Understanding Native Language Identification for Brazilian Indigenous Languages

Paulo Cavalin, Pedro H. Domingues, Julio Nogima, Claudio Pinhanez
IBM Research
Rio de Janeiro, RJ, Brazil
pcavalin@br.ibm.com

Abstract

We investigate native language identification (LangID) for Brazilian Indigenous Languages (BILs), using the Bible as training data. Our research extends from previous work, by presenting two analyses on the generalization of Bible-based LangID in non-biblical data. First, with newly collected non-biblical datasets, we show that such a LangID can still provide quite reasonable accuracy in languages for which there are more established writing standards, such as Guarani Mbya and Kaigang, but there can be a quite drastic drop in accuracy depending on the language. Then, we applied the LangID on a large set of texts, about 13M sentences from the Portuguese Wikipedia, towards understanding the difficulty factors may come out of such task in practice. The main outcome is that the lack of handling other American indigenous languages can affect considerably the precision for BILs, suggesting the need of a joint effort with related languages from the Americas.

1 Introduction

Brazil is home to about 270 indigenous languages, referred to as Brazilian Indigenous Languages (BILs) hereafter. All of those language are endangered, spoken by at most 30 thousand people, and are quite understudied. Serious effort should be put onto creating resources and tools to help vitalize the culture of such underrepresented communities. The creation of AI tools, in special language models and applications such as language translators, next-word predictors, spell checkers, can be key for this endeavour, since all could be used as learning-aid tools.

One main issue in building AI tools for understudied languages, which is the case of BILs, is the lack of data. There is almost no data available in ready-to-use formats, such as parallel corpora and labelled datasets, even monolingual data is scarce. Finding data for such languages is very difficult, since documents are stored in varied repositories and there is no indexing in search engines for such languages.

Native language identification (LangID) represent of a crucial approach to help in the task of gathering and augmenting data for BILs and many other indigenous languages. Not only LangID can be helpful to mine data from the web, it can be used as a tool to validate data that is generated synthetically with back-translation or self-training (Feldman and Coto-Solano, 2020; He et al., 2020). Before putting a LangID system into practice, though, it is very important to have a clear understanding of its capabilities, such as the expected accuracy on unseen domains.

Apart from an evaluation of LangID with indigenous languages in isolation (Lima et al., 2021), or the addition of some language in a publicly-available LangID dataset (Brown, 2014), in both cases with only Bible data, the potential of LangID for BILs in non-biblical, open-world data is quite understudied. That, again, owns to the lack of data, since the only source of data available to build a LangID for BILs is the Bible. And that is quite limiting in terms of understanding of the usefulness of a LangID for BILs in varied domains.

In this work we focus on expanding the horizons on a LangID for BILs, and present a deeper investigation on the quality and practical issues of Bible data for LangID on such languages. For that, we collected and appended 1.5M sentences from 51 BILs to the existing WiLi2018 LangID dataset, with 235 languages (Thoma, 2018), to train a machine learning-based LangID approach, and test it on different scenarios. We focused on answering two main research questions: **RQ1)** what is the level of accuracy achieved by this Bible-based LangID on a sample of out-of-domain, non-biblical data sets? and **RQ2)** if we apply this LangID on a large set of texts in the wild, what are the main difficulty factors?

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 12–18

For LangID, we implemented an approach considering bag-of-words on tokens computed with SentencePiece, and Support Vector Machines (SVMs) as the classifier. Results show an accuracy of 73.3% and 95.1% for BILs and non-BIL languages, respectively. To answer RQ1, we built a dataset with almost three thousand sentences, comprising seven monolingual dataset in six different BILs. The results indicate that our LangID classifier generalizes quite well to most languages, reaching up to about 90% accuracy. But we see also that there might be a drop to about 37% with Apurinã, for which writing standards are not quite well established. To answer RQ2, we applied our LangID on about 13.5M sentences extracted from the Portuguese Wikipedia. As much as 3,821 sentences were pointed out with a BIL as the most probable class, but most of them with very low probability scores, below 0.1. A further manual inspection showed a precision of 7% only, uncovering a fundamental issue that needs to be overcome in the future to improve the prediction of LangID in in-the-wild data, which the need to handle other american languages to reduce false positive hits.

