahasa Inggris. Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan pendekatan Instruksional Linguistik berbasis *Continual Instruction Tuning* (CIT) untuk melatih model LLM ke bahasa Melayu Kupang, dengan memanfaatkan fitur leksikal dan semantik eksplisit dari kamus bilingual. Empat jenis instruksi dikembangkan—berbasis Konteks, Pemetaan Semantik, Fonetik, dan List-Group-Label—yang secara bertahap digunakan dalam pelatihan. Hasil eksperimen menunjukkan bahwa model yang berbasis Instruksional Linguistik menghasilkan peningkatan kinerja signifikan: 12,99 SacreBLEU, 32,01 chrF++, 31,55 ROUGE-L, dan 90,36 TER, dibandingkan dengan model tanpa Instruksional Linguistik yang hanya mencapai 8,97 SacreBLEU, 27,3 chrF++, 25,94 ROUGE-L, dan 97,13 TER. Selain itu, model ini juga cukup efektif dalam skenario *zero-shot* dan *few-shot* prompting, dengan peningkatan 10-13 poin pada metrik evaluasi dibandingkan dengan beberapa model NMT dan LLM. Sebagai tambahan, kemampuan *multitasking* dan *multilingual* model tetap terjaga, termasuk untuk tugas-tugas seperti *sentiment analysis* dan *question answering*. Terakhir, evaluasi secara *human* menunjukkan bahwa terjemahan umumnya akurat secara makna (*adequacy)* meskipun aspek kefasihan (*fluency)* masih dapat diperbaiki. Pendekatan ini menunjukkan potensi pelatihan LLM berbasis CIT untuk meningkatkan performa model terjemahan dalam bahasa *low-resource* seperti Melayu Kupang.

Kata Kunci: Melayu Kupang, Instruksional Linguistik, *Large Language Model*, *Continual Instruction Tuning*, *Machine Translation*

ABSTRACT

AN INSTRUCTIONAL LINGUISTIC APPROACH FOR TRAINING CONTINUAL INSTRUCTION TUNING-BASED TRANSLATION MODELS IN KUPANG MALAY LANGUAGE

By

JOANITO AGILI LOPO
23/530909/PPA/06751

Neural Machine Translation (NMT) models heavily rely on parallel data, making them less effective for low-resource languages translation. In contrast, Large Language Model (LLM) offer new potential for translation tasks but experience performance degradation when translating languages that are semantically distant from English. To address this challenge, this study proposes an Instructional Linguistics approach based on Continual Instruction Tuning (CIT) in the LLM training process using the Kupang Malay language, leveraging explicit lexical and semantic features derived from a bilingual dictionary. Four types of instructional prompts were developed—Context-based, Semantic Mapping, Phonetic, and List-Group-Label—which were introduced incrementally during the training process. Experimental results indicate that the Instructional Linguistics-based model achieved notable scores of 12.99, 32.01, 31.55, and 90.36 on SacreBLEU, chrF++, ROUGE-L, and TER metrics, respectively. In comparison, the baseline model without Instructional Linguistics scored only 8.97 on SacreBLEU, 27.30 on chrF++, 25.94 on ROUGE-L, and 97.13 on TER. Furthermore, the proposed model proved effective in zero-shot and few-shot scenarios, outperforming several NMT and LLM baselines by 10–13 points across the same evaluation metrics. Additionally, the model retained its multitasking and multilingual capabilities, including for tasks such as sentiment analysis and question answering. Lastly, human evaluation indicated that the translations were generally accurate in terms of adequacy, although fluency could still be improved. This approach highlights the potential of CIT-based LLM training for enhancing translation performance in low-resource languages such as Kupang Malay.

Keywords: Kupang Malay, Instructional Linguistics, Large Language Models, Continual Instruction Tuning, Machine Translation