Overview of TweetLID: Tweet Language Identification at SEPLN 2014

Introducción a TweetLID: Tarea Compartida sobre Identificación de Idioma de Tuits en SEPLN 2014

Arkaitz Zubiaga[1], Iñaki San Vicente[2], Pablo Gamallo[3], José Ramom Pichel[4] Iñaki Alegria[5], Nora Aranberri[5], Aitzol Ezeiza[5], Víctor Fresno[6]

[1] University of Warwick, [2] Elhuyar, [3] USC
[4] imaxin|software, [5] University of the Basque Country, [6] UNED
tweetlid@elhuyar.com

Resumen: Este artículo presenta un resumen de la tarea compartida y taller TweetLID, organizado junto a SEPLN 2014. Resume brevemente el proceso de colección y anotación de datos, el desarrollo y evaluación de la tarea compartida, y por último, los resultados obtenidos por los participantes.
Palabras clave: identificación de idioma, tuits, textos cortos, multilingüismo

Abstract: This article presents a summary of the TweetLID shared task and workshop held at SEPLN 2014. It briefly summarizes the data collection and annotation process, the development and evaluation of the shared task, as well as the results achieved by the participants.
Keywords: language identification, tweets, short texts, multilingualism

1 Introduction

Recent research shows that while Twitter's predominant language was English in the early days, the global growth and adoption of the social media platform in recent years has increased the diversity in the use of languages (Lehman, 2014). This has in turn sparked an increasing interest of the scientific community in automatically guessing the languages of tweets (Carter, Weerkamp, and Tsagkias, 2013). The identification of the language of a tweet is crucial for the subsequent application of NLP tools such as machine translation, sentiment analysis, or information extraction. This kind of NLP tools tend to be crafted with resources specifically trained for a language or some languages. Hence, accurately identifying the language of a tweet would facilitate the application of NLP resources suitable to the language in question.

Twitter itself does provide a language id along with each tweet's metadata, but as we show in this article it leaves much to be desired in terms of accuracy. Besides, it is intended to detect major languages, and does not identify other languages with lesser presence on the platform such as Catalan, Basque or Galician, which account for millions of native speakers within the Iberian Peninsula. Following up on a recent shared task on normalization of tweets (Alegria et al., 2013; Alegria et al., 2014), we have organized a new shared task on tweet language identification. This task focuses specifically on the 5 top languages of the Iberian Peninsula (Spanish, Portuguese, Catalan, Basque and Galician), and English. These languages are likely to co-occur along with many news and events relevant to the Iberian Peninsula, and thus an accurate identification of the language is key to make sure that we use the appropriate resources for the linguistic processing. This task has intended to bring together contributions from researchers and practitioners in the field, to develop and compare tweet language identification systems designed for the aforementioned languages, which can potentially later be extended to a wider variety of languages.

This shared task has enabled the

development of an annotated tweet corpus that set out participants to deal with three novel aspects with respect to previous research in tweet language identification: (i) some of the languages considered in the task belong to the same language family, which makes the distinction of these languages an extra challenge, (ii) tweets can be occasionally multilingual, with parts of it written in different languages –manually annotated as lang1+lang2–, and (iii) tweets can be ambiguous occasionally given their brevity, i.e., it is not possible to determine which of two (or more) languages a tweet is written in, and therefore any of them should be deemed correct. This article serves as an introductory overview of the shared task, describing the process of generation of the corpus, providing a brief description of the systems developed by the participants, as well as the performance of these systems evaluated in comparison with human assessments.

2 Language Identification

Language identification consists in determining the language in which a text is written. It has usually been tackled as a classification problem in previous research, and the best known approaches make use of n-grams to learn the model for each of the languages, as well as to represent each of the documents to be categorized into one of the languages (Cavnar, Trenkle, and others, 1994).

Language identification has progressed significantly in recent years. The task has been considered solved for certain situations (McNamee, 2005), assuming among others that documents are long enough and that are written in a single language. However, the emergence of social media and the chatspeak employed by its users has brought about new previously unseen issues that need to be studied in order to deal with these kinds of texts. Three key issues posited in the literature and that, as of today, cannot be considered solved include: (i) distinguishing similar languages (Zampieri, 2013), (ii) dealing with multilingual documents (Lui, Lau, and Baldwin, 2014), and (iii) language identification for short texts (Bergsma et al., 2012; Carter, Weerkamp, and Tsagkias, 2013; Laboreiro et al., 2013). The shared task organized at TweetLID has considered these three unresolved issues, and has enabled participants to compare the performance of their systems in these situations.

