Language identification : A Study of the GlotLID-M Model - NLP Mini-Project

Hugo Capot
ENSAE Paris
hugo.capot@ensae.fr

Abstract

Language identification is a crucial task for construction of correct datasets, and so extremely important for better model in NLP. The purpose of this mini-project is to study the GlotLID-M model, introduced in the paper "GlotLID: Language Identification for Low-Resource Languages", Kargaran et al., 2023, [4], which improves language identification of low-resources languages (*i.e.* languages with only a few samples available online). Our study aims at better understanding the construction of the model by easing its architecture, and then to evaluate its performances on new test datasets. The code is given in the form of a Jupyter Notebook, available on the following GitHub repo : https://github.com/HugoCapot/Natural-Language-Processing-ENSAE-3A.git.

1 Introduction and state-of-the-art

When performing NLP tasks—such as the language models studied in class—it is essential to operate within a specific language (for example, one wouldn't expect a model to predict Spanish words after English ones). Therefore, having access to well-labeled textual datasets by language is crucial for enabling more effective NLP applications.

However, collecting large-scale data from the internet and automating the labeling process requires algorithms capable of language identification. This is the core of the following study here: developing an algorithm that, given a document, can accurately determine the language it is written in. The language identification methods is a sub-task of language classification. The first intuition when it comes to identify a language is to use a statistical method by was based on the prevalence of certain function words (e.g., "the" in English ; technique defined in [1] as "Trigram technique"), or distinctive diacritics or punctuation [6]. The method that prevails today is inspired of this intuitions : it consists in representing sentences by sequences of n-grams (which is to say that each word is represented as a set of sub-words of length n), and then to apply classification on the embedded vectors.

If many progress have been made in improving accuracy of language identification for the most spoken languages (with FT176,5 CLD3, NLLB (NLLB Team et al., 2022) and OpenLID), the question of the accuracy of language identification for low-resources languages (ie languages with only a few documents available) remain to be solved. To tackle this issue, two aspects are very important : the construction of a good dataset with as many labeled languages as possible, and the construction of a good classification model. A good language identification model then must be able to cover a large number of languages with limited resources, must be easy to use, and its performances must be easily and reliably measured.

This project is based on the GlotLID-M model, introduced in the paper "GlotLID: Language Identification for Low-Resource Languages", Kargaran et al., 2023 [4]. The paper aims at improving language identification for underrepresented languages in online content—such as dialects that lack a large number of written documents. The authors first construct a corpus (GlotLID-C) of labeled sentences with the associated language, and then use it with a FastText classifier to construct a new language identification model (GlotLID-M). The conclusion of the paper is that the GlotLID-M model outperforms four classic language identification systems—CDL3, FT176, OpenLID, and NLLB—in terms of F1 score and false positive rate, particularly for low-resource languages.

The objective of this project is double. First, pedagogical : in order to present the core of the algorithm, a simplified version of GlotLID-M is built, with a simplified implementation of a FastText-based classifier, and a training made on a much smaller dataset than GlotLID-C, using only WiLI-2018 [7]. Second, experimental : the original GlotLID-M model is evaluated on the same datasets used in the paper (e.g., reproduce their results), and then test it on another public language identification dataset to compare performances.

2 Simplified Construction of GlotLID-M

The first part of the project aims at constructing a small version of the GlotLID-M model from scratch, in order to better understand it, searching for both a simple dataset and a simple architecture.

2.1 Small construction of FastText classifier

FastText is a ML algorithm designed for text classification. It has been introduced by the Facebook AI team in 2017 [3]. The point is first to represent the text by a sequence of words, each word being represented as a series of n-grams. This n-grams are then embed in a "word latent space", and then the text are themselves embed in a "text latent space", using the average of the embeddings of the n-grams in the text text. Classification is then just made in order to classify the embeddings. The chosen loss is the negative log-likelihood.

The structure of FastText is quite simple (just a linear classifier), which enables the execution time to be short, compared to other DeepLearning techniques [3].

The objective of the first part of the project was to implement a manageable version of FastText from scratch. For this, two main classes were implemented : the first one aim are converting a text into a sequence of n-grams, making it possible to convert the text into a sequence of n-grams, and then another one which mimics the FastText implementation.

2.2 Training Dataset

For building the training dataset of this project, the permission was asked to the authors of the article [4] to use their corpus. They accepted the demand, but even with the data available the dataset was too big to handle. The decision has then been made to take a smaller corpus: WiLI-2018, a subset of corpus GlotLID-C [1].