2 A LangID Dataset for BILs

We built a dataset containing 51 BILs, with data extracted from the Bible. Although this covers only a sample of the total of about 270 existing BILs, according to the last comprehensive assessment of linguistic diversity in Brazil (IBGE, 2010)[1], this set represents about a third of the estimate of 90 languages that have established standards of writing (Diniz, 2007). Additionally, we expect that the results expand to languages that belong to same families and branches in which the BILs are organized (Storto, 2019; Rodrigues, 1986).

Besides the languages spoken solely in Brazil, we include languages that are mostly spoken outside of Brazil but with some speakers in the country, such as the version of Guarani spoken in Paraguay, and languages that are relatives to some BILs, such as the eastern and western versions of Guarani spoken in Bolivia.

For data splitting, the test set was composed of all sentences from the Matthews New Testament book, for which we tokenized all chapters with the NLTK sentence tokenizer. Then we perform the same procedure to create the training set, with all remaining books from the New Testament, and books from the Old Testament, when available. As a result, the total number of training samples is 1,330,457 samples, and 199,128 test examples. The average number of samples per language is of 26,087 for the training set, and 3,904 in the test set.

Additional details are presented in Appendix A.

3 The LangID Classifier for BILs

We developed a LangID system using a linear SVM classifier with Bag-of-words (BOW) features, relying on the SentencePiece tokenizer[2], with 100K tokens. Note that we have evaluated different configurations for vocabulary size and other classifiers, but found that the linear SVM with 100K tokens presented the highest mean accuracy in the two test sets available, i.e. one for the BILs and another from WiLi-2018. Detailed results are provided in Appendix B.

As the training set, we considered the concatenation of our Bible-based dataset for BILs and the WiLi-2018 dataset, which contains 235K samples, evenly distributed over the 235 languages in the dataset. The accuracy on those sets are, respectively, 73.3% and 95.1%. Notice that our LangID approach excels pretty well on the WiLi2018 test set, almost 5 percentage points better than the 89.42% accuracy reported in Thoma (2018). But the accuracy presented on the BILs test set is 22 percentage points lower, which we believe is related to the inherited difficulties of doing LangID for such languages.

4 Accuracy on non-biblical datasets

In order to validate the quality of the LangID system proposed in this work, and to answer RQ1, we performed an evaluation on non-biblical data. We built seven new datasets, comprising six different BILs, to measure the accuracy of LangID on domains that are quite unrelated to the training set. Furthermore, this analysis also helps understand if the orthography of the training samples match what is expected in unseen domains.

This data has been collected either from PDF files, available in repositories in the web, or from annotation efforts such as the Universal Dependencies Parsing (UDP). For the former, the task basically consisted of cleaning up any annotation and generating a file only with the sentences in the corresponding BIL. But for the PDFs, we had to

[1]There is some discussion about the accuracy of those numbers, see Franchetto (2020); Storto (2019).

[2]https://github.com/google/sentencepiece

Table 1: Results on out-of-domain, non-biblical datasets.

Language	Source	#sent	Acc(%)
gun	Books	1,400	88.2
gun	Tales	1,022	88.8
myu	UDP	157	91.7
kgp	Books	146	81.8
urb	UDP	83	72.3
apu	UDP	59	37.3
xav	UDP	20	75.0
mean		412.4	76.4

either copy and paste the contents in the PDF to text files, or even retype the content given the lack of standard in encoding for such languages and the lack of standard for PDFs files. In both cases, tough, manual inspection of the conversion results proved necessary to handle special characters such as some combinations of letters and accents that are not very usual in non-indigenous languages. Once the blocks of texts have been inspected and converted to a text file, we then applied a sentence tokenizer to split paragraphs into individual sentences. Finally, we filtered all sentences with less than three tokens, to avoid dealing with such very short sentences.

