A Low-Resourced Peruvian Language Identification Model

Alexandra Espichán Linares[1] and **Arturo Oncevay-Marcos**[2]
[1]Facultad de Ciencias e Ingeniería, [2]Departamento de Ingeniería
Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada
Pontificia Universidad Católica del Perú, Lima, Perú
a.espichan@pucp.pe,arturo.oncevay@pucp.edu.pe

Abstract

Due to the linguistic revitalization in Perú through the last years, there is a growing interest to reinforce the bilingual education in the country and to increase the research focused in its native languages. From the computer science perspective, one of the first steps to support the languages study is the implementation of an automatic language identification tool using machine learning methods. Therefore, this work focuses in two steps: (1) the building of a digital and annotated corpus for 16 Peruvian native languages extracted from documents in web repositories, and (2) the fit of a supervised learning model for the language identification task using features identified from related studies in the state of the art, such as n-grams. The obtained results were promising (97% in average precision), and it is expected to take advantage of the corpus and the model for more complex tasks in the future.

1 Introduction

In Perú, there are 4 million people that are speakers of a native language. They are part of the rich linguistic diversity in the country, with a presence of 47 original languages divided by 19 linguistic families. These peruvian languages are distributed across the highlands and jungle (Amazon) regions, and most of them are very unique, in spite of their geographical or linguistic closeness (Ministerio de Educación, Perú, 2013).

The linguistic diversity calls for equal opportunity across the different native communities, and this could be supported by high-level bilingual education and a deep knowledge about these languages. For that reason, there is a need to support the linguistic research from an informatics point of view, and one of the first required tools is an automatic language detector for written text (in different levels, such as a complete document, a paragraph or even a sentence) (Malmasi et al., 2015).

To develop an automatic language identifier, a basic natural language processing (NLP) task, an annotated textual corpus for the languages is required first. However, not all the languages have large enough digital corpus for any computational task, so they are known as low-resourced languages from a computer science point of view (Forcada, 2006).

In this way, it is a must to build a digital repository of textual corpora for these languages. That will be a previous step to the develop of an automatic language model identification.

In the next section, the Peruvian native languages used in this work are presented. Then, in Section 3 some related works are described. After that, Section 4 presents the corpus building and the details of the dataset obtained for the study. Then, Section 5 contains the implementation of the language identification model. Finally, the results and discussions are included in Section 6, while the conclusions and future work for the study are presented in Section 7.

2 Peruvian native languages

Among the 47 languages spoken by peruvian people, 43 are Amazonian (from the jungle) and 4 are Andean (from the highlands). These languages are considered prevailing languages because they have live speakers. Therefore, there are 19 linguistic families (a set of languages related to each other and with a common origin): 2 Andean (Aru and Quechua) and 17 Amazonian (Ministerio de Educación, Perú, 2013).

Table 1: Basic information of the languages within the scope of the study.

Linguistic Family	Language	ISO-639-3	Speakers
Arawak	Ashaninka	cni	88 703
	Asheninka	cjo	8 774
	Matsigenka	mcb	11 275
	Yine	pib	3 261
Aru	Aymara	aym	443 248
Jíbaro	Awajún	agr	55 366
Pano	Cashinahua	cbs	2 419
	Kakataibo	cbr	1 879
	Matses	mcf	1 724
	Shipibo-konibo	shp	22 517
Quechua	Quechua Wanca	qxw	37 559
	Quechua de Lambayeque	quf	21 496
	Quechua de Yauyos	qux	456 225
	Quechua del Callejon de Huaylas	qwh	451 789
	Quechua del Cusco	quz	566 581
	Quechua del Este de Apurimac	qve	266 336

The 47 original native languages are highly agglomerative, unlike Spanish (Castillan), the main official language in the country. Even though, most of them presents more than 100 morphemes for the word formation process. For instance, *Quechua del Cusco* contains 130 suffixes (Rios, 2016), meanwhile *Shipibo-konibo* uses 114 suffixes plus 31 prefixes (Valenzuela, 2003).

In this work, the language identification task was performed on 16 languages (from 5 families) including 6 dialects of Quechua. The ISO-639-3 codes and the approximate number of speakers of each language are presented in Table 1.

3 Related Work

Given that Peruvian languages can be considered as low-resourced ones, a systematic search for studies focused on automatic language identification for low-resourced languages was carried out. The results are described as follows.