WiLI-2018 (Wikipedia Language Identification) is a publicly available benchmark dataset designed for monolingual written language identification tasks [7]. It consists of 1,000 paragraphs for each of 235 languages, sourced from Wikipedia articles. Each paragraph contains at least 140 Unicode characters and is stripped of markup, punctuation, and metadata to ensure clean and standardized

[1] To see all the resources of the GlotLID-M model, look at the following GitHub repo : https://github.com/cisnlp/GlotLID.git

input. All languages are labeled using their ISO 639-3 codes (ie the official international code for languages), allowing compatibility with multilingual NLP pipelines.

The dataset is well-suited for supervised training and evaluation of language classifiers. It offers a relatively balanced distribution across languages, although the quality and representativeness of data can vary due to inconsistencies in Wikipedia content volume across languages [2]. Because of its manageable size and broad coverage, WiLI-2018 is frequently used as a baseline dataset for benchmarking models like FastText, langid.py, and more recently, GlotLID.

WiLI-2018 is really small (235 languages) compared to the corpus GlotLID-C (1832 including 1665 well-performing languages). While WiLI-2018 covers a wide range of scripts and language families, it does not include code-switching, and its coverage of dialects and minority languages is limited compared to newer datasets such as GlotLID-C. However, its ease of use and standardized format make it an ideal dataset for prototyping and testing language identification models in academic and industrial settings.

2.3 Evaluation

In order to quantify the performances of the model, a metric had to be chosen. In classification, classic metrics are accuracy, precision, recall and F1 score.

In the paper, Kargaran *et al.* use the F1 score and the false positive rate, defined this way - for each class (ie each language):

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{and} \quad \text{FP-rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

where

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \text{and} \quad \text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In order to make our results comparable to the ones of the paper, it has been decided to use the same metrics.

Remarks :

- In the paper, the authors choose not to use the accurary ratio because the classes are highly imbalanced. Here, because the WiLI-2018 dataset use way less languages than GlotLID-M and because the number of examples in each class is equal (see below), we could have use the accuracy ratio too.
- Because of the large number of classes, a raw confusion matrix wouldn't be really helpful here, because one wouldn't be able to read it.

As it appears on Figure 1, classification performs quite well on the majority of languages, with more than 320 languages whose performances are above 0.8, and really good prediction for a lot of them (see Table 1 for examples). However, there are still 15 languages with a F1 score lower than 0.8, with some of them really near 0 (see Table 2).

Remark : Of course, for a further study, the names of the languages in the Tables 1 and 2 should had been add instead of the WiLI-2018 labels only, but there has been an obvious problem in associating each language with its name, and the problem couldn't be solved.

[2] https://www.theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikip

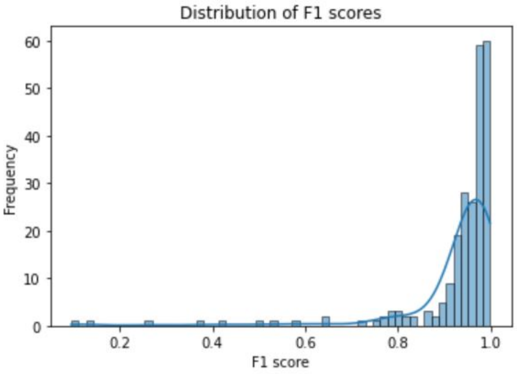

Figure 1: Distribution of F1 scores on the 235 classes of the WiLI-2018 test set.

Label	F1 score	Precision score	Recall score	False positive rate
153	0.999	1	0.998	0
22	0.999	1	0.998	0
70	0.999	1	0.998	0
215	0.999	1	0.998	0
189	0.997	0.996	0.998	0

Table 1: Some of the best performances when it comes to F1 score on the WiLI-2018 test set.

Label	F1 score	Precision score	Recall score	False positive rate
230	0.096	0.192	0.064	0.001
84	0.13	0.259	0.88	0.001
179	0.26	0.408	0.19	0.001
155	0.37	0.470	0.31	0.001
67	0.42	0.527	0.354	0.001

Table 2: Worst performances when it comes to F1 score on the WiLI-2018 test set.

2.4 Discussion

Much improvement has to made on this construction to reach the construction of the GlotLID-M model.

First of all, concerning the FastText architecture, the code hasn't been optimized and is quite long to run. In order to improve efficiency, it could be possible to use an implementation by batch, and not sentence by sentence, or to use a GPU instead of a CPU.

