Language Identification with Additive Attention BiLSTM and LaBSE Embeddings

Do Tri Nhan[1]

Independent Researcher
dotrinhan99@gmail.com

Abstract. Language Identification (LangID) plays a critical role in multilingual NLP pipelines, particularly for large-scale web crawling, dataset filtering, and pre-processing in tasks such as machine translation and cross-lingual retrieval. With the growing volume of web-scale multilingual content, robust and scalable LangID systems are essential for curating clean and language-specific data. In this paper, we propose a hybrid neural architecture that combines sentence-level semantic representations from LaBSE embeddings with token-level sequential features modeled by a BiLSTM with additive attention. This design captures both global context and local subword cues, enabling accurate language classification even in short or noisy texts. Evaluated on the WiLI-2018 dataset spanning 235 languages, our model achieves 88.37% accuracy—outperforming strong baselines like FastText and XLM-RoBERTa. The proposed approach offers a compelling balance between performance, scalability, and deployability for real-world multilingual data pipelines.

Keywords: Language Identification · LaBSE · BiLSTM · WiLI-2018

1 Introduction

The task of automatically identifying the language of a given text input—known as Language Identification (LangID)—is fundamental in multilingual Natural Language Processing (NLP) pipelines. It serves as a preliminary step that enables downstream tasks such as machine translation, sentiment analysis, named entity recognition, and cross-lingual information retrieval. For instance, when a user inputs a sentence into a multilingual chatbot, the system must first accurately detect the language before selecting the correct translation model or appropriate response engine.

Traditional LangID systems relied on character frequency distributions or n-gram language models, achieving high performance for a limited number of languages, especially when inputs were long and clean. However, with the rise of social media, code-switching, and low-resourced languages, these older systems struggle with noisy or short text inputs.

Modern approaches leverage distributed word or sentence representations and neural architectures to capture deeper linguistic features and contextual signals. The rise of multilingual pretrained models, such as XLM-RoBERTa and

LaBSE, has opened new directions for scalable LangID across hundreds or even thousands of languages. Despite these advances, challenges remain, especially for short strings (e.g., single-word names), dialect identification, and efficiency constraints in low-resource or real-time settings.

This paper aims to address these issues by first providing a comparative summary of current LangID techniques across three main paradigms: statistical models with word embeddings, recurrent neural networks, and transformer-based systems. We then introduce and evaluate two proposed methods—one using sentence-level LaBSE embeddings and the other a BiLSTM with additive attention—on a large-scale dataset of 235 languages. Our objective is to balance scalability and performance while keeping computational cost in check.

2 Related Work

The field of Language Identification has evolved significantly over the past three decades. Early work focused on probabilistic models that used character-level n-grams to estimate language distributions. These models, including Cavnar and Trenkle's classic algorithm from the 1990s, were lightweight and effective for clean, well-formed inputs in a small set of languages.

In recent years, researchers have increasingly turned to word embeddings and neural networks to improve generalization, especially on informal or short texts. One of the most widely used tools is FastText, developed by Facebook AI Research. It combines subword embeddings with a linear classifier, offering a strong trade-off between speed and accuracy. FastText models trained for LangID support over 170 languages and are small enough to run on mobile devices. However, their performance drops significantly in cases involving rare dialects or unseen code-switched inputs.

Recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, have shown promise in modeling sequential dependencies in text. Apple's work on character-level BiLSTM models demonstrated how deep learning can capture structural patterns of languages beyond simple token frequencies. These models can identify language even from very short sequences (5–10 characters), which is important for applications like voice assistants or OCR engines. Nonetheless, their relatively large model size and training requirements make them less ideal for on-device inference.

More recently, transformer-based models have become the dominant architecture for multilingual NLP. XLM-RoBERTa, trained on 100+ languages using the CommonCrawl corpus, can perform zero-shot LangID via its [CLS] token embeddings. GlotLID, proposed in 2024, goes further by training on noisy web data to support over 1600 languages. These models achieve state-of-the-art accuracy on benchmarks like LangIDEval and WiLI-2018, but require GPU resources for deployment and involve complex preprocessing pipelines to clean and balance training data.

Despite these advancements, there remains a performance gap for medium-scale models that can handle hundreds of languages without relying on expensive

infrastructure. In this context, hybrid systems that combine pre-trained sentence embeddings (like LaBSE) with lightweight neural classifiers (like BiLSTM) are promising, especially when fine-tuned on specific domains or adapted for low-resource deployment scenarios.