Malmasi et al. (2015) present the first study to distinguish texts between the Persian and Dari languages at the sentence level. As Dari is a low-resourced language, it was developed a 28 thousand sentences corpus for this task (they used 14 thousand for each language). Characters and sentences n-grams were considered as language features. Finally, using a SVM (Support Vector Machine) implementation within a classification ensemble scheme, they discriminate both languages with 96% accuracy.

Botha and Barnard (2012) research the factors that may determine the performance of text-based language identification, with a special focus in the 11 official languages of South Africa, using n-grams as language features. In the study 3 classification methods were tested: SVM, Naive Bayes and n-gram rank ordering on different training and test text sizes. In this way, it was found that the 6-gram Naive Bayes model has the best performance in general, obtaining 99.4% accuracy for large training-test sets and 83% for shorter sets.

Selamat and Akosu (2016) propose a language identification algorithm based on lexical features that works with a minimum amount of training data. For this study, a dataset of 15 languages, mostly low-resourced, extracted from the Universal Declaration of Human Rights was used. The used technique is based on a spelling checker-based method (Pienaar and Snyman, 2011) and the improvement proposed in this research was related to the indexation of the vocabulary words according to its length. In this way, the average precision of the method was 93% and an improvement of 73% in execution time was obtained.

Grothe et al. (2008) compare the performance of three feature extraction approaches for language identification using the Leipzig Corpora Collection (Quasthoff et al., 2006) and randomly selected Wikipedia articles. The considered approaches for features were short words (SW), frequent words (FW) and n-grams (NG). Meanwhile, the employed classification method was Ad-Hoc Ranking. Hence, the best obtained results for each approach were: FW 25% (99.2%), SW 4 (94.1%) and NG with 3-grams (79.2%).

4 Corpus Development

To build the corpus used in this study, digital documents containing Peruvian native languages texts were retrieved from the web, while others one were obtained directly from private repositories or books. In this way, it was possible to collect and annotate documents from 16 different native languages. It may be considered that these documents must be annotated, i.e., the language in which they are written must be known.

Then, as almost all the documents were in PDF format, the text content was extracted and some manual corrections were made if it was necessary. Next, a preprocessing program was developed to clean the punctuation, to lowercase the text and to split the sentences. After that, Spanish and English sentences were discarded using the resources of a language generic spell-checking library[1], remaining only Peruvian native languages sentences.

Table 2 contains the total amount of files, plus the number of sentences/phrases and tokens split for each Peruvian language used in this study. This preprocessed collection is partially available in a project site, including details of the sources of each language text[2].

Moreover, Figures 1 and 2 presents some statistics regarding the distribution of the total of characters per word and per sentence, respectively, in each processed language.

The first boxplot in Figure 1 supports the rich morphology feature of the Peruvian native languages, as a high number of characters is observed for the word length value in most of them. Also, it can be noticed that most of the words are formed by 5 to 10 characters. Nevertheless, there are very large words from *Matses* (cbf), such as *cuishonquededcuishonquededtsëcquiec* or *tantiabentantiabentsëccondaidquio*, with 35 and 33 characters, respectively. Although, most words from *Matses* presents a word-length value between 5 to 10 characters.

On the other hand, on average, the language with longer words is *Matsigenka* (mcb), while the language with shorter words is *Kakataibo* (cbr). Moreover, the distribution among languages of the Quechua family is pretty similar.

Table 2: Retrieved corpus information: $|D|$ = document collection size; $|S|$ = sentences/phrases collection size; $|\mathcal{V}|$ = word vocabulary size, without considering punctuation; $|\mathcal{C}|$ = character vocabulary size; T = number of tokens.