Concerning the corpus, much has yet to be done in order to get similar performances to the GlotLID-C corpus. Indeed, the choice of the training dataset is key for the use of the model :the article from Kargaran *at al.* explains that even if language datasets—including lesser-known languages—have been proposed, such as CC100, mc4, and Oscar, they all had quality issues before GlotLID-C, particularly with low-resource languages. The most frequent errors were found in these underrepresented languages, which inevitably led to noisy data and degraded performance in downstream NLP tasks. The authors' conclusion is clear: for NLP to perform well on low-resource languages, high-quality datasets are essential, which in turn requires high-quality language identification. Then here, with the choice of WiLI-2018, only 235 languages can be identifies at most, instead of the 1,665 languages available in the GlotLID-M model. In fact, our implementation here with a small corpus is closer to the FT176[3] model than to GlotLID-M.

[3] https://fasttext.cc/docs/en/language-identification.html

In addition to this, granularity is not handled at all in the proposed model: macro-languages and their variants aren't distinguished at all in the WiLI-2018 way of labeling sentences. A further study of the ISO 639-3 nomenclature should be conducted if it were to manage this issue. With such an additional structure, it would have been interesting to study the robustness of the model to out-of-distribution cousin languages, and to check that the model really helps improve performance even on high-resource languages, with a choice of the same corpus with and without the cousin languages.

3 GlotLID-M evaluation

The second part of the project is to try to evaluate the model GlotLID-M as constructed on the paper of Kargaran *et al.*, first by reproducing the results on the UDHR dataset, and then by applying the model to a new test set.

3.1 Evaluation on UDHR

The Universal Declaration of Human Rights (UDHR) is a foundational international document adopted by the United Nations General Assembly in 1948. It outlines a broad range of fundamental human rights and freedoms to which all individuals are entitled, regardless of nationality, race, religion, or language. As one of the most widely translated documents in the world, the UDHR serves as a valuable resource for multilingual and cross-cultural studies, including language identification tasks.

The UDHR-LID dataset [4], developed by CIS at LMU Munich and available on Hugging Face, is a multilingual corpus specifically designed for language identification (LID) tasks. It comprises approximately 27,800 sentences, each corresponding to a translation of a sentence from the Universal Declaration of Human Rights (UDHR) in one of 419 languages. Each data entry includes the sentence itself, the ISO 639-3 language code, and the ISO 15924 script code, making the dataset highly structured and useful for multilingual NLP research. With broad linguistic coverage—including low-resource languages such as Tigrinya and Balkan Romani—this dataset supports the development and evaluation of robust LID models across diverse linguistic typologies. It is distributed in CSV format and can be accessed programmatically using the Hugging Face datasets library. The dataset is released under a CC0-1.0 license, encouraging open use and research.

Following the paper's methodology (and as it has been done in the first part), the model was evaluated on the UDHR corpus using macro F1 score, with detailed precision and recall. The evaluation was performed using the top-1 prediction only.

The performances of the GlotLID-M model are excellent for a high majority of the languages, as one can see on Figure 2. The worst performances of the model are detailed on Table 3.

Label	F1 score	Precision score	Recall score	False positive rate
Chimborazo Highland Quichua	0.011	0.33	0.005	0
Quechua	0.018	0.010	0.09	0
Waorani	0.026	1	0.01	0
Secoya	0.029	0.14	0.02	0
Central Atlas Tamazight	0.037	0.02	1	0

Table 3: Worst performances of the GlotLID-M model when it comes to F1 score on the UDHR test set.

[4] https://huggingface.co/datasets/cis-lmu/udhr-lid

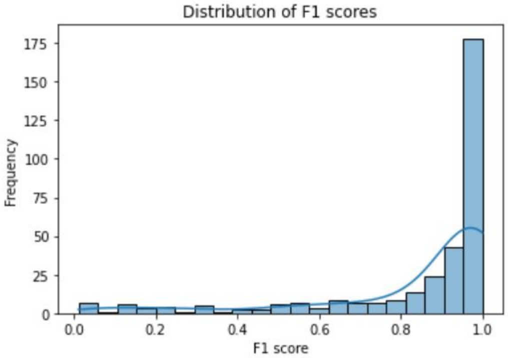

Figure 2: Distribution of F1 scores of the GlotLID-M model on the 235 classes of the UDHR test set.