This task includes the five top languages of the Iberian Peninsula, which are spoken in different regions, and four of them –Spanish, Portuguese, Catalan, and Galician– are romance languages with certain similarities among them, which makes the task more challenging. The fifth language –Basque–, and English, belong to different language families, and therefore are rather different from the rest. Still, their cultural proximity, and the fact that many users in the area are bilingual, entails that they often mix words and spellings across languages. For instance, a Basque native might naturally write something like *"nos vemos, agur!"* (see you later, bye!), when *"nos vemos"* is in Spanish, while *"agur"* is Basque to say good bye; similarly, a Catalan speaker might often misspell the Spanish word *"prueba"* (test) as *"prueva"*, given that the Catalan translation of the word (*"prova"*) is written with *v*. These characteristics are common in bilingual areas, and have been considered in the definition of this task in order to carefully develop the annotation guidelines and to pursue the final annotation of the corpora.

3 Description of the Task

The TweetLID shared task consists in identifying the language or languages in which tweets are written. It is focused on the most widely used languages of the Iberian Peninsula, which provides an ideal context where news and events are likely to be shared and discussed in multiple languages.

To the end of setting up a common evaluation framework to enable comparison of different language identification systems, we have put together an annotated corpus of nearly 35,000 tweets and defined a methodology to evaluate the multi-label output of the language identification systems. Splitting the corpus into a training set with 15k tweets, and a test set with 20k tweets, the participants had a month to develop their language identification systems making use of the training set. They then had 72 hours to work on the test set and submit their results.

Besides the challenge of dealing with the short and often informal texts found in tweets, the task has considered that a tweet is

not necessarily written in a single language. This is especially true in bilingual regions, where speakers that feel equally comfortable with either of their two native languages tend to switch between them and mix them in a sentence quite frequently. Hence, the task has also considered a number of cases where the response is not basically one of the languages in the list: (i) a tweet can combine two –or occasionally three– languages in a tweet, e.g., when a tweet has parts in Catalan and Spanish, (ii) given the similarity and cultural proximity between some of the languages, it is not possible to determine which of two –or more– languages a tweet is written in, e.g., some tweets might be written equally in Catalan or Spanish, (iii) despite the geographical restriction of the tweets in the task, it is also likely that tweets in other languages occur, such as French, and (iv) it is not possible to determine which of the 6 languages considered in the task a tweet is written in, e.g., when a tweet only mentions entities, smileys, or onomatopoeias.

4 Data Collection

To collect an unrestricted set of tweets, but rather focused on the set of languages within the scope of TweetLID, we relied on geolocation to retrieve tweets posted from areas of interest. We used Twitter's streaming API's `statuses/filter` endpoint to collect geolocated tweets posted within the Iberian Peninsula from March 1 to 31, 2014. While this stream is limited to tweets explicitly providing geolocation metadata, it allows to track a diverse set of tweets that is not restricted to a specific set of users or domain. Having collected these tweets, we used Nominatim[1] to obtain specific location information for each tweet. Given the coordinates of a tweet as input, Nominatim queries OpenStreetMap for the specific address associated with those coordinates, i.e., region, city, and street (if available) from which the tweet has been sent. This led to the collection of 9.7 million tweets with location details associated. From this set of tweets, we sampled tweets from **Portugal** and the following **3 officially bilingual regions**:

- **Basque Country**, where Basque and Spanish are spoken. Tweets from the province of Gipuzkoa were chosen here to represent the Basque Country.
- **Catalonia**, where Catalan and Spanish are spoken. Tweets from the province of Girona were chosen to represent Catalonia.
- **Galicia**, where Galician and Spanish are spoken. Tweets from the province of Lugo were chosen.