Lang.	$\|D\|$	$\|S\|$	$\|\mathcal{V}\|$	$\|\mathcal{C}\|$	$\mathcal{T}$
cni	4	7 516	10 125	35	25 119
cjo	1	555	1 308	35	2 691
mcb	1	2 502	4 276	33	10 092
pib	1	106	299	21	465
aym	5	16 431	16 216	39	53 115
agr	5	14 258	18 631	36	47 127
cbs	1	33	129	26	161
cbr	195	6 970	6 584	38	37 117
mcf	2	16 356	14 722	36	64 779
shp	4	15 866	24 597	35	203 988
qxw	6	1 259	2 782	38	6 640
quf	8	442	1 289	36	2 027
qux	2	20 496	25 105	35	85 124
qwh	3	665	2 029	36	3 448
quz	2	3 496	5 866	37	13 592
qve	9	635	1 744	31	2 957

On Figure 2, it can be noticed that the longest collected sentences are from *Shipibo-konibo* (shp) while the shortest are from *Aymara* (aym). The reason for the first case is the origin of the *Shipibo-konibo* corpus: a parallel one built for a SMT experiment, which legal and educational text domain sources contains longer sentences than the ones found in dictionary or lexicon samples (Galarreta et al., 2017).

5 Language Identification Model

As it is proposed to perform language identification at the sentence level, the aim was to learn a classifier or classification function (γ) that maps the sentences from the corpus (S) to a target language class (L):

$$\gamma : S \rightarrow L \tag{1}$$

In order to identify which γ classifier is most suited in the task, each sentence $s \in S$ will be represented in a feature vector space model: $s_i = \{w_{1,i}, w_{2,i}, ..., w_{t,i}\}$, where t indicates the number of dimensions or terms to be extracted.

Character-level n-grams was one of the most

[1] libenchant: https://github.com/AbiWord/enchant

[2] chana.inf.pucp.edu.pe/resources/multi-lang-corpus

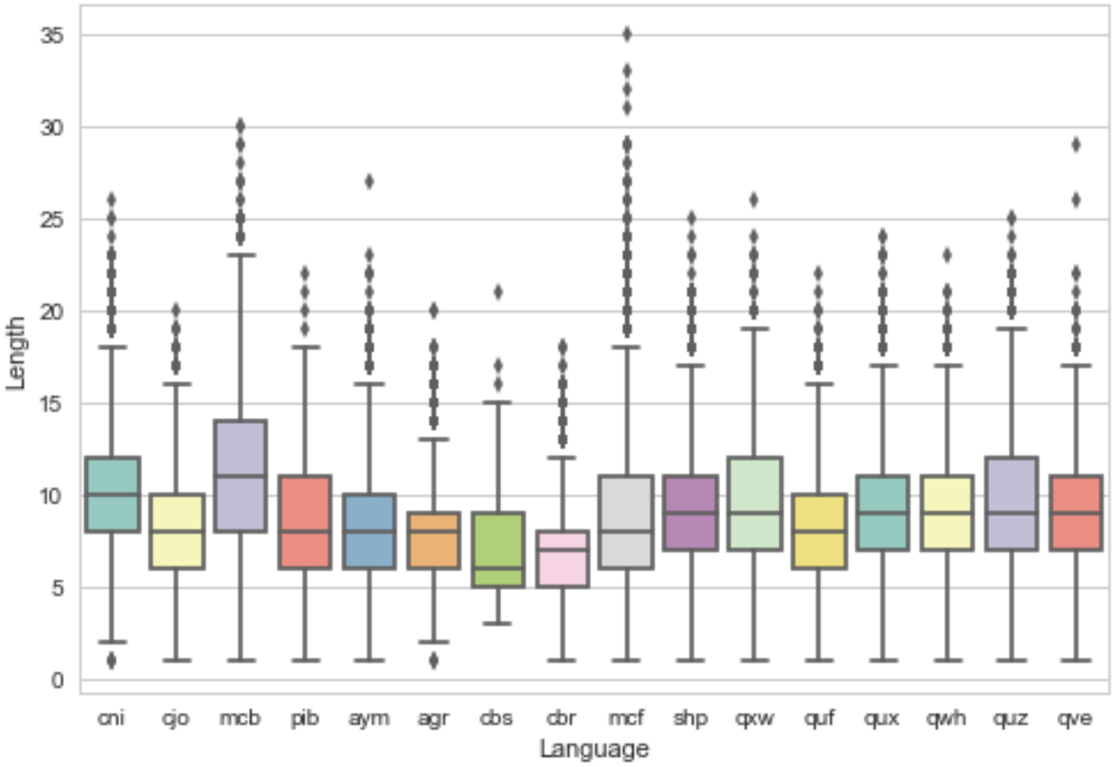


Figure 1: Boxplots representing the distribution of the word length in number of characters per each language.