3.2 Evaluation on GlotSparse

The second goal of this second part was to use a new test dataset. The idea was to use one of the dataset presented in the paper *which hasn't been used* for the construction of the training dataset. The absence of overlapping with the training dataset is really important, to avoid an overestimation of the performances of the model. Among the datasets which fulfilled this condition, the goal was to choose a dataset with low-resources languages, because that is exactly the type of data the model has been constructed for. Hence, the choice has been made to take the GlotSparse dataset [5].

The GlotSparse benchmark, introduced alongside GlotLID, is designed to measure performance on closely related (or cousin) languages and in sparse-resource conditions. It consists of ten languages, with both monolingual and code-switched samples, making it particularly suitable to test the model's robustness and granularity. To simulate real-world noisy data, the GlotSparse corpus contains sentences where language boundaries are ambiguous or where multiple dialects are involved. The evaluation was conducted by comparing predicted language codes against the gold-standard ISO 639-3 tags.

As we can see on Table 4, The performances of the GlotLID-M model are excellent for five out of the ten languages, which is encouraging for the performances of the model on low-resources languages. For the others however, the model hasn't be able to identify them at all, which calls for a deepen analysis of the predictions.

Language	F1 score	Precision score	Recall score	False positive rate
Southern-Kurdish	0.004	0.93	0.002	0
Balochi	0.554	0.99	0.38	0
South-Azerbaijani	0.705	0.59	0.88	0.1
Gilaki	0.78	0.93	0.68	0
Twi	0.926	1	0.86	0
Fanti	1	1	1	0

Table 4: Best performances of the GlotLID-M model when it comes to F1 score on the GlotSparse test set.

3.3 Discussion

A further study would have to be made in order to really compare the contribution of the model on low-level languages. A deep comparison with a benchmark of LID models such as OpenLID would also help to see the real benefit of GlotLID-M even for high-resource languages.

[5] https://huggingface.co/datasets/cis-lmu/GlotSparse

It was underlined that the lack of overlapping between the test set and the training set was really important. However, in the previous work, this lack of overlapping was *supposed* (when both UDHR and GlotSparse have been used), but not checked. A more rigorous study could use the GlotLID-C model (which we have access to, but which was too big to handle), and check that there is no sentence that appears in both dataset.

Other modifications of the experiments suggested in the paper - like adding the variation of temperature- could be good ideas to enrich the study.

4 Conclusion

This project provided an in-depth exploration of GlotLID-M, a model specifically designed to address the challenge of language identification for low-resource languages. The project was structured around two core components: (i) pedagogical replication with a simplified FastText-based classifier and (ii) empirical evaluation of GlotLID-M on public benchmarks.

The simplified FastText model was useful to understand how n-gram embeddings and averaging can be leveraged for language classification, although its performance remains limited compared to larger, more robust models like GlotLID-M. The training dataset (WiLI-2018) proved suitable for experimentation, but its coverage (235 languages) is significantly smaller than GlotLID-C's (1665), limiting the ability to test the full generalization capacity.

The evaluation on UDHR and GlotSparse confirmed good performances of the model on many languages, even if a further study would have to be conducted in order to confront all the paper's claims.

However, several limitations remain:

- The full training on GlotLID-C was not feasible due to hardware constraints. Future work could explore training subsets with better sampling strategies or cloud-based training.
- Further experiments could focus on evaluating GlotLID-M's ability to handle code-switching, which is especially common in real-world multilingual contexts.
- It would also be relevant to perform qualitative error analysis, for example by identifying language pairs that are frequently confused, and investigate why.

One interesting future direction would be to fine-tune GlotLID-M on specific language groups (e.g., Niger-Congo, Indo-Aryan) to improve specialization, or to incorporate unsupervised clustering to detect unseen or unlabeled languages.

Acknowledgments: The author thanks the authors of the GlotLID-M paper for granting access to the training corpus. Special thanks also go to the NLP teaching staff for their guidance throughout the project, and for this very enriching semester.

References

[1] Gregory Grefenstette. Comparing two language identification schemes. In *Proceedings of JADT*, volume 95, pages 263–268, 1995.

[2] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg, 1998. Springer.

[3] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[4] Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. GlotLID: Language Identification for Low-Resource Languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore, December 2023. Association for Computational Linguistics.

[5] Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. GlotScript: A Resource and Tool for Low Resource Writing System Identification, March 2024. arXiv:2309.13320 [cs].

[6] Wolfgang G. Stock and Mechtild Stock. *Handbook of Information Science*. De Gruyter Saur, 2013.

[7] Martin Thoma. The WiLI benchmark dataset for written language identification, January 2018. arXiv:1801.07779 [cs].