In Table 1, we present further details on each dataset, such as the language, the source, and the resulting number of sentences. Note that some datasets consist of groupings of different sources, such Books in gun and kgp, which are composed of sentences extracted from multiple school books in PDF formats, such as Dooley (1985), and Tales in gun, which comprises several PDF files containing short indigenous tales (Dooley, 1988a,b).

The accuracy rates, also presented in Table 1, show that the results vary greatly from language to language. For the datasets in Guarani Mbya (gun), our LangID approach was able to achieve an accuracy of 88.6% on average, which is quite higher than the 73.3% achieved on held-out bible data. And the approach was able to achieve accuracy as high as 91.7% on myu. For urb and xav, we observe accuracy that are comparable to what we found on bible data, i.e. 72.3% and 75%. And for apu, there is a significant drop to 37.3% . We suspect that such drop in accuracy is due to differences in orthography from what is in the Bible and what is in these test sets, but further inspection with linguists or native speakers is necessary to check this assumption. It is worth mentioning that gun and kgp have quite established written forms, and for those languages we do not see such a drastic drop in classification quality.

An additional evaluation was then performed to understand if the classification of BILs is affected by non-indigenous languages. For that, we checked which languages where misclassified the most with gun in the respective datasets for this language. In the Books dataset, from the 162 misclassifications, 63 (39%) were associated to languages belonging to the Tupi-guarani family, which is the same family of gun. From those 63 samples, 38 were detected as kgk, and 25 as gug. Similarly, in the Tales dataset, from the 118 errors, 78 (66%) were from Tupi-guarani family languages: 48 in kgk and 30 in gug. Thus, considering the high similarity of such languages from the Tupi family, it is likely that the results with gun can be improved with further development of the LangID classifier, in order to handle better the classification among these similar languages.

5 Bringing LangID closer to practice

Aiming at answering RQ2, we expanded the evaluation of the previous chapter to a large, unsupervised set. Our goal was to understand the main challenges in a scenario that is closer to practical application, which is applying our LangID on in-the-wild data, to mine for sentences written in one of the 51 BILs. For that, we considered about 13.5M sentences extracted from the Portuguese Wikipedia. Although that data presents limitations, since most pages are supposedly written in Portuguese and a totally open set such as Common Crawl represents better the real world, that is also an advantage since we can discard all sentences detected as Portuguese and manually inspect only the remaining smaller set. And the associated Wikipedia pages can be used as ground-truth for the results of the classifier.

This evaluation considered an exact total of 13,573,101 sentences, from which our proposed LangID was able to identify 3,821 sentences as one of the 51 BILs considered in this work. That corresponds to 0.03% of total sentences in the dataset. We observed, though, the very low prediction score for such detected sentences, with a mean of around 0.03, and decided to discarded all sentences with a prediction score below 0.1, resulting in a set of only 129 sentences. That is a quite small set, but this number was somewhat expected given the data

in Portuguese. One the other hand, that allowed us to conduct a manual inspection on the results.

We manually inspected all of the detected 129 sentences, marking all sentences that 'looked like being correctly classified'. That is, we inspected the 129 sentences and marked all sentences that were written in a latin scripts, but with words that did not belong to any of the non-indigenous languages known by the authors, such as Portuguese, English, Spanish, German, and French, to name a few. With that approach, we found a total of 50 sentences that could likely be from a BIL. Then, for each of those 50 sentences, we searched for the original Wikipedia page of the sentence, by using its text as a query on Google, and inspected the resulting pages. The results were, on one hand, disappointing, since very few sentences were correctly classified. But on the other hand, they were quite useful in understanding some particular difficulties of this task, and how to approach this problem better in the future. Details are provided next.