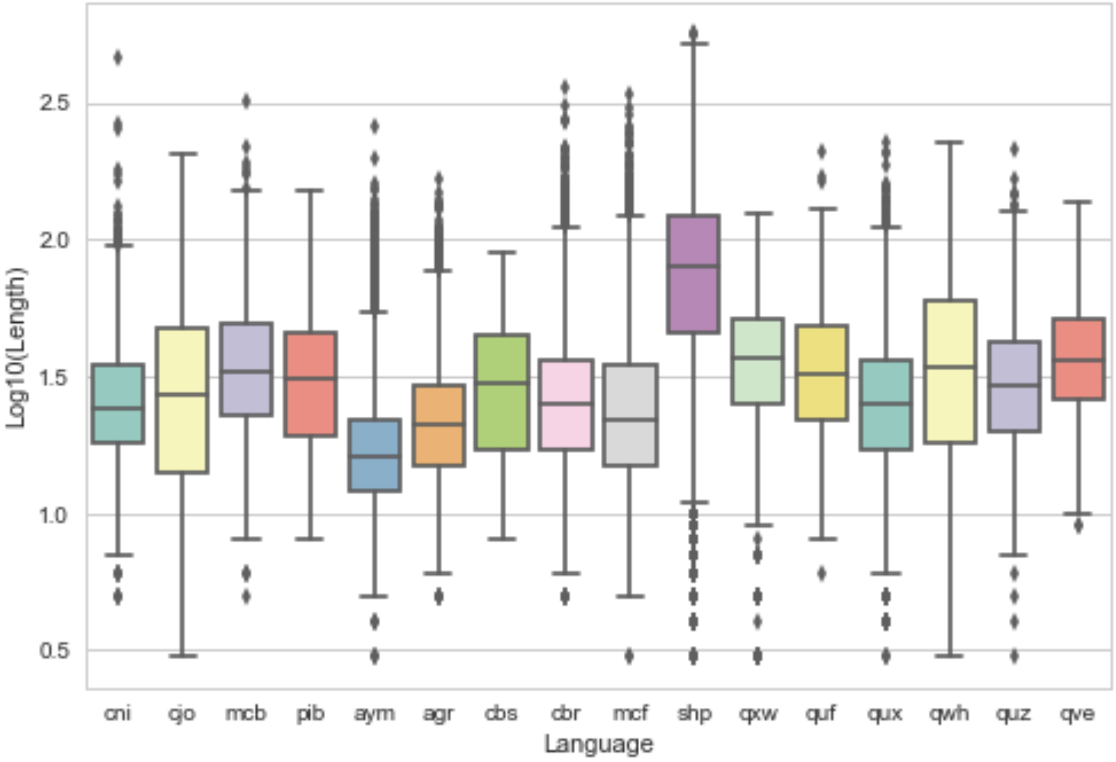


Figure 2: Boxplots representing the distribution of the sentence length in number of characters per each language. The vertical axe (lenght) is in a log10 scale.

used language features in the revised works for this task (Grothe et al., 2008; Botha and Barnard, 2012; Malmasi et al., 2015). Hence, the dimensionality of each vector in the space model will be equal to the number of distinct subsequences of n characters in a given sequence from the corpus S (Cavnar and Trenkle, 1994).

In this experiment, bigrams and trigrams were used to built the vector space model, and a term frequency - inverse document frequency (TF-IDF) matrix from the aforementioned n-grams scheme was calculated (Prager, 1999).

After that, the matrix was split in train and test sub-datasets (70%-30%) and some classification methods identified in the related works (Grothe et al., 2008) were fit using a 5-fold cross-validation schema on the training sub-dataset. The obtained results are shown in Table 3.

Table 3: Results of the 5-fold cross-validation classification on the train sub-dataset

Method	**Accuracy (%)**
SVM (linear kernel)	96.22
Multinomial Naive Bayes	92.76
SGD Classifier	94.52
Perceptron	95.05
Passive Aggressive Classifier	95.89

As the SVM classifier with a linear kernel got the best accuracy result, this method was used to fit the main model on the entire train sub-dataset. Next, this model was validated on the test sub-dataset. A report of the performance of this model at classifying each language was made and is shown in Table 4 (where Support indicates the number of samples that were classified). Furthermore, the confusion matrix of this model is presented in Figure 3.

