Twitter Language Identification using Rational Kernels and its potential application to Sociolinguistics

Identificación de lengua en Tuiter con kernels racionales y su potencial aplicación a la sociolingüística

Jordi Porta
Departamento de Tecnología y Sistemas
Centro de Estudios de la Real Academia Española
c/ Serrano 187-189. Madrid 28002
porta@rae.es

Resumen: Este artículo presenta las técnicas empleadas por el sistema presentado a la tarea compartida TweetLID para la identificación de lengua en Tuiter. Se describen tanto el uso de máquinas de soporte vectorial con kernels racionales como el algoritmo para el etiquetado de varias lenguas. También se incluye una evaluación y una aplicación a la sociolingüística.

Palabras clave: Tuiter, Identificación de lengua, Máquinas de soporte vectorial, Kernels racionales, Cambio de código.

Abstract: This paper describes the techniques used by the system presented at the TweetLID shared task for Twitter language identification. The system is based on Support Vector Machines and Rational Kernels. An algorithm for multilanguage labeling is described. Its evaluation and application to Sociolinguistics is also included.

Keywords: Twitter, Language Identification, Support Vector Machines, Rational Kernels, Code-switching.

1 Introduction

The TweetLID shared task[1] consists in identifying the language or languages in which tweets are written, focusing on events and news generated within the Iberian Peninsula (San Vicente et al., 2014). However, despite language identification (LI) has reached a great success in discriminating between distant languages, fine-grained identification is still a challenge for language technologies and there remain two major bottlenecks according to Zampieri (2013): the discrimination between similar languages, varieties and dialects; and multilingualism, code-switching and moreover, noisy or non-standard features in texts. The first problem was addressed by the author with maximum entropy models at word level (Porta and Sancho, 2014) and others in the DSL shared task (Zampieri et al., 2014). Multilingualism was addressed by Lui, Jey Han Lau, and Baldwin (2014) using probabilistic mixture models and the identification of language in short texts by Vatanen, Väyrynen, and Virpioja (2010) with n-gram language models. However, in this paper, to address the task defined in TweetLID, related with the second of the aforementioned bottlenecks in LI, we will use n-grams of characters and support vector machines (SVMs) with rational kernels.

2 System Description

Kernel functions are commonly used to extend statistical learning methods such as SVMs to define non-linear decision boundaries. The most widely used kernels are the linear, polynomial or Gaussian ones which are applied over vector spaces (Cortes and Vapnik, 1995). In the case of natural language processing (NLP), it is common practice to represent text sequences into vector spaces as bags of words or n-grams of words or characters. However, it is possible to use string kernels to define other similarity measures between sequences as the number of common substrings of two sequences, allowing mismatches, gaps or wildcards, or the weights assignment to particular substrings (Lodhi et al., 2002; Leslie, Kuang, and Bennett, 2004). Rational kernels are a family of sequence kernels constructed from weighted finite state transducers covering all string

[1] http://komunitatea.elhuyar.org/tweetlid/

kernels commonly used in machine learning applications in bioinformatics and NLP (Cortes et al., 2004).

Lang.	#Ex.	Len.	%Acc1	%Acc2
ca	1,435	2–5	96.44	98.22
en	1,058	2–5	97.73	98.26
es	7,670	1–5	88.47	93.38
eu	478	1–5	98.13	99.37
gl	696	3–4	95.24	97.02
pt	1,920	1–5	95.62	97.66

Table 1: Minimum and maximum lengths of the n-grams used for each language classifier and estimated accuracies using cross-validation on the unambiguous examples in the training dataset applying different preprocessing. In %Acc1 only hashtags, user mentions and URLs are removed, and tokens and punctuation are split. Additionally, in %Acc2, text is lowercased and reduplicates are removed.

