Optimizing a Supervised Classifier for a Difficult Language Identification Problem

Yves Bestgen
Laboratoire d'analyse statistique des textes - LAST
Institut de recherche en sciences psychologiques
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

This paper describes the system developed by the Laboratoire d'analyse statistique des textes for the Dravidian Language Identification (DLI) shared task of VarDial 2021. This task is particularly difficult because the materials consists of short YouTube comments, written in Roman script, from three closely related Dravidian languages, and a fourth category consisting of several other languages in varying proportions, all mixed with English. The proposed system is made up of a logistic regression model which uses as only features n-grams of characters with a maximum length of 5. After its optimization both in terms of the feature weighting and the classifier parameters, it ranked first in the challenge. The additional analyses carried out underline the importance of optimization, especially when the measure of effectiveness is the Macro-F1.

1 Introduction

Identifying the language in which something is written is a prerequisite for many NLP systems, for example in information retrieval and machine translation, but also for applications as common as spell checkers. This task therefore captured the attention of researchers many years ago and, as highlighted by Jauhiainen et al. (2019), it was quickly considered to be very largely solved. This conclusion, obtained in ideal situations (only one language per text, reduced number of potential languages, relatively long documents) was far too optimistic as shown by the series of VarDial evaluation campaigns (Zampieri et al., 2020). The Dravidian Language Identification (DLI) shared task of VarDial 2021 (Chakravarthi et al., 2021) is undoubtedly a perfect example of a particularly complex situation even if the number of languages to be distinguished is small since it focuses on three Dravidian languages. The major problems it poses are:

- As stated by the task organizers, the three target Dravidian languages are closely related and some of the words are common in all these languages.
- The material includes a fourth category labeled *Other-language* which appears to contain several other languages in varying proportions.
- The material consists of short YouTube comments, written in Roman script and not in native script, which can make different languages more similar, but also standardize the way authors write. A significant number of them contain only one or two words like *Neerali* (Malayalam), *BGM chindi* (Kannada) or even a series of letters such as *A S U R A N* (Tamil).
- If two languages are represented by a large number of comments (almost 11,000 instances for Tamil and 4,000 for Malayalam), the learning material contains only 493 comments in Kannada and 1,008 in Other-language.
- Above all, these comments are one of the most extreme instances of code-mixed sentences, including inter-sentential and intra-sentential switches between a Dravidian (or other) language and English (Chakravarthi et al., 2020b,a). Both the syntax and the lexicon of the two languages are mixed and often in several points of a comment as in the following examples: *But ore oru varutham ennana Sunday release pannirukkalam yesterday aavathu announcement pannirukkalam* (Tamil) and *lalettan marana mass..intha movie bayangara hit adikkum...love from tamilnadu* (other-language).

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 96–101

Another difficulty is that all the instances to be classified are YouTube comments. They are therefore very similar in nature, structure and by the author's aims regardless of the language in which they are expressed. For example, many of them relate only to the number of views (*326K likes in 1 HOUR*, Tamil) or to a portion of the video that is considered by the author to be particularly interesting (*2.08 to 2.10 gun rotating style semma super*, Tamil).

In a previous edition of VarDial (Zampieri et al., 2017), I obtained good results with a relatively simple system, based mainly on character n-grams and a supervised classifier (Bestgen, 2017). The present study was aimed at determining whether this approach could be competitive in the DLI shared task. This report describes the system proposed by the Laboratoire d'analyse statistique des textes (LAST), simpler than the one used for VarDial 2017, but better optimized. It ranked first out of four systems, but it should be noted that the differences between the systems were small.

The remainder of this report first presents the data provided by the task organizers and describes the proposed system. The following section analyzes the performance obtained by comparing it to other participating systems and assesses the importance of each component and parameter of the system to its efficiency.

It must be mentioned that the LAST also participated in this year extension of the ULI 2020 task (Jauhiainen et al., 2020). I decided to not discuss this participation here partly because the system used is the same as that proposed for the DLI 2021 task and because it was not able to end up with better scores than the organizers' baseline. This is a classic criterion for eliminating underperforming systems.

2 Data provided for the challenge

The learning materials for this shared task consist of 16,674 YouTube comments from three South Dravidian subgroup language: Tamil (ISO 639-3: tam), Malayalam (ISO 639-3: mal), and Kannada (ISO 639-3: kan). All these comments were written in Roman script and most of them contain some code switching: Dravidian languages grammar with English lexicon or English grammar with Dravidian languages lexicons. The task was made even more complex by adding a fourth category of comments, also in Roman script, labeled *Other-language*, whose contains was not made clear to the participants. The average length of the comments is 8.5 tokens and 58 characters. A quarter of the comments contain 6 tokens or less. An important feature of the materials is the presence of a strong imbalance in class frequencies, with two thirds of the instances belonging to the Tamil category and one quarter belonging to the Malayalam category (see Table 1).