6 Results and Discussions

In this study, a straightforward experiment was performed for the automatic identification of some Peruvian languages, showing that they can be distinguishable with 96% accuracy. This is a new result for languages that have not previously been worked with.

The acceptable overall result was obtained although there was a great disadvantages to face: the unbalanced corpus, because it was not possible to extract many more sentences from some languages than from others, and even some languages were left with too few data. For instance, for *Yine* (pib) it was only collected 106 sentences, from which at most 39 ones were to the test part. For that language, a precision and recall of 100% and 85% respectively were obtained. This may indicate an acceptable low-resourced language identification model, but to avoid the possibility of overfitting there must be additional tests when more textual documents can be retrieved.

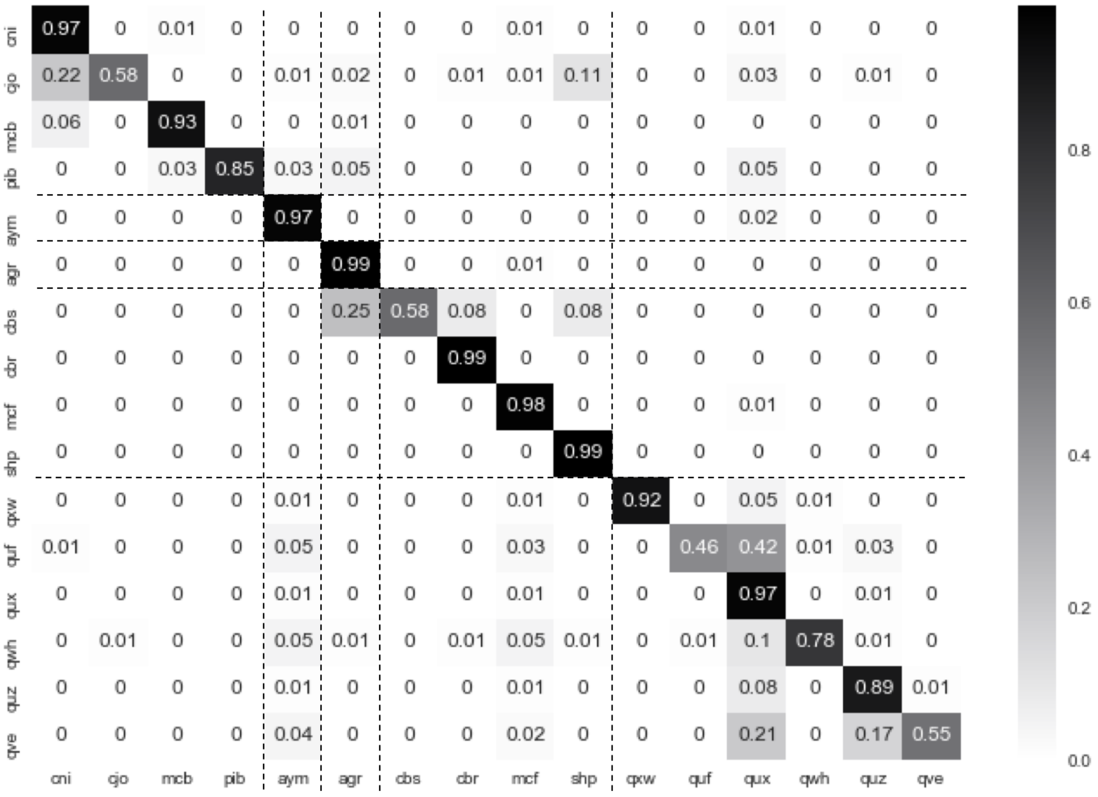
Table 4: Main classification results for each language (SVM with a linear kernel)

Lang.	**Precision**	**Recall**	**Support**
cni	0.94	0.97	2 225
cjo	0.83	0.58	158
mcb	0.96	0.93	753
pib	1.00	0.85	39
aym	0.97	0.97	4 894
agr	0.99	0.99	4 340
cbs	1.00	0.58	12
cbr	0.99	0.99	2 157
mcf	0.97	0.98	4 984
shp	0.99	0.99	4 795
qxw	0.99	0.92	391
quf	0.93	0.46	142
qux	0.94	0.97	5 991
qwh	0.92	0.78	198
quz	0.91	0.89	1 024
qve	0.86	0.55	173
avg/total	**0.97**	**0.97**	**32 276**

On the other hand, as seen in Figure 3, for closely-related languages like *Ashaninka* (cni) and *Asheninka* (cjo), there was a considerable confusion in the model since 22% of the *Asheninka* test sentences were misclassified as *Ashaninka* and only 58% of them were correctly identified.