For TweetLID, the problem of labeling multiple languages has been tackled with binary classifiers trained with the so-called one-versus-all technique: learning k binary classifiers, discriminating one language from the rest. Different variable length n-gram kernels have been used for each language. The best parameters for each kernel have been estimated from the results on the unambiguous examples in the training dataset by cross-validation. A preprocessing step is carried out with a transducer that removes URLs, hashtags ('#Buzz'), and username mentions ('@justinbieber'); converts the text to lower case; splits words and punctuation; normalizes blanks; and removes reduplicates ('hoooola' → 'hola'). Other manipulations like diacritics removal has not found to improve results. This transducer is incorporated into the rational kernel in order to be applied before the n-gram kernel. Classifiers have been implemented with the OpenKernel library[2] with default parameters. Accuracy results for each language are shown in Table 1, where the improvement due to the adding of the preprocessing step can be noticed.

To assign more than one language to a tweet, the classifying function $c_k(x) = \text{sgn}(s_k(x))$, mapping the score of each example to $\{-1, +1\}$, has been replaced with the underlying scoring function s_k, whose values can be interpreted as confidence scores and can be used to compare classifiers without calibration. The algorithm for assigning one or more languages to a tweet combines the output of the classifiers with heuristic criteria in the form of a decision tree as follows: For each text sample x and each language k, the scores $s_1(x), \ldots, s_k(x)$ are computed. There are three situations: (a) there exists only one $s_k(x) > 0$; (b) there is more than one $s_k(x) > 0$; and (c), there is no $s_k(x) > 0$. In (a), language k is assigned to x. Case (b) has several subcases, depending on the number of languages with positive scores, the length in words of x and the difference in scores, x is finally labeled either with many languages or it is assigned the 'und(efined)' label. In case (c), x is classified with the higher scored language, if its value is over a given empirically determined threshold, or as 'other', otherwise.

3 Evaluation

The distribution of errors of the classifier on the test dataset is shown in Table 3. At the level of single language tweets, Galician (gl) obtains the worst results. As can be seen in Table 2, most Galician tweets are incorrectly classified as Portuguese (pt) or Spanish (es), but Portuguese, which is genetically most related to Galician[3], does not suffer from this problem. A shift in classification from underrepresented to overrepresented languages could be caused by the unbalanced representation of languages in the training set. Precision and recall of 'other' is rather low when compared to specific language figures.

Due to restrictions on the distribution of Twitter content, the tweets of the TweetLID corpus were provided through their IDs. Unfortunately, a number of tweets of the reference were not always available to the participants for different reasons, but are taken into consideration in the official final evaluation, affecting negatively recall and F-score (see Table 3). An alternative evaluation considering only the tweets in the reference available to the system is shown in Table 4. Results on both tables are similar, indicating that performance estimation is sound.

The evaluation of the multiple language labelings has led to the following section.

[2] `http://www.openkernel.org`

[3] Galician is genetically related to Portuguese but orthographically related to Spanish.

	Predicted								
Actual	ca	en	es	eu	gl	pt	other	amb	und
ca	1,248	2	102	1	0	6	48	18	1
en	18	773	54	2	0	10	41	10	2
es	58	81	11,264	13	44	54	200	30	8
eu	3	1	34	298	0	1	10	9	2
gl	1	0	201	0	101	40	59	21	0
pt	3	3	112	1	1	1,736	52	19	1
other	70	30	105	6	4	15	155	0	0
amb	26	24	226	32	2	14	20	9	0
und	85	29	407	11	2	110	156	4	10

Table 2: Confusion matrix of the test dataset

Lang.	**Prec.**	**Rec.**	**F-score**
ca	0.838	0.850	0.844
en	0.840	0.737	0.786
es	0.921	0.952	0.936
eu	0.905	0.746	0.818
gl	0.665	0.284	0.398
pt	0.912	0.898	0.905
amb	1.000	0.746	0.855
und	0.366	0.298	0.328
Total	0.806	0.689	0.734

Table 3: Results taking into account all the 18,423 tweets in the reference. Unavailable tweets of the reference (72) affect both recall and F-score negatively.

Lang.	**Prec.**	**Rec.**	**F-score**
ca	0.838	0.855	0.846
en	0.840	0.741	0.787
es	0.921	0.955	0.938
eu	0.905	0.747	0.819
gl	0.665	0.284	0.398
pt	0.912	0.905	0.908
amb	1.000	0.749	0.856
und	0.366	0.298	0.328
Total	0.806	0.692	0.735

Table 4: Results taking only into account submitted results in the reference (18,351 tweets in common).