One province was picked from each of the regions to avoid cases such as that of the province of Barcelona in Catalonia, which is much more diverse in terms of languages due to tourism. These three bilingual regions enabled us to sample tweets in Basque, Catalan, Galician, and Spanish, and we could sample Portuguese tweets from Portugal. English is the sixth language in the corpus, which can be found all across the aforementioned regions. For the final corpus to be manually annotated, we picked 10k tweets from each of the bilingual regions, and 5k from Portugal. The tweets picked here had to contain at least one word (i.e., string fully made of a-z characters), so that there is some text, and tweets with e.g. only a link are not considered. The next section describes the manual annotation performed on this corpus with 35k tweets.

5 Manual Annotation

The collection of 35k tweets resulting from the aforementioned process was then manually annotated. Each of the tweets was associated with its corresponding language code in the manual annotation process. The manual annotation was conducted by annotators who were native or proficient speakers in at least three languages considered in the task. This enabled us to distribute the tweets from each of the four regions to different annotators, so that each annotator was a native or proficient speaker of the languages spoken in the region in question.

The annotators were instructed to assign codes to tweets according to the language in which they were written. We asked them to ignore #hashtags and @user mentions, as well as references to named entities in another language. For instance, in the tweet *Acabo de ver el último capítulo de la temporada de 'the walking dead', muy bueno!*

[1] http://wiki.openstreetmap.org/wiki/Nominatim

(Spanish: I just saw the season finale of 'the walking dead', it's amazing!), only Spanish should be annotated.

They had to assign codes to the tweets as follows: *eu* for Basque, *ca* for Catalan, *gl* for Galician, *es* for Spanish, *pt* for Portuguese, and *en* for English. When a different language was found in a tweet –e.g., French or German–, they had to annotate it as *other*. Additionally, when the text of a tweet included words that are widely used in any of the languages in the task –e.g., onomatopoeias such as 'jajaja' or 'hahaha', or internationalized words such as 'ok'–, which makes it impossible to determine the language being used in that specific case, they were asked to annotate it as *und*(eterminable).

In the above situations, the annotators had to mark a tweet as either being written in one of the 6 languages, *other* or *und*. However, two more cases were identified and included in the annotation guidelines: multilingual tweets, and ambiguous tweets.

Multilingual tweets contain parts of a tweet in different languages, where the annotators were instructed to annotate all of the languages being used. For instance, *Qeeeee matadaaa* (Spanish: that was exhausting) *da Biyar laneaaaa...* (Basque: and gotta go to work tomorrow) should be annotated as *es+eu*, and *Acho que vi a Ramona hoje* (Portuguese: man, I've seen Ramona today) *but im not sure* (English) should be annotated as *pt+en*. Occasionally, three languages were also found, e.g., *Egun on! Buenos días! Good morning!* (Good morning in Basque, Spanish and English), annotated as *eu+es+en*. The annotation had to consider all the languages being used, in no specific order, except when a single word or term was used as a constituent of a sentence in another language, e.g., *es un outsider* (Spanish: he is an outsider), where only one language is annotated.

Ambiguous tweets were defined as the tweets that can be categorized into the list of languages being considered, but may have been written in at least two of them. Given the similarity and cultural proximity of some of the languages, it is likely that some short texts are written equally in some languages. For instance, *Acabo de publicar una foto* (I just published a photo) can be either Spanish or Catalan, and cannot be disambiguated in the absence of more context. This case had to be annotated as *es/ca*.

6 Annotated Corpus and Evaluation Measures

The annotated corpus is composed of 34,984 tweets, with manually annotated language labels following the procedure described above. Table 1 shows the distribution of the manual annotation, where it can be seen that Spanish is the predominant language, which amounts to 61.22% of the tweets. This is why we use a macroaverage approach to evaluate the systems, as we describe later, which rewards the systems that perform well for all the languages rather than just for the predominant language.

For the purposes of the shared task, the corpus was split into two random sets of tweets: a training set with 14,991 tweets, and a test set with 19,993 tweets. However, due to restrictions on the use of the Twitter API[2], we distributed the corpora to the participants by including only the tweet IDs. We also provided them with a script to download the content of the tweets having the IDs, which scrapes the web page of each tweet to retrieve the content.

Once the participation period ended we checked the set of tweets in the test set that were still available at the moment. This was done specifically on the 7th of July, with the submission deadlines closed for all the participants. This final check found that 18,423 out of the initial 19,993 tweets, i.e., 92.1%, were available at the moment. For further details into the composition of the corpora, Table 2 shows the distribution of categories for the train and test datasets.