The results were disappointing in the sense that very few sentences were correctly classified, i.e. very low precision. From the 50 sentences that we suspected were correct, only 9 sentences were extracted from a Wikipedia page that related to the actual predicted language. That gives a precision of only 7%. Besides, we uncovered that those nine sentences consisted of samples of the Lord's prayer, which is a content that is very close to what is in the training for such languages, so these results do not help in clarifying the potential of LangID in non-religious content.

Nevertheless, some interesting findings of this study consist of a better understanding of the main difficulties that we may face when applying LangID to mine data for BILs. One clear drawback of our proposed approach, is the limited handling of similar low-resource languages, such as indigenous languages from other South and North American countries besides Brazil. Most of the classification mistakes involved Wikipedia entries of languages spoken in countries such as Peru, Colombia, Mexico, and the United States. Some other few mistakes involved languages from more distant locations, such as Indonesia and the African continent. These results show that, in order to perform accurate LangID for BILs, it is important to include as much languages as possible in the training set to have a more precise classification, or to implement some mechanism to deal with out-of-scope detection.

This evaluation also showed that searching for webpages using sentences in a target language as a query for a search engine can be helpful to find for additional data, such as PDFs with additional content such as the one found for the Amarakaeri language[3]. Even though Amarakaeri is not included in the set of BILs, with more accuracy in LangID, we could search for PDF documents is such languages with greater precision. Furthermore, we found that misclassifications can be useful to find content in additional related languages, such as the language Cocama[4], which is spoken in Brazil and belongs to the Tupi family, but was not included our set of BILs for LangID.

6 Conclusions and future work

In this paper we present an evaluation of LangID for Brazilian Indigenous Languages (BILs), using the Bible as the only source for training data. We demonstrate that on non-biblical, labeled datasets, the approach is able to achieve even accuracy in languages with more estabilised written forms, such as Guarani Mbya and Kaigang, but the performance may drop considerable for less studied languages. By applying the LangID classifier in an almost in-the-wild dataset, we saw that the precision is quite affected by related American indigenous languages that are not handled by our LangID approach, so a joint effort must be made to handle as much american languages as possible, together, to improve the quality of the LangID in practice.

As future work, we believe that expanding the LangID training set, to consider as much languages as possible, is mandatory. Furthermore, an inspection of the orthography of some languages should also be done, by partnering with linguists and/or native speakers. And we think that we could further develop the study in in-the-wild data, either by searching for BIL data on a more comprehensive dataset, such as Common Crawl[5] and BrWac[6], and by including the search for PDF documents, which is the most commonly used format containing data for such languages.

[3]https://www.ohchr.org/sites/default/files/UDHR/ Documents/UDHR_Translations/amr.pdf

[4]https://pt.wikipedia.org/wiki/Língua_cocama

[5]https://commoncrawl.org/

[6]https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC

Limitations

One limitation of this work is the lack of a more comprehensive study of LangID methods, which could impact slightly the results. Another limitation is the number of non-BIL languages, which can be increased to more than 1,000 languages with the datasets proposed in (Brown, 2014). Furthermore, the use of Wikipedia data limits the search of samples, since all pages are supposedly written in Portuguese. So, relying on a broader set can bring a more realistic estimate on the in-the-wild search for data. In addition, a major limitation of this work is the lack of inspection of the results with native speakers. We are already engaging with one mbya guarani community, but it is quite difficult to extend such engagement to other communities.

References

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar. Association for Computational Linguistics.

Kollontai Cossich Diniz. 2007. Notas sobre tipografias para línguas indígenas do brasil. *InfoDesign: Revista Brasileira de Design da Informação*, 4(1).

Robert Dooley. 1985. Nhanhemboe aguã nhandeayvu py [1-5].

Robert Dooley. 1988a. Arquivo de textos indígenas – guaraní (dialeto mbyá) [1].

Robert Dooley. 1988b. Arquivo de textos indígenas – guaraní (dialeto mbyá) [2].

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bruna Franchetto. 2020. Língua (s): cosmopolíticas, micropolíticas, macropolíticas. *Campos-Revista de Antropologia*, 21(1):21–36.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

IBGE. 2010. Censo demográfico 2010. Accessed = 2022-12-30.

