

INTISARI

Optimalisasi Pemilihan Pasangan Bahasa *Source-Target* dalam *Transfer Learning* untuk Mesin Penerjemah Bahasa Daerah Indonesia sebagai *Low-Resource Language* dengan Pendekatan Metrik Kemiripan Linguistik

Oleh

Coveeta Kosambi
21/474894/PA/20505

Penurunan penggunaan bahasa daerah di Indonesia, seperti Bahasa Ngaju, Madura, dan Banjar mendorong perlunya pendekatan inovatif untuk pelestarian bahasa melalui teknologi. Salah satu pendekatan yang potensial adalah pemanfaatan mesin penerjemah yang dirancang khusus untuk bahasa dengan sumber daya rendah atau *low-resource language* (LRL). Tantangan utama dalam menerjemahkan bahasa daerah adalah keterbatasan data paralel yang cukup untuk melatih model secara efektif.

Penelitian ini mengusulkan pengembangan strategi pemilihan pasangan bahasa sumber–target dalam skenario *transfer learning* untuk penerjemahan bahasa daerah Indonesia yang tergolong *low-resource*. Pendekatan ini menggunakan metrik kemiripan linguistik, seperti *Jaccard Similarity*, *Levenshtein Distance*, *FastText Similarity*, dan kombinasinya, untuk menentukan bahasa sumber yang paling relevan secara linguistik sebelum dilakukan proses *transfer learning*. Untuk menguji efektivitas strategi ini, penelitian mengimplementasikan sistem penerjemah berbasis Transformer dengan model arsitektur *pre-trained* mBART50, yang dilatih pada pasangan bahasa Indonesia–beberapa bahasa daerah sebagai *baseline* tanpa *transfer learning*, serta dibandingkan dengan dua skenario, yaitu *Naive Transfer*, dimana pemilihan bahasa sumber berdasarkan jumlah data terbesar dan transfer berbasis metrik kemiripan linguistik.

Eksperimen ini menunjukkan bahwa pemilihan bahasa sumber berbasis kemiripan linguistik memberikan peningkatan performa yang konsisten namun terbatas dibandingkan *baseline* tanpa transfer, dengan kenaikan skor BLEU sebesar +0.25–0.48, ROUGE-L hingga +0.38, dan chrF hingga +0.69 pada beberapa skenario. Hasil terbaik bervariasi menurut bahasa target, namun strategi berbasis *FastText* dan kombinasi metrik cenderung memberikan hasil yang lebih stabil dan unggul dibandingkan *Jaccard* atau *Levenshtein* secara tunggal. Sehingga dapat disimpulkan bahwa mempertimbangkan kemiripan semantik antar bahasa dalam pemilihan pasangan *transfer learning* dapat meningkatkan kualitas terjemahan bahasa daerah *low-resource*, meskipun peningkatannya relatif kecil pada arsitektur mBART50.

Kata-kata kunci: *Neural Machine Translation*, *Natural Language Processing*, *Transfer Learning*, Transformer, mBart *Pre-trained*, *Low-Resource Language*, Bahasa Daerah, BLEU Score

ABSTRACT

Optimizing The Selection of Source-Target Language Pairs in Transfer Learning for Indonesian Regional Language Machine Translation As a Low Resource-Language Using A Linguistic Similarity Metric Approach

By

Coveeta Kosambi
21/474894/PA/20505

The decline in the use of regional languages in Indonesia, such as Ngaju, Madura, and Banjar, has prompted the need for innovative approaches to language preservation through technology. One potential approach is the use of translation engines specifically designed for low-resource languages (LRLs). The main challenge in translating regional languages is the limited amount of parallel data available to effectively train the model.

This study proposes the development of a strategy for selecting source–target language pairs in a transfer learning scenario for translating Indonesian regional languages that are classified as low-resource. This approach uses linguistic similarity metrics, such as Jaccard Similarity, Levenshtein Distance, FastText Similarity, and combinations thereof, to determine the most linguistically relevant source language before the transfer learning process is carried out. To test the effectiveness of this strategy, the study implemented a Transformer-based translation system with a pre-trained mBART50 architecture model, which was trained on Indonesian–several regional language pairs as a baseline without transfer learning, and compared it with two scenarios: Naive Transfer, where the source language was selected based on the largest amount of data, and transfer based on linguistic similarity metrics.

The experiment showed that source language selection based on linguistic similarity provides consistent but limited performance improvements compared to the baseline without transfer, with increases in BLEU scores of +0.25–0.48, ROUGE-L up to +0.38, and chrF up to +0.69 in some scenarios. The best results vary depending on the target language, but strategies based on FastText and metric combinations tend to yield more stable and superior results compared to Jaccard or Levenshtein alone. Therefore, it can be concluded that considering semantic similarity between languages in the selection of transfer learning pairs can improve the quality of low-resource language translations, although the improvement is relatively small on the mBART50 architecture.

Keywords: Neural Machine Translation, Natural Language Processing, Transfer Learning, Transformer, mBart Pre-trained, Low-Resource Language, Regional Languages, BLEU Score