6.1 Evaluation Measures

The fact that the corpora (as well as the reality of Twitter itself) is unbalanced, and some languages are far more popular than others is an important issue to be considered when defining the evaluation measures. Besides, given that the language identification task has been defined as a classification problem where tweets can be either multilingual, with more than a language per tweet, or ambiguous, where it is not possible to disambiguate among a set of target languages, the evaluation measures

[2]https://dev.twitter.com/terms/api-terms

Language	Tweets	% Tweets
Spanish (es)	21,417	61.22
Portuguese (pt)	4,320	12.35
Catalan (ca)	2,959	8.46
English (en)	1,970	5.63
Galician (gl)	963	2.75
Basque (eu)	754	2.16
Undeterm. (und)	787	2.25
Multilingual (a+b)	747	2.14
Ambiguous (a/b)	625	1.79
Other	442	1.26

Table 1: Distribution of the manual annotation

Language	%Tweets Train	%Tweets Test
Spanish (es)	57.11 (8,562)	64.02 (11,794)
Portuguese (pt)	14.35 (2,151)	10.55 (1,943)
Catalan (ca)	9.78 (1,466)	7.79 (1,435)
English (en)	6.66 (999)	4.97 (914)
Galician (gl)	3.38 (507)	2.30 (423)
Basque (eu)	2.53 (380)	1.94 (358)
Undeterm. (und)	1.25 (188)	3.01 (555)
Multilingual (a+b)	2.47 (371)	1.93 (356)
Ambiguous (a/b)	2.31 (346)	1.41 (260)
Other	0.14 (21)	2.09 (385)

Table 2: Distribution of the manual annotation in train and test data sets.

need to be carefully defined to take these into account.

To deal with the imbalance, we compute the precision, recall, and F1 values for each language, and the macroaveraged measures for all languages afterwards. This is intended to provide higher scores to systems that perform well for many languages, rather than those performing very well in the most popular languages such as Spanish and Portuguese.

Given the characteristics of the task, we rely on a concept-based evaluation methodology for multi-label classification (Nowak et al., 2010), and adapt it to the specific purposes of the task. We compute Precision (P), Recall (F) and F1 measures as defined in Equations 1, 2, and 3.

$$P = \frac{1}{|C|} \sum_{i \in C} \frac{TP_i}{TP_i + FP_i} \quad (1)$$

$$R = \frac{1}{|C|} \sum_{i \in C} \frac{TP_i}{TP_i + FN_i} \quad (2)$$

$$F_1 = \frac{1}{|C|} \sum_{i \in C} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \quad (3)$$

Where $C = \{ca, en, es, eu, gl, pt, amb, und\}$ is the set of labels defined in our classification task, and TP, FP and FN refer to the counts of true positive, false positive and false negative answers respectively.

To determine whether a system's output for a tweet is correct, we compare it with the manually annotated ground truth. Given that tweets are not simply multilingual, the TP, FP and FN values are computed as follows:

- For monolingual tweets, the TP count is incremented by 1 if the answer is correct, and FP is incremented by 1 for the language output by the system otherwise. If a system's prediction contains more than one language, incorrect languages will be penalized, e.g., for a tweet annotated as "pt", a system that outputs "pt+en" will increment TP for "pt" but also FP for "en". FN will be incremented for the language in the ground truth if the answer does not contain the correct language. Hence, the system that outputs "eu" for a tweet that is actually "pt", will count as an additional FP for "eu", and as a FN for "pt".
- For multilingual tweets, we apply the same evaluation methodology as for the multilingual tweets above repeatedly for each of the languages in the ground truth, e.g., for a tweet annotated manually as "ca+es", a system that outputs just "ca" will count as TP for "ca" and as FN for "es".
- For ambiguous tweets that could have been written in any of a set of languages, any of the responses in the ground truth is deemed correct, e.g., for a tweet annotated as "ca/es", either "ca" or "es" is deemed correct as a response, counting as TP of the "amb" category

in either case. If, instead, the system outputs "pt", which is not among the languages listed in the ground truth of the ambiguous tweet, the evaluation counts as a FP for "pt", and as a FN for "amb".

Finally, note that we decided to merge tweets annotated as "other" or "und" for evaluation purposes. We did not differentiate between them as those are the tweets that need to be ruled out for being out of the scope of the task. If a system determines that a tweet is "other", and the ground truth is "und", or vice versa, it is deemed correct.