Tiago Lima, André Nascimento, Pericles Miranda, and Rafael Mello. 2021. Analysis of a brazilian indigenous corpus using machine learning methods. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 118–129, Porto Alegre, RS, Brasil. SBC.

Aryon Dall'Igna Rodrigues. 1986. *Línguas brasileiras: para o conhecimento das línguas indígenas*. Edições Loyola.

Luciana Raccanello Storto. 2019. *Línguas indígenas: tradição, universais e diversidade*. Mercado de Letras.

Martin Thoma. 2018. The wili benchmark dataset for written language identification. *CoRR*, abs/1801.07779.

A Details on languages and datasets

In Table 2 we present the full list of Brazilian Indigenous Languages (BILs) considered for this work, with the corresponding ISO 639 codes, their geo-linguistc classification in terms of branches and families, the estimated number of speakers, and the number of samples for training and test sets.

B Detailed results on classifier evaluation

In Table 3 we present the detailed accuracy on each methods and dataset evaluated in this work. In terms of classifier, we evaluated two approaches: Logistic Regression and Support Vectors Machines (SVMs). For feature extraction, we evaluate the use of bag of words (BoW) and corpus-based vocabulary extraction with SentencePiece (SP), with varied number of tokens: 10K, 50K, 100K, and 250K.

Table 2: Details on the indigenous languages and datasets used in the study.

Languages					# Aligned Sentences		
Name	**Acron**	**Branch**	**Family**	**Speakers**	**Train**	**Test**	**Total**
Apalaí	apy	No Branch	Karib	252	27,763	4,401	32,164
Apinayé	apn	Macro Jê	Jê	1,386	28,069	4,354	32,423
Apurinã	apu	No Branch	Aruak	824	28,629	4,403	33,032
Ashaninka	cni	No Branch	Aruak	302	19,564	2,943	22,507
Bakairí	bkq	No Branch	Karib	173	27,314	4,206	31,520
Boróro	bor	Macro Jê	Boróro	1,035	32,392	5,206	37,598
Desána	des	No Branch	Tukano	95	26,115	4,019	30,134
Guajajára	gub	Tupi	Tupi-Guarani	8,269	33,188	4,818	38,006
Guarani Eastern Bolivia	gui	Tupi	Tupi-Guarani	NA	22,681	3,342	26,023
Guarani Kaiowá	kgk	Tupi	Tupi-Guarani	24,368	31,523	4,711	36,234
Guarani Mbya	gun	Tupi	Tupi-Guarani	3,248	18,245	2,857	21,102
Guarani Paraguay	gug	Tupi	Tupi-Guarani	2,464	16,891	2,841	19,732
Guarani Western Bolivia	gnw	Tupi	Tupi-Guarani	NA	22,281	3,264	25,545
Hixkaryána	hix	No Branch	Karib	52	37,893	5,797	43,690
Jamamadí-Kanamanti	jaa	No Branch	Arawá	217	21,169	3,121	24,290
Ka'apor	urb	Tupi	Tupi-Guarani	1,241	44,969	6,678	51,647
Kadiwéu	kcb	No Branch	Guaikurú	649	19,773	3,020	22,793
Kaiabi	kyz	Tupi	Tupi-Guarani	673	36,118	5,145	41,263
Kaingáng	kgp	Macro Jê	Jê	19,905	27,778	4,070	31,848
Kanela	ram	Macro Jê	Jê	488	18,342	731	19,073
Karajá	kpj	Macro Jê	Karajá	3,119	22,721	3,646	26,367
Kaxinawá	cbs	No Branch	Pano	3,588	14,590	2,099	16,689
Kayapó	txu	Macro Jê	Jê	5,520	34,066	5,631	39,697
Kubeo	cub	No Branch	Tukano	171	25,216	3,650	28,866
Kulina Madijá	cul	No Branch	Arawá	3,043	27,744	4,318	32,062
Makúna	myy	No Branch	Tukano	6	27,568	4,000	31,568
Makuxí	mbc	No Branch	Karib	4,675	26,942	4,199	31,141
Matsés	mcf	No Branch	Pano	1,144	23,754	3,772	27,526
Mawé	mav	Tupi	Mawé	8,103	27,034	3,035	30,069
Maxakali	mbl	Macro Jê	Maxakali	1,024	20,663	3,045	23,708
Mundurukú	myu	Tupi	Mundurukú	3,563	32,880	5,146	38,026
Nadëb	mbj	No Branch	Makú	326	24,653	3,821	28,474
Nambikwára	nab	No Branch	Nambikwára	951	29,089	4,377	33,466
Nheengatu	yrl	Tupi	Tupi-Guarani	3,771	15,236	2,321	17,557
Palikúr	plu	No Branch	Aruak	925	28,322	4,228	32,550
Paresí	pab	No Branch	Aruak	122	20,759	3,043	23,802
Paumarí	pad	No Branch	Arawá	166	30,389	4,550	34,939
Piratapúya	pir	No Branch	Tukano	81	25,721	4,030	29,751
Rikbaktsa	rkb	Macro Jê	Rikbaktsa	10	35,777	4,841	40,618
Sanumá	xsu	No Branch	Yanomámi	1,788	25,118	3,749	28,867
Siriáno	sri	No Branch	Tukano	2	24,247	3,626	27,873
Tenharim	pah	Tupi	Tupi-Guarani	32	30,277	5,145	35,422
Teréna	ter	No Branch	Aruak	6,314	20,713	3,170	23,883
Tikúna	tca	No Branch	No Family	30,057	20,101	3,218	23,319
Tukáno	tuo	No Branch	Tukano	4,412	26,826	3,952	30,778
Tuyúca	tue	No Branch	Tukano	263	23,973	3,572	27,545
Wanana	gvc	No Branch	Tukano	236	25,487	3,983	29,470
Wapishana	wap	No Branch	Aruak	3,154	20,561	2,930	23,491
Xavante	xav	Macro Jê	Jê	11,733	24,714	3,737	28,451
Yamináwa	yaa	No Branch	Pano	222	24,808	3,680	28,488
Yanomámi	guu	No Branch	Yanomámi	12,301	29,811	4,687	34,498
TOTAL				**176,463**	**1,330,457**	**199,128**	**1,529,585**