	Learn		Test	
Language	Freq.	%	Freq.	%
Kannada	493	3.0	63	1.4
Malayalam	4,204	25.2	1,171	25.5
Tamil	10,969	65.8	3,049	66.5
Other-language	1,008	6.0	305	6.6
Total	16,674	100	4,588	100

Table 1: Frequency distribution of the languages according to the material sets.

The test materials consisted of 4,588 comments in one of these four categories. As shown in Table 1, the imbalance between categories is even greater in the test material, an observation confirmed by a chi-square test for a contingency table ($chi^2(3) = 37.0, p < .0001$). The difference is mainly observed in the Kannada category which represents 3% of the learning set, but only 1.37% of test set.

3 Developed System

In the *Closed* submission type in which I participated, only materials provided by the organizers could be used. The main consequence was to make the automatic identification of words written in English more or less difficult and any syntactic analysis impossible. I thus chose to base the proposed system on character n-grams alone, an approach frequently used for this kind of task (Zampieri et al., 2020). The system was optimized during the training phase by means of a 5-fold cross-validation procedure, the folds of which being stratified according to the categories. During this phase, the official measure to rank the challenge participants was unknown. Due to the strong imbalance of the categories (see Table 1), I assumed that the Macro-F1 would be chosen (Opitz and Burst, 2021) and therefore used it during optimization.

This cross-validation step led to design the system as described below.

- The character n-grams of length 1 to 5 were extracted from the lowercased comments, keeping every character including spaces between tokens and punctuation marks. The n-grams located at the beginning or at the end of a comment were distinguished from the others by the presence of a specific symbol. All n-grams observed less than twice in the materials were discarded.

- Each n-gram character was weighted by BM25 (Robertson and Zaragoza, 2009; Bestgen, 2017), which is considered as one of the most effective weighting schema (Manning et al., 2008). It is a kind of TF-IDF that takes into account the length of the document. The formula is provided in Bestgen (2017).

- For each instance, the features thus formed were normalized by the L2 norm.

- The supervised classifier used was the LIBLINEAR L2-regularized logistic regression (dual) model (Fan et al., 2008). Two parameters were optimized: the regularization parameter C, which was set at 9, and the -wi options for adjusting the parameter C of different categories, which was set at 300 for Kannada, 24 for Malayalam, 1 for Tamil and 310 for Other-language.

4 Analyses and Results

The organizers of the challenge decided to use as performance measure the weighted average F1-score (WA-F1), which weights the F1-score for each class by its support and is very close to the Micro-F1. When it is important that the systems be the most effective in the most populated categories, the choice of the WA-F1 is evident. Its main weakness is to give very little importance to rare categories, as it is the case with Kannada and, to a lesser extent, with "Other-language". As the proposed system has been optimized for the Macro-F1, the results will be presented by means of both measures as did the task organizers in another study on these same languages (Chakravarthi et al., 2020a). It is important to note, however, that when some classes are very rare, a small change in effectiveness on them can have a big impact on the Macro-F1 while it will hardly change the WA-F1.

Team	WA-F1	Macro-F1
LAST	0.928	0.810
HWR	0.923	0.793
Nayel	0.915	0.765
Phlyers	0.896	0.728

Table 2: Results of the four participating teams on the test set.

4.1 Shared Task Results

As shown in Table 2, the proposed system ranked first of the four participating teams for both metrics. Looking at the system performance for the different languages, Table 3 shows that the Other-language and Kannada categories are the most difficult to identify. The performance on the Other-language category is even very poor and this is the case for all the systems submitted. On the other hand, the effectiveness for the two most frequent categories is excellent.

Language	Prec.	Recall	F1
Kannada	0.857	0.659	0.745
Malayalam	0.939	0.947	0.943
Tamil	0.961	0.959	0.960
Other-language	0.577	0.605	0.591

Table 3: Results for the four languages.

	Kan	Mal	Tam	Oth	Total
Kan	54	2	4	3	63
Mal	1	1,100	38	32	1,171
Tam	12	28	2,929	80	3,049
Oth	15	31	83	176	305
Total	82	1,161	3,054	291	4,588

Table 4: Confusion matrix.

The confusion matrix (Table 4) shows that the system distinguishes very well between the Kannada and Malayalam categories. It succeeds in identifying the true Kannada instances, but tends to assign instances of two other categories to it, as confirmed by the very clear difference between precision and recall for this category (Table 3). The errors for the Other-language category fall into the other three. It is difficult to go further in the analysis of this category without additional information

on how it was constituted.

