Firoz Shaik CS 521 09/03/2024

Paper Citation: Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. *Improving Low-Resource Languages in Pre-Trained Multilingual Language Models?*

This research presents an unsupervised technique to enhance low-resource language representation in multilingual language models that have already undergone pre-training. It improves language alignment without requiring parallel data by bootstrapping word translation pairs from monolingual corpora. Extensive experiments in nine languages on tasks such as contextual word retrieval and zero-shot named item identification illustrate the usefulness of the strategy with notable gains in both intrinsic and downstream task performance. This work represents a major step forward with its unsupervised method of utilizing non-parallel resources to improve low-resource language support in multilingual models. The thorough testing of the procedure in several languages demonstrates how versatile and widely applicable it is. The method shows significant gains in downstream task performance as well as intrinsic cross-lingual word representation quality. Additionally, by improving language model support for underrepresented languages, the research helps to broaden the inclusivity of NLP technology.

The research identifies areas for development, particularly in improving performance for unknown languages, where preliminary findings are encouraging but modest. The pre-trained models (mBERT and eBERT) beginning quality limits the mining process effectiveness, indicating the necessity for techniques that can make use of more easily accessible cross-lingual resources to produce appreciable improvements. Furthermore, the need for parallel sentences to yield significant benefits highlights the difficulty of limited resources for many low-resource languages, suggesting the need to investigate alternate approaches that do not rely on large amounts of parallel data. The research methodology employs an iterative, unsupervised strategy to enhance low-resource language representations in multilingual language models, particularly without the need for parallel data. The strategy efficiently aligns language representations by mining word translation pairs from monolingual corpora and repeatedly fine-tuning model parameters with a dedicated linear layer. This procedure marks a considerable departure from the usual reliance on parallel corpora and emphasizes the creative use of non-parallel resources while focusing on improving cross-lingual quality through straightforward yet effective strategies.

Two main evaluation methods are used in the study: **zero-shot named entity recognition (NER)** and **contextual cross-lingual word retrieval**. The quality of cross-lingual word representations is evaluated by the former, which gauges the model's capacity to match word pairs across languages. To show how the enhanced language representations are useful in real-world scenarios, the latter assesses the model's ability to identify named entities in a language it wasn't expressly trained on. The effectiveness of the strategy in improving multilingual models for low-resource languages is demonstrated by utilizing accuracy for word retrieval and F1 scores for NER to quantify performance. The research paper's conclusion highlights the effectiveness of an unsupervised technique for improving multilingual language models for low-resource languages in the absence of parallel data. One interesting conclusion is that the amount rather than the quality of the word pairs that are mined has a major impact on model improvements. To support even more languages, the report also recommends future research into the use of easily accessible cross-lingual materials. By presenting an unsupervised technique that improves the representation of low-resource languages in multilingual models and shows notable performance improvements, the research makes a major contribution to NLP. It makes efficient use of non-parallel resources and provides a scalable way to increase NLP inclusivity.