Likewise, although the Quechua family obtained an acceptable overall precision, a not so good recall is shown for those with less data. As seen in Figure 3, for *Quechua de Lambayeque* (quf), which is the variety of Quechua with the least amount of extracted sentences, only 46% of the test sentences of this variety was properly classified, and the model misclassified 42% of them as *Quechua de Yauyos* (qux). Also, there is confusion at discriminating *Quechua del Este de Apurímac* (qve) since 21% of the sentences of this variety was misidentified as *Quechua de Yauyos* (qux) and 17% as *Quechua del Cusco* (quz).

Both scenarios may indicate the need to go



	cni	cjo	mcb	pib	aym	agr	cbs	cbr	mcf	shp	qxw	quf	qux	qwh	quz	qve
cni	0.97	0	0.01	0	0	0	0	0	0.01	0	0	0	0.01	0	0	0
cjo	0.22	0.58	0	0	0.01	0.02	0	0.01	0.01	0.11	0	0	0.03	0	0.01	0
mcb	0.06	0	0.93	0	0	0.01	0	0	0	0	0	0	0	0	0	0
pib	0	0	0.03	0.85	0.03	0.05	0	0	0	0	0	0	0.05	0	0	0
aym	0	0	0	0	0.97	0	0	0	0	0	0	0	0.02	0	0	0
agr	0	0	0	0	0	0.99	0	0	0.01	0	0	0	0	0	0	0
cbs	0	0	0	0	0	0.25	0.58	0.08	0	0.08	0	0	0	0	0	0
cbr	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0	0	0
mcf	0	0	0	0	0	0	0	0	0.98	0	0	0	0.01	0	0	0
shp	0	0	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0
qxw	0	0	0	0	0.01	0	0	0	0.01	0	0.92	0	0.05	0.01	0	0
quf	0.01	0	0	0	0.05	0	0	0	0.03	0	0	0.46	0.42	0.01	0.03	0
qux	0	0	0	0	0.01	0	0	0	0.01	0	0	0	0.97	0	0.01	0
qwh	0	0.01	0	0	0.05	0.01	0	0.01	0.05	0.01	0	0.01	0.1	0.78	0.01	0
quz	0	0	0	0	0.01	0	0	0	0.01	0	0	0	0.08	0	0.89	0.01
qve	0	0	0	0	0.04	0	0	0	0.02	0	0	0	0.21	0	0.17	0.55

Figure 3: Confusion matrix obtained by the main language identification model. The dashed lines separate the different linguistic families.

deeper in the representation features used for languages within the same linguistic family, and to consider a hierarchical classifying scheme.

Additionally, *Cashinahua* (cbs) was confused as *Awajun* (agr) 25% of the time. This is an interesting result since both languages are from different families: Pano and Jibaru, respectively. However, as *Cashinahua* was the language with the least amount of collected sentences (only 33), it was expected that its results were not as precise as the obtained for the other ones.

7 Conclusions and Future Works

For this study, a corpus for 16 Peruvian native languages was built through web and private repositories. Also, it was performed a straightforward classification experiment with it, using n-grams as features in a tf-idf vector model space. The obtained results (97% in overall precision) were in the expected range regarding the state of the art of language identification in a low-resource scenario.

The fit model may be exploited for other tasks, such as the automatic increasing of the corpus through web and document search (Martins and Silva, 2005). As there are 68 Peruvian native languages preserved, it is essential to expand the corpus to cover most of them. The Bible will be targeted first, as it is translated in some of the left unworked languages, and is a very important resource in NLP for minority cases (Christodouloupoulos and Steedman, 2015).

Also, as the corpus may be growing, other recent methods could be tested on it, such as the bidirectional recurrent neural network proposed by Kocmi and Bojar (2017) or other similar deep architectures (Bjerva, 2016; Mathur et al., 2017). Although in our scenario, this kind of algorithms may face the low-resourced and unbalanced corpus, so there must be an adaptive and tuning steps. However, those methods could help to decrease the window approach of the classification to a phrase or word-level.