4.2 Factors Affecting the System Performance

As the organizers provided the gold-label for the test set, a series of analyzes were performed to assess almost all of the parameters and decisions made during the system development. As a reminder, the values of these paramters in the submitted system were a maximum n-gram length of 5, a frequency treshold of 2, a BM25 weighting, a L2 normalisation and a C equal to 9 with -wi. As above, both the WA-F1 and the Macro-F1 are provided in Table 5. It is interesting to note that the conclusions resulting from the analysis of the test set are generally identical to those obtained by the cross-validation approach. The main results are as follow:

- The maximal length of the character n-grams should at least be 4, but 5 is needed to get the best Macro-F1.
- Setting the minimum frequency threshold above 2 affects only the Macro-F1.
- Bestgen (2017) observed that the BM25 weighting was more effective than the sublinear TF-IDF. In the present study, the sublinear TF-IDF (Zampieri et al., 2015) and the log-entropy weighting schema (Jarvis et al., 2013) are as effective as BM25 for the WA-F1. On the other hand, for the Macro-F1, BM25 is more effective. It is also observed that a logarithmic weighting of the n-gram frequencies is desirable.
- L2 normalization clearly improves performance, especially on the Macro-F1.
- Using the -wi parameter improves the Macro-F1. This result was expected since this parameter was designed to improve LIBLINEAR processing of unbalanced data. On the other hand, the C parameter has only a limited impact, except when it is as small as 1.

In the preceding analyzes, only one parameter was modified at a time in order to avoid a combinatorial explosion if a complete design had been evaluated. This last analysis presents the performance of a system that has not been fully optimized to compare it to the other three systems which participated in the shared task. This system employs n-grams from 1 to 5 characters, a frequency threshold of 10, a logarithmic weighting and no normalization. As the two LIBLINEAR parameters have little impact on its effectiveness, it is the version using the optimized C and wi parameters, which narrowly obtains the best scores, that is presented. Its WA-F1 was 0.915 (-0.014) and its Macro-F1 was 0.772 (-0.032). It would have been tied for 3rd, quite far from the system that came second (see Table 2). Above all, its Macro-F1 is much lower.

Parameter	WA-F1	Macro-F1
N-gram length		
3	**-0.021**	**-0.049**
4	-0.005	-0.017
6	-0.001	-0.005
7	-0.001	-0.005
8	-0.003	-0.009
Freq. threshold		
3	-0.001	-0.004
5	-0.002	-0.010
10	**-0.003**	**-0.013**
Weighting		
Binary	-0.006	**-0.017**
Freq.	**-0.010**	**-0.017**
Logarithmic	-0.005	-0.012
TF-IDF	0.000	-0.003
Log-entropy	0.000	-0.004
Normalisation		
None	**-0.007**	**-0.030**
Range [0,1]	**-0.007**	**-0.030**
C without -wi		
1	**-0.038**	**-0.140**
9	-0.013	-0.038
50	-0.007	-0.017
200	-0.006	-0.017
1000	-0.005	-0.016
C with -wi		
1	**-0.014**	**-0.019**
5	-0.002	-0.005
8	-0.001	-0.002
10	0.000	-0.001
15	-0.001	-0.003

Table 5: Impact of the parameter values on the test set performance: difference from the submitted system. The most negative differences, corresponding to the worst value of a parameter, are in bold.

In these analyzes, the differences are clearly greater on the Macro-F1 than on the WA-F1. Moreover, the comparison of these two scores shows that, if one can have the impression that the automatic identification of the three Dravidian languages is on the way to being solved when using the WA-F1, it is very far from being the case if the Macro-F1 is privileged. Measuring performance in automatic classification on unbalanced data has already caused much debate and the present study was not designed to provide an answer. At the very least, it underlines how important it is that the measure used in a shared task is carefully considered and that the participants are informed before the start of the training phase.

5 Discussion and Conclusion

Given the difficulty of the DLI 2021 task, the level of performance achieved by the system is appreciable. Identifying the Other-language category was particularly difficult because it may be thought that it is not homogeneous but composed of different languages in varying proportions. It is not even certain that all the other languages present in the test set were also present in the learning set. The additional analyzes highlighted the positive impact of a series of decisions made during the development of the system, most notably a maximum length of n-grams greater than 3, a logarithmic weighting of the frequency, an L2 normalization and optimising LIBLINEAR parameters in cross validation.

The designed system completely neglects an important dimension of the task: code-switching. The main reason is that since a system participating in the *Closed* submission type was not allowed to use any data or linguistic resources other than those provided by the organizers (see the *Submission Types* section in https://sites.google.com/view/vardial2021/evaluation-campaign), trying to identify the English words might have required a lot of effort while the simple use of an English word lists would have made this task much simpler. Taking code-switching into account is certainly a development path.

A question that could be interesting to address in the future is to try to determine if certain parameters are generalizable to the previous VarDial shared tasks (see Jauhiainen et al. (2019) for an extensive survey). BM25 seems a good choice. One might think that logarithmic weighting and L2 normalization are too. Setting the n-gram length at 5 based on the present study is certainly much more questionable. A priori, the language to be identified should have a significant impact, but on the other hand, it's still about character n-grams with the same supervised classifier.

Acknowledgments

The author is a Research Associate of the Fonds de la Recherche Scientifique (FRS-FNRS).

References

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain.

Bharathi Raja Chakravarthi, Mihaela Găman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial Evaluation Campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of The 8th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*, pages 111–118.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic Language Identification (ULI) 2020 shared task dataset and the Wanca 2017 corpus. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 688–698.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge University Press.

Juri Opitz and Sebastian Burst. 2021. Macro F1 and macro F1.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. 2015. Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 66–72, Hissar, Bulgaria.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.