7 Results and Description of Participating Systems

Out of the initially registered 16 participants, 7 groups submitted their results for either one or both of the tracks. Participants had a 72 hour window to work with the test set and submit up to two results per track. Next, we first summarize the types of approaches that the participants utilized, and further detail the technique used by each of the participants afterwards.

7.1 Overview of the Techniques and Resources Employed

The participants have relied on very diverse and different techniques in their systems. They have employed different classification algorithms, different methods to learn the models for each language, as well as different criteria to determine the languages of a tweet. This diversity of approaches enables us to broaden the conclusions drawn from the analysis of the performance of different systems. One aspect that the participants agreed upon is the need to preprocess tweets by removing some tokens that do not help for the language identification task such as URLs and user mentions, as well as by lowercasing and reducing the repetition of characters, among others.

The participants have used different classification algorithms to develop their systems. The classification algorithms used by most participants include Support Vector Machines (SVM), and Naive Bayes, which have proven effective in previous research in language identification for longer texts.

Not all the participants have developed multilabel techniques that can deal with multilingual tweets. Only two of them actually did, mostly by defining a threshold that determines the languages to be picked for the output when the classifier provides a higher confidence score for them.

Table 3 summarizes the characteristics of the approaches developed by each of the participants.

7.2 Brief Description of the Systems

Citius-imaxin (Gamallo et al., 2014) submitted two different systems to each of the tracks. On the one hand, a system they called Quelingua builds dictionaries of words ranked by frequency for each language. New tweets are categorized by weighing the ranked words in it, as well as specific suffixes that characterize each language. On the other hand, they build another system based on a state-of-the-art bayesian algorithm, which has proven accurate in recent research. For the unconstrained track, they fed the systems with news corpora extracted from online journals for all six languages. Their systems do not pick more than one language per tweet, hence not dealing with multilingual tweets. Their bayesian system achieved the best performance for the unconstrained track. Moreover, it was the only system in the task that outperformed its constrained counterpart.

RAE (Porta, 2014) submitted two systems only to the constrained track. Their systems rely on n-gram kernels of variable length for each language. The best parameters for each kernel were estimated from the results on the unambiguous examples in the training dataset by cross-validation. They then used Support Vector Machines (SVM) to categorize each new tweet. They relied on a decision tree to interpret the output of the one-vs-all SVM approach, and thus deciding whether the confidence values for more than one language exceeded a threshold (multilingual tweet), only one did (monolingual tweet), or none did (undeterminable).

UB/UPC/URV (Mendizabal, Carandell, and Horowitz, 2014) submitted one system to each of the tracks. They developed a different type of system in this case for each track. The first system, submitted to the constrained track, makes use of a linear interpolation smoothing

TEAM	Classifier	Representation	Ext. Resources	Multiling.
Citius-imaxin	1) ranked n-grams 2) naive bayes	words & n-grams & suffixes	news corpora	no
RAE	support vector machines	n-grams	-	yes
UB/UPC /URV	1) linear interpolation 2) out-of-place measure	n-grams	-	no
IIT-BHU	n-gram distances	n-grams	-	no
CERPAMID	n-gram distances	3-grams	Europarl corpus Wikipedia	no
ELiRF @ UPV	1) support vector machines 2) Freeling	words & 4-grams	Wikipedia	yes
LYS @ UDC	TextCat & langid.py & langdetect	-	Yali	no

Table 3: Summary of the systems developed by the participants

method (Jelinek, 1997) to compute the probabilities of each n-gram to belong to a language, and weigh new tweets using those probabilities. The second system, submitted to the unconstrained track, is an out-of-place approach that builds a ranked list of n-grams for each language in the training phase, and compares each new tweet with these ranked lists to find the language that resembles in terms of n-gram ranks.

IIT-BHU (Singh and Goyal, 2014) only submitted a run to the constrained track. They adapted a system that they previously created for other kinds of texts (Singh, 2006), which is a simple language identification system that makes use of n-grams, and based on that created by (Cavnar, Trenkle, and others, 1994), to the context of Twitter. Basically, they integrated a preprocessing module that removes noisy tokens such as user mentions, hashtags, URLs, etc., and then uses a symmetric cross entropy to measure the similarity or distance between each new tweet and the models learned for each language in the training phase.