Moreover, regarding the confusion presented in languages within the same family, there must be specific considerations in the following experiments with the hierarchy nature in the peruvian linguistic context (Koller and Sahami, 1997; McCallum et al., 1998; Jaech et al., 2016).

Finally, it is desired to develop and integrate a way to discriminate languages that are not part of the scheme, in order to not misclassify out of model languages to a Peruvian one.

Acknowledgements

The authors are thankful to J. Rubén Ruiz, bilingual education professor at NOPOKI, for providing access to some private books written in native languages (Universidad Católica Sedes Sapientiae, 2015; Díaz, 2012). Likewise, it is appreciated the collaboration of Dr. Roberto Zariquiey, linguistic professor at PUCP, for allowing the use of his own corpus for the Panoan family (Zariquiey Biondi, 2011).

Furthermore, it is acknowledged the support of the "Concejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC Perú) under the contract 225-2015-FONDECYT.

References

Johannes Bjerva. 2016. Byte-based language identification with deep convolutional networks. *arXiv preprint arXiv:1609.09004* .

Gerrit Reinier Botha and Etienne Barnard. 2012. Factors that affect the accuracy of text-based language identification. *Computer Speech & Language* 26(5):307–320.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*. pages 161–169.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation* 49(2):375–395.

Darinka Pacaya Díaz, editor. 2012. *Relatos de Nopoki*. Universidad Católica Sedes Sapientiae.

Mikel Forcada. 2006. Open source machine translation: an opportunity for minor languages. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC*. Citeseer, volume 6, pages 1–6.

Ana-Paula Galarreta, Andres Melgar, and Arturo Oncevay-Marcos. 2017. Corpus creation and initial SMT experiments between spanish and shipibo-konibo. In *RANLP*. ACL Anthology. In-press.

Lena Grothe, Ernesto William De Luca, and Andreas Nürnberger. 2008. A comparative study on language identification methods. In *LREC*.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A Smith. 2016. Hierarchical character-word models for language identification. *arXiv preprint arXiv:1608.03030* .

Tom Kocmi and Ondřej Bojar. 2017. LanideNN: Multilingual language identification on character window. *arXiv preprint arXiv:1701.03338* .

Daphne Koller and Mehran Sahami. 1997. Hierarchically classifying documents using very few words. Technical report, Stanford InfoLab.

Shervin Malmasi, Mark Dras, et al. 2015. Automatic language identification for persian and dari texts. In *Proceedings of PACLING*. pages 59–64.

Bruno Martins and Mário J. Silva. 2005. Language identification in web pages. In *Proceedings of the 2005 ACM symposium on Applied computing*. ACM, pages 764–768.

Priyank Mathur, Arkajyoti Misra, and Emrah Budur. 2017. LIDE: Language identification from text documents. *arXiv preprint arXiv:1701.03682* .

Andrew McCallum, Ronald Rosenfeld, Tom M. Mitchell, and Andrew Y. Ng. 1998. Improving text classification by shrinkage in a hierarchy of classes. In *ICML*. volume 98, pages 359–367.

Ministerio de Educación, Perú. 2013. *Documento nacional de lenguas originarias del Perú*. URI: http://repositorio.minedu.gob.pe/handle/123456789/3549.

Wikus Pienaar and DP Snyman. 2011. Spelling checker-based language identification for the eleven official south african languages. In *Proceedings of the 21st Annual Symposium of Pattern Recognition of SA, Stellenbosch, South Africa*. pages 213–216.

John M. Prager. 1999. Linguini: Language identification for multilingual documents. *Journal of Management Information Systems* 16(3):71–101.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on language resources and evaluation*. volume 17991802, page 21.

Annette Rios. 2016. A basic language technology toolkit for quechua .

Ali Selamat and Nicholas Akosu. 2016. Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University-Computer and Information Sciences* 28(4):457–469.

Universidad Católica Sedes Sapientiae. 2015. *Relatos Matsigenkas*. Universidad Católica Sedes Sapientiae.

Pilar Valenzuela. 2003. *Transitivity in shipibo-konibo grammar*. Ph.D. thesis, University of Oregon.

Roberto Zariquiey Biondi. 2011. *A grammar of Kashibo-Kakataibo*. Ph.D. thesis, La Trobe University.