ANLP Competition Report

Lina Ben-Younes, Mattéo Debart, Nathan Janiec, Paul Massey, Samuel Sithakoul

Abstract

Language Identification is an essential aspect of almost every Natural Language Processing (NLP) pipeline. However, this task appears to be extremely fastidious when it comes to rarer languages, for which obtaining a clear and unbiased dataset is particularly difficult. This work compares the performance and computational cost of different methods for language identification.

Keywords– NLP - Language Identification

1 Introduction

We analyse a dataset containing 190,599 labelled phrases and 190,567 unlabelled phrases, covering 389 languages. The dataset includes widely spoken languages (only 2 % of the labelled data) as well as regional dialects. Notably, 92% languages have 200 or more sentences, and over 98 % of phrases contain less than 100 words. The distribution is sensibly similar in the test dataset.

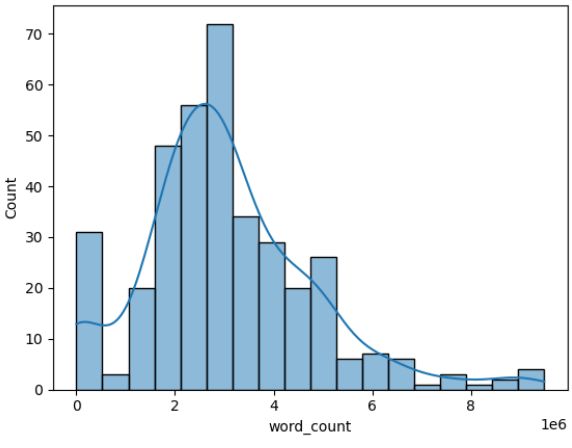

Figure 1: Distribution of word count per language in the training dataset.

The vast majority of languages are represented by less than a million words, and close to 10% of the dataset is represented by less than a hundred thounsand words.

2 Explored Approaches

2.1 Statistics study

The first two approaches weren't expected to have great performances. They were mostly tried to establish a lower bound of the results that could be achieved. Those methods are purely deterministic approaches based on statistics computed from the empirical observations among the training Dataset.

2.1.1 Word frequency

The first naïve approach considered to solve this problem was to look at the "small words". The intuition is that these are the most relevant words to look for in a language because they are often the most frequent ones in a given language. We then built a dictionary listing the frequency of languages for each given word containing less than n letters. For inference, we just had to sum all the frequencies for every small word in the sentence to find the most likely language predicted by the model. After splitting our dataset into training, validation, and testing datasets, we were able to determine that $n = 7$ letters was the optimal length for the words to consider, achieving an accuracy of 45%.

2.1.2 Inverse Word frequency

Some languages in the dataset share strong etymological roots. To address this, we utilized the Term Frequency-Inverse Document Frequency method to weight the most frequent words within each language, taking into account their specificity to the language. The training set was reorganized into a corpus of 389 documents, one for each language and we performed a cosine similarity search for each sentence in the testing set. This approach yielded an accuracy of 72%. To address the sparsity of the TF-IDF matrix, we applied a PCA. We reduced the dataset to 389 dimensions, while maintaining the accuracy of the model.

2.2 Finetuned Language Model

To try and improve the results for this task, we decided to use a Language Model and adapt it to our specific task. XLM-RoBERTa (Conneau et al., 2020) is a multilingual model trained on 100 different languages from CC-100 corpus and its tokenizer has a vocabulary size of 250 000 tokens which we deemed sufficient to represent the words in the competition dataset. It is based on the architecture of RoBERTa with masked language modeling objective (MLM) and translation language modeling (TLM) objectives that are alterned during the pre-training. Unlike some other multilingual models, it does not require lang tensors added in the input to understand which language is used, and should determine the correct language directly from the input tokens.

Other models were considered within the state-of-the-art models in multilingual classification (essentially based on the XTREME benchmark) like mT5 or RemBERT but XLM-RoBERTa was selected for its simplicty of use in classification task as an encoder model, relatively small size compared to encoder-decoder or decoder models (550M parameters) and overall good performance.

We decided to finetune this model on our training dataset to achieve the desired task. We trained the model on 80% of the training dataset for 20 epochs and compute the accuracy at each epoch on 10% if the data as the validation split. We also compute the accuracy on a 10% test dataset before submitting our results on the public dataset. At the end of the hyperparameters selection process, we finally achieved a **90.2%** accuracy score on the test dataset.

3 Results

To compare the results of our approaches, we evaluated each model on the same validation and test sets extracted each from 10% of the labelled training set.

Model	Accuracy (%)
Word frequency	45
Inverse Word frequency	72
RoBERTa	90.2

Table 1: Comparison of model accuracies on a test set

Looking at the RoBERTa results in particular, we found that the correlation between the F1 score and the number of words in our training dataset, or the number of tokens in the XLM-RoBERTA dataset was too weak to be considered significant (below 5%). However, when examining the correlation between the F1 score and the number of words, focusing only on languages with fewer than 1,000 words, the correlation was found to be 20 %. These results suggest that, beyond a certain threshold, the number of words in the training set no longer is the decisive factor in the classifier's performance for a given language.

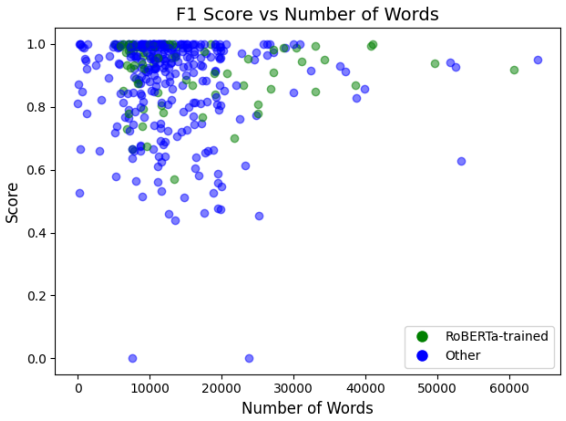

Figure 2: Relationship Between F1 Score and Word Count Across Different Languages.

To further investigate low F1 scores, we examined the confusion matrices for languages with a large number of sentences. We found that predictably they had a high rate of confusion with languages that share strong etymological roots. For example, Croatian, Bosnian, and (to a lesser extent) Serbian, which are closely related, were frequently confused with each other [2]. Both Croatian and Bosnian are mutually intelligible as part of the Serbo-Croatian language family. Similarly, Shona and Kinyarwanda are both Bantu languages, with Kinyarwanda classified as a Shona language [3]. There were many such cases.

4 Conclusion

To conclude, our goal in this competition was to implement an efficient method for language identification. We compared a few pipelines to find one that was efficient enough for this task. Even if naive and deterministic approaches achieved some interesting performances, the solution we kept was the fine-tuning of the RoBERTa model, which is an advanced language understanding model. This method led us to a final accuracy of 90.2%.

5 References

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

6 Appendix

Language	Croatian	Bosnian
F1 Score	0.463	0.438
RoBERTa trained on lang.	Yes	Yes
Sentence Count	500	500
Word Count	6846	6786
Tokens (Million)	3297	14
Size (GiB)	20.5	0.1

Table 2: Language Information for Croatian and Bosnian

Language	Shona	Kinyarwanda
F1 Score	0.455	0.460
RoBERTa trained on lang.	No	No
Sentence Count	500	500
Word Count	12489	17024

Table 3: Language Information for Shona and Kinyarwanda