CERPAMID (Zamora, Bruzón, and Bueno, 2014) submitted two systems to each of the tracks. They extract n-grams of three characters to represent the tweets, and use three different weighting methods to weight the n-grams. Then, they give a score to each new tweet for all the languages in the collection using the three weighting schemes, and pick the final language given as output by the system through simple majority voting. As their systems only output one language, they did not develop any solutions to deal with multilingual tweets. For the unconstrained track, they used the Europarl corpus (Koehn, 2005) for English, Spanish, and Portuguese, and Wikipedia for Basque, Catalan, and Galician.

ELiRF @ UPV (Hurtado et al., 2014) submitted two systems to each of the tracks. For the constrained track, the authors made use of a one-vs-all classifier combining method using SVM. The two approaches submitted to the constrained track differ in the way they deal with multilingual tweets: on one of the approaches, they consider each combination of languages as a new category, while in the other approach they defined a threshold so that the output included all the languages for which the SVM classifier returned a higher confidence value. For the unconstrained track, they developed a classifier using SVM, which used Wikipedia to train the system but did not return multilabel outputs, and another classifier using Freeling's language identification component (Padró and Stanilovsky, 2012), which includes its own models of 4-grams for the languages in the corpus, except for Basque that the authors created themselves. The constrained method that relies on a threshold to pick the languages for the output achieved the best performance for the constrained track.

LYS @ UDC (Mosquera, Vilares, and Vilares, 2014) submitted two systems to each of the tracks. They used three different classifiers to develop their systems: TextCat (Cavnar, Trenkle, and others, 1994), langid.py (Lui and Baldwin, 2012), and langdetect (Shuyo, 2010). The two different systems they developed for both tracks differ in that one determines the final output by relying on the classifier with

#	TEAM	P	R	F1
1	ELiRF @ UPV II	0.825	0.744	0.752
2	ELiRF @ UPV I	0.824	0.730	0.745
3	UB/UPC/URV	0.777	0.719	0.736
4	RAE II	0.806	0.689	0.734
5	RAE I	0.811	0.687	0.733
6	Citius-imaxin I	0.824	0.685	0.726
7	Citius-imaxin II	0.689	0.743	0.699
8	CERPAMID I	0.716	0.681	0.666
9	LYS @ UDC I	0.732	0.734	0.638
10	IIT-BHU	0.605	0.670	0.615
11	CERPAMID II	0.704	0.578	0.605
12	LYS @ UDC II	0.610	0.582	0.498

Table 4: Constrained

#	TEAM	P	R	F1
1	Citius-imaxin I	0.802	0.748	0.753
2	ELiRF @ UPV II	0.737	0.723	0.697
3	ELiRF @ UPV I	0.742	0.686	0.684
4	Citius-imaxin II	0.696	0.659	0.655
5	LYS @ UDC I	0.682	0.688	0.581
6	UB/UPC/URV	0.598	0.625	0.578
7	LYS @ UDC II	0.588	0.590	0.571
8	CERPAMID I	0.694	0.461	0.506
9	CERPAMID II	0.583	0.537	0.501

Table 5: Unconstrained

higher confidence, while the other determines the output by majority voting. For the unconstrained track, they used the corpus provided with Yali (Majliš, 2012). Their systems return a single language as output, and does not deal with multilingual tweets.

7.3 Results

Table 4 shows the results for the *constrained* track, and Table 5 shows the results for the *unconstrained* track. The **ELiRF @ UPV** group performed best for the constrained track with an F1 of 0.752, and **Citius-imaxin** presented the most accurate system for the unconstrained track, with a very similar F1 value, 0.753.

One of the aspects that stands out from the results of the participants is the fact that most of the systems performed better in the constrained track, and the lower performance of their unconstrained counterparts suggests that either the external resources used are not suitable for the task, or they were not properly exploited. Surprisingly, only the unconstrained version of Citius-imaxin's bayesian technique outperformed its constrained counterpart. This posits an important caveat of the presented systems, which needs to be further studied in the future.