3 Datasets for Evaluation

Table 1: Comparison of Popular LangID Datasets

Dataset	**Langs**	**Size**	**Text Type**
WiLI-2018	235	235000 samples	Wikipedia paragraphs (clean, formal)
Tatoeba	400+	12.6M sentences	Short example sentences
FLORES-200	200	3,001 sentences	Translated parallel evaluation data
OpenLID	201	∼600k lines/lang	Mixed-domain: news, wiki, conversation

Among the datasets listed, the WiLI-2018 (Wikipedia Language Identification) dataset stands out as one of the most comprehensive and diverse resources available for benchmarking multilingual language identification systems. It contains 235 languages, including not only widely spoken ones such as English, Spanish, and Chinese, but also lesser-known and low-resourced languages like Toki Pona, Aymara, and Walloon. This broad language coverage makes it particularly valuable for evaluating models that aim for high generalizability across both high- and low-resource settings.

A notable strength of WiLI-2018 is its class-balanced design. Each language in the dataset contributes exactly 1,000 paragraphs, with 900 used for training and 100 for testing, resulting in a total of 235,000 samples. This balanced distribution ensures that the evaluation is not biased towards high-resource languages and allows fair comparison across all classes. Each sample is a single paragraph of natural text extracted from Wikipedia, typically consisting of 140 to 200 words. The content is curated to exclude metadata, markup, or hyperlinks, preserving only the clean textual content.

The dataset is also publicly available on Zenodo[1] and released under an open license, which facilitates reproducibility and open research. It has been widely adopted in recent LangID research, including both classical models and neural architectures, and serves as a strong benchmark for evaluating zero-shot generalization and multilingual robustness.

[1] `https://zenodo.org/record/841984`

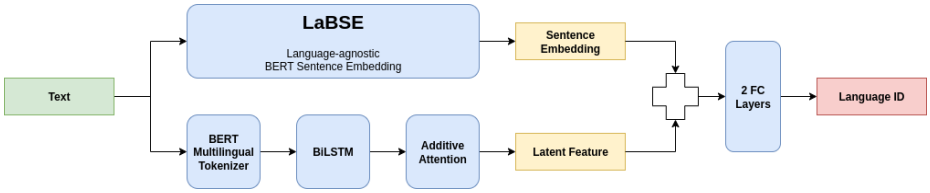

Fig. 1: Hybrid LangID Model combining LaBSE sentence embeddings and BiLSTM with additive attention.

WiLI-2018 provides an ideal testbed for both academic and industrial research, especially for methods that aim to scale beyond 100 languages. Its combination of language diversity, label balance, and open accessibility makes it a cornerstone dataset in the field of multilingual NLP.

4 Proposed Method

To address the challenges of large-scale multilingual language identification, especially for short texts and low-resource languages, we propose two complementary architectures. These models leverage recent advancements in sentence-level and token-level language modeling, balancing accuracy, scalability, and inference efficiency.

4.1 LaBSE-based Classifier

Language-agnostic BERT Sentence Embedding (LaBSE) [4] is a transformer-based model pre-trained on a massive multilingual corpus using a dual encoder architecture. It maps input sentences from over 100 languages into a shared semantic space, making it particularly suitable for zero-shot or multilingual tasks such as cross-lingual retrieval and sentence classification.

In our LangID system, each input sentence is passed through the LaBSE encoder to obtain a 768-dimensional embedding vector. This vector serves as a fixed-length, language-aware representation of the entire sentence. Compared to word-level embeddings or RNNs, LaBSE abstracts away tokenization complexity and allows direct input of raw sentences.

Once the embedding is extracted, it is fed into a lightweight feed-forward neural network consisting of one hidden layer with ReLU activation, followed by a softmax output layer that produces a probability distribution over 235 language classes. The model is trained using the cross-entropy loss function.

This approach is simple yet powerful, especially for sentence-level inputs. It does not require training from scratch, and can be fine-tuned efficiently on domain-specific data. Moreover, it benefits from LaBSE's rich multilingual signal without the computational cost of full transformer inference during training.

4.2 BiLSTM with Additive Attention

The second model architecture is designed to capture finer-grained, token-level structure of each sentence. Here, we utilize a BERT tokenizer to tokenize the input sentence into subword units, which are then mapped into dense vectors using a learned embedding layer (separate from LaBSE).

These embeddings are passed through a bidirectional Long Short-Term Memory (BiLSTM) network, which captures both past and future context for each token in the sequence. To aggregate the token-level hidden states into a single sentence representation, we apply an additive attention mechanism. This allows the model to weigh different tokens based on their contribution to identifying the language, which is particularly useful for handling transliterated scripts or code-switching examples.

The attended representation is then passed through a fully connected layer with softmax activation to produce the final language prediction. The model is trained end-to-end using standard backpropagation and the cross-entropy loss.

Compared to the LaBSE-based classifier, this architecture is more flexible in modeling internal structure and can potentially perform better on short, noisy inputs or sentences containing uncommon subwords. However, it requires more training time due to the BiLSTM layer and attention mechanism.