4 Application to Sociolinguistics

Language contact has been a hot issue in linguistics since the publication of *Languages in contact* (Weinreich, 1953) and represents one of the most common scenarios for the study of language variation and change, including code-switching (CS). In informal communication, CS is a pervasive phenomenon by which multilingual speakers switch back and forth between their languages. CS is present at the inter-sentential, intra-sentential and even morphological levels. The system presented in this paper could be applied to CS to unveil part of the underlying sociolinguistic structure of communities and, at the same time, when this structure is known in advance, it can also be used to evaluate the predictive power of the method used by the system. In the case of the Iberian Peninsula, it is the westernmost southern European peninsula separated from the rest of Europe at the north-east edge by the Pyrenees. In the Iberian Peninsula, the six top languages found in tweets are Basque, Catalan, Galician, Spanish, Portuguese and English. Except for English, which is a global language, and Basque, which is a language isolate, the rest of Iberian languages descend from Vulgar Latin spoken in the Peninsula. Spain has an official language, Castilian (also known as Spanish) but the central government has transferred some of its powers to regional governments, known as autonomous communities, some of them having co-official languages. Table 5 contains two matrices with the number of pairs of languages cooccurring in tweets. Table 5.a is computed using the manually labeled examples of the training and test datasets while Table 5.b is computed from the predictions on the test dataset. There are four identifiable blocks in those matrices: (1) English (en), which is a global language, and cooccurs with the rest of languages; (2) Portuguese (pt), which is a national language with little contact with Spanish; (3) Spanish (es) a national language cooccurring with the Spain's co-official languages: Catalan (ca), Galician (gl) and Basque (eu); and (4) the block of the co-officials, which are not seen together in tweets because they are not languages in contact.

	es	pt	en	ca	gl	eu
es	20,356	1	275	111	25	231
pt	1	4,094	34	-	5	-
en	275	34	1,913	44	4	16
ca	111	-	44	2,901	-	-
gl	25	5	4	-	930	-
eu	231	-	16	-	-	738

(a) Labeled examples in TweetLID datasets

	es	pt	en	ca	gl	eu
es	12,546	21	12	34	25	13
pt	21	2,003	3	2	3	-
en	12	3	949	3	1	-
ca	34	2	3	1,517	-	-
gl	25	3	1	-	155	-
eu	13	-	-	-	-	365

(b) Predictions on the TweetLID test dataset

Table 5: Matrices with the language cooccurrences on tweets. For tweets containing more than two languages, (e.g., 'en+es+eu'), all their pairs have been computed (e.g., 'en+es', 'en+eu' and 'es+eu'). The matrix in (a) has been computed from the labeled examples in the datasets of TweetLID and the system's predictions for the test dataset in (b).

5 Conclusions and Future Work

Results from the Evaluation Section suggest there is still room for potential improvements. A more balanced representation of languages or the introduction of a cost matrix could improve the performance of underrepresented languages as Galician. The labeling of tweets in the language category 'other' could improve both precision and recall of other languages. Finally, it is also left as future work to combine the output of the individual classifiers with multilabel learning methods (Zhang and Zhou, 2014), in order to improve and replace the heuristic presented in this paper for multilanguage labeling.

References

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Leslie, Christina, Rui Kuang, and Kristin Bennett. 2004. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5:1435–1455.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, March.

Lui, Marco, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

Porta, Jordi and José-Luis Sancho. 2014. Using maximum entropy models to discriminate between similar languages and varieties. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial-14)*, Dublin, Ireland.

San Vicente, Iñaki, Arkaitz Zubiaga, Pablo Gamallo, José Ramon Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2014. Overview of TweetLID: Tweet language identification at SEPLN 2014. In *TweetLID @ SEPLN 2014*.

Vatanen, Tommi, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-10)*. European Language Resources Association (ELRA).

Weinreich, Uriel. 1953. *Languages in contact. Findings and Problems.* Mouton, Hague and Paris.

Zampieri, Marcos. 2013. Using bag-of-words to distinguish similar languages: How efficient are they? In *Proceedings of the 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI-13)*, pages 37–41.

Zampieri, Marcos, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland, August.

Zhang, M.-L. and Z.-H. Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.