7.3.1 Results by Language

Figure 1 summarizes in a boxplot the distribution of precision values achieved by the 21 submitted systems for the different categories. It can be seen that the systems performed poorly especially for Galician (gl); this can be due to its similarity to Spanish (es) and Portuguese (pt), and its small presence in the corpus. Because of this similarity, and of course the cultural proximity where users tend to mix up spellings, the system might have had a tendency to picking the most popular languages in these cases as output. The systems performed better for the rest of the languages, but still surprisingly there is a high variation of performances for Basque (eu), where we can see that some of the systems performed poorly. This is rather surprising given that Basque is very different from the rest of the languages, being an isolate language. It also stands out that all the systems performed very well for Spanish, being this the majority language with over 60% of the tweets in the corpora.

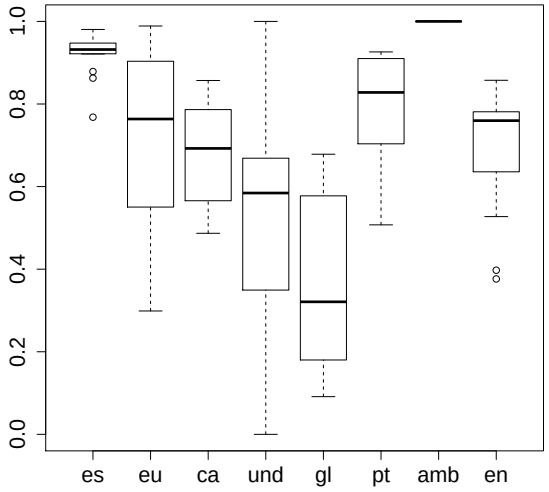

Figure 1: Distribution of precision scores by language for the 21 submitted systems

7.3.2 Alternative Microaveraged Evaluation

For the sake of comparison with the performance reported in other research works, we also show here the microaveraged evaluation of the three best systems in each track. Note that the micro-averaged evaluation favors the overall performance of the systems, regardless of their likely poor performance for some of the languages. Tables 6 and 7 show the microaveraged results, with an overall boost in the results for all the contestants. Still, the best results obtained in this shared task are from the 99.4% accuracy score reported for formal text, or the 92.4% accuracy score reported for microblogs by Carter et al. (Carter, Weerkamp, and Tsagkias, 2013). However, it is worth mentioning that Carter et al's scores rely on a monolingual tweet language identification task for major languages including Dutch, English, French, German, and Spanish. The fact that TweetLID has introduced multilingual tweets, as well as tweets from underrepresented languages led to slightly lower performances scores of 89.8% accuracy in the best case. This only reflects a 2.6% accuracy loss when compared to Carter et al's best results for tweets.

#	TEAM	P	R	F1
1	ELiRF @ UPV II	0.891	0.886	0.889
2	ELiRF @ UPV I	0.897	0.880	0.888
3	Citius-imaxin I	0.891	0.871	0.881

Table 6: Constrained (micro average)

#	TEAM	P	R	F1
1	Citius-imaxin I	0.898	0.878	0.888
2	ELiRF @ UPV II	0.839	0.854	0.847
3	ELiRF @ UPV I	0.820	0.802	0.811

Table 7: Unconstrained (micro average)

7.3.3 Comparison with Baseline Approaches

Table 8 includes two additional results as baselines that we computed using the following two solutions: (i) Twitter's metadata, which the system itself provides with each tweet, but it does not recognize Basque, Catalan, and Galician, and (ii) TextCat, a state-of-the-art n-gram-based language identification system developed for formal texts, which can deal with the six languages considered in the task. Note that TextCat was run after cleaning up the tweets by removing hashtags, and user mentions, as well lower-casing the text. The low performance of both solutions, with F1 values below 0.5, emphasizes the difficulty of the task, as well as the need for proper alternatives for social media texts.

System	P	R	F1
Twitter	0.457	0.498	0.463
TextCat	0.586	0.480	0.447

Table 8: Results of baseline systems

8 Discussion

The shared task organized at TweetLID has enabled us to come up with a benchmark corpus of nearly 35,000 tweets with manual annotations of the language in which they are written, as well as to define an evaluation methodology that allowed participants to compare their systems. For this task, we have considered the five top languages of the Iberian Peninsula –Spanish, Portuguese, Catalan, Basque, and Galician– as well as English. This has allowed participants to compare their systems with four romance languages that share similarities with one another, and two more languages that are substantially different from the rest, i.e., English and Basque.