Table 3: Detailed results considering different classifiers and feature extraction methods.

		Classifier									
		Logistic Regression					Support Vector Machine				
		BoW	**SP10K**	**SP50K**	**SP100K**	**SP250K**	**BoW**	**SP10K**	**SP50K**	**SP100K**	**SP250K**
Test set	WiLi2018	73.87	93.13	92.30	91.37	89.70	89.45	94.15	94.94	95.07	94.63
	Bibles-BILs	69.59	72.81	72.31	71.96	71.47	72.85	73.18	73.19	73.28	73.14
	mean	71.73	82.97	82.31	81.67	80.59	81.15	83.67	84.07	**84.18**	83.89
	gun Books	71.01	86.67	86.51	85.73	86.06	82.21	89.10	89.25	88.20	89.32
	gun Tales	82.19	83.95	83.66	82.88	83.76	86.20	89.53	90.12	88.85	88.65
	myu UDP	73.97	95.54	91.08	89.81	90.45	87.67	96.18	91.72	91.72	91.08
	kgp Books	73.55	81.25	73.43	70.63	69.23	84.30	89.58	86.01	81.82	81.82
	urb UDP	43.84	60.24	61.45	59.04	62.65	63.01	65.06	74.70	72.29	72.29
	apu UDP	14.29	28.81	33.90	28.81	32.20	25.00	25.42	33.90	37.29	45.76
	xav UDP	47.37	75.00	75.00	60.00	50.00	63.16	75.00	70.00	75.00	18.75
	mean	58.03	73.07	72.15	68.13	67.76	70.22	75.70	76.53	76.45	69.67