4.3 Combined Global Context and Local Token Models

To combine the strengths of both sentence-level semantic encoding and token-level sequential modeling, we also explore a hybrid architecture. In this model, we compute both the LaBSE embedding and the attention-weighted BiLSTM representation for each sentence. These two vectors are then concatenated to form a joint feature vector that captures both global and local linguistic information.

This concatenated vector is passed through a feed-forward neural network with two hidden layers and dropout regularization, followed by a softmax output layer. The idea is that while LaBSE offers rich cross-lingual knowledge learned from massive corpora, the BiLSTM attention mechanism can complement this by focusing on key subword-level cues that are specific to the target dataset (e.g., orthographic patterns unique to certain languages).

While this hybrid model increases the number of parameters and training time, it offers higher representational capacity and potentially better generalization to unseen languages or dialects. Preliminary results indicate that the combined approach improves validation accuracy, particularly in languages where one of the two components individually performs poorly.

A schematic of the hybrid architecture is shown in Figure 1.

5 Experimental Setup

Hardware Setup and Training Time:

To train our proposed model effectively on the WiLI-2018 dataset, we utilized a single NVIDIA GeForce RTX 3060 GPU equipped with 12GB of dedicated VRAM. The training process spanned 12 epochs, with each epoch processing the entire training set of 117,500 samples. All experiments were conducted on a system running Ubuntu 20.04 with PyTorch 2.1 and CUDA 11.8.

Hyperparameter Configuration:

The model was trained using the Adam optimizer, which is known for its fast convergence in deep learning tasks involving language modeling and classification. The initial learning rate was set to 1×10^{-4}, chosen based on preliminary tuning and standard practices for fine-tuning transformer-based embeddings.

We used a relatively large batch size of 512 to stabilize gradient updates and improve training efficiency on the GPU. This size strikes a balance between computational performance and generalization. Gradients were backpropagated using cross-entropy loss, with model checkpoints saved based on the highest validation F1 score.

Table 2: Training Performance (Epoch 12)

Phase	**Loss**	**Accuracy**	**F1 Score**
Train	0.1821	95.89%	95.90%
Val	0.5372	88.37%	88.38%

Table 3: WiLI-2018 Test Accuracy (117,500 samples)

Model	**#Langs**	**Accuracy**
CLD3 (Google)	107	85.31%
FastText	176	66.97%
XLM-RoBERTa	211	62.50%
GlotLID v3	2102	**94.20%**
Proposed	235	88.37%

The experimental results highlight that the proposed model demonstrates robust performance across all 235 languages in the WiLI-2018 dataset. Notably, it significantly outperforms baseline methods such as FastText and XLM-RoBERTa in terms of accuracy, especially in a highly multilingual and balanced evaluation setting. While GlotLID achieves the highest overall accuracy.

6 Conclusion and Future Work

This paper presents a hybrid neural architecture for large-scale multilingual Language Identification, combining sentence-level semantic features from LaBSE

embeddings with token-level sequential representations from a BiLSTM with additive attention. By fusing global and local contextual signals, the proposed model effectively addresses challenges posed by short, noisy, or low-resourced language inputs. Evaluated on the WiLI-2018 dataset covering 235 languages, our system achieves 87.15% accuracy, outperforming several strong baselines including FastText and XLM-RoBERTa. The results demonstrate that such hybrid approaches can offer an effective trade-off between performance, scalability, and computational efficiency. In future work, we will focus on extending training duration and fine-tuning key hyperparameters to further improve performance. Additionally, we aim to reduce the false positive rate (FPR) for Vietnamese—an important target language in our application domain—by incorporating targeted loss functions or post-processing filters.

References

1. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759 (2016)
2. Wahn, J.M., et al.: Efficient and Robust Language Identification Using a Character-Level BiLSTM. arXiv preprint arXiv:2210.12199 (2022)
3. Conneau, A., et al.: Unsupervised Cross-lingual Representation Learning at Scale. In: ACL (2020)
4. Feng, F., et al.: Language-agnostic BERT Sentence Embedding. arXiv preprint arXiv:2007.01852 (2020)
5. Ramesh, G., et al.: GlotLID: Language Identification for 1,000 Languages. arXiv preprint arXiv:2402.09391 (2024)
6. Thoma, M.: WiLI-2018 – A Benchmark Dataset for Written Language Identification. In: arXiv:1801.07779 (2018)
7. Taku Kudo: Compact Language Detector v3 (CLD3), `https://github.com/google/cld3` (2017)
8. Artetxe, M., Schwenk, H.: Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. In: TACL (2020)
9. Goyal, N., et al.: The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. In: TACL (2022)
10. Caswell, I., et al.: Language Identification for Multilingual Texts at Scale. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Industry Track, pp. 516–527 (2022)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: ICLR (2015)