The participants have applied state-of-the-art language identification techniques designed for other kinds of texts such as news articles, as well as adapted approaches that take into account the nature of the brevity and chatspeak found in tweets. Still, the performance of the systems posits the need of further research to come up with more accurate language identification systems for social media. Some of the key shortcomings that the shared task has brought to light include the need for a better choice of external resources to train the systems, the low accuracy of the systems when dealing with underrepresented languages which are very similar to others –as occurred with Galician here–, and the inability to identify multilingual tweets. Future work on tweet language identification should look into these issues to develop more accurate systems. A thorougher analysis of the task and performance of the participating systems will follow in an extended version of this paper.

Acknowledgments

This work has been supported by the following projects: PHEME FP7 project (grant No. 611233), QTLeap FP7 project (grant No. 610516), Spanish MICINN projects *Tacardi* (Grant No. TIN2012-38523-C02-01) and *Skater* (Grant No. TIN2012-38584-C06-01), Galician HPCPLN project (Grant No. EM13/041), Celtic (Innterconecta program, Grant No. 2012-CE138).

References

Alegria, Inaki, Nora Aranberri, Pere R Comas, Vıctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2014. Tweetnorm_es corpus: an annotated corpus for spanish microtext normalization. In *Proceedings of LREC*.

Alegria, Inaki, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2013. Introducción a la tarea compartida tweet-norm 2013: Normalización léxica de tuits en español. In *Tweet-Norm@SEPLN*, pages 1–9.

Bergsma, Shane, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Workshop on Language in Social Media*, pages 65–74. ACL.

Carter, Simon, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Cavnar, William B, John M Trenkle, et al. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

Gamallo, Pablo, Marcos Garcia, Susana Sotelo, and José Ramom Pichel. 2014. Comparing ranking-based and naive bayes approaches to language detection on tweets. In *TweetLID@SEPLN*.

Hurtado, Lluís-F., Ferran Pla, Mayte Giménez, and Emilio Sanchis. 2014. Elirf-upv en tweetlid: Identificación del idioma en twitter. In *TweetLID@SEPLN*.

Jelinek, Frederick. 1997. *Statistical methods for speech recognition*. MIT press.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Laboreiro, Gustavo, Matko Bošnjak, Luís Sarmento, Eduarda Mendes Rodrigues, and Eugénio Oliveira. 2013. Determining language variant in microblog messages. In *Proceedings of SAC*, pages 902–907. ACM.

Lehman, Brian. 2014. The evolution of languages on twitter. `http://blog.gnip.com/twitter-language-visualization/`.

Lui, Marco and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of ACL*, pages 25–30. ACL.

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.

Majliš, Martin. 2012. Yet another language identifier. In *Student Research Workshop at EACL'12*, pages 46–54. ACL.

McNamee, Paul. 2005. Language identification: A solved problem suitable for undergraduate instruction. *J. Comput. Sci. Coll.*, 20(3):94–101.

Mendizabal, Iosu, Jeroni Carandell, and Daniel Horowitz. 2014. Tweetsafa: Tweet language identification. In *TweetLID@SEPLN*.

Mosquera, Yerai Doval, David Vilares, and Jesus Vilares. 2014. Identificación automática del idioma en twitter: Adaptación de identificadores del estado del arte al contexto ibérico. In *TweetLID@SEPLN*.

Nowak, Stefanie, Hanna Lukashevich, Peter Dunker, and Stefan Rüger. 2010. Performance measures for multilabel evaluation: a case study in the area of image classification. In *Proceedings of ICMR*, pages 35–44. ACM.

Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of LREC*.

Porta, Jordi. 2014. Twitter language identification using rational kernels and its potential application to sociolinguistics. In *TweetLID@SEPLN*.

Shuyo, Nakatani. 2010. Language detection library for java.

Singh, Anil Kumar. 2006. Study of some distance measures for language and encoding identification. In *Workshop on Linguistic Distances*, pages 63–72. ACL.

Singh, Anil Kumar and Pratya Goyal. 2014. A language identification method applied to twitter data. In *TweetLID@SEPLN*.

Zamora, Juglar Díaz, Adrian Fonseca Bruzón, and Reynier Ortega Bueno. 2014. Tweets language identification using feature weighting. In *TweetLID@SEPLN*.

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