Automatic Detection of Regional Words for Pan-Hispanic Spanish on Twitter *

Sergio Jimenez[1], George Dueñas[1], Alexander Gelbukh[2], Carlos A. Rodriguez-Diaz[1], and Sergio Mancera[2,3]

[1] Instituto Caro y Cuervo, Bogotá D.C., Colombia
{sergio.jimenez|george.duenas|carlos.rodriguez}@caroycuervo.gov.co
http://www.caroycuervo.gov.co
[2] Centro de Investigación en Computación, Instituto Politécnico Nacional, México
[2]gelbukh@gelbukh.com, [3]sergiomanceranom@gmail.com
http://www.cic.ipn.mx/

Abstract. Languages, such as Spanish, spoken by hundreds of millions of people in large geographic areas are subject to a high degree of regional variation. Regional words are frequently used in informal contexts, but their meaning is shared only by a relatively small group of people. Dealing with these regionalisms is a challenge for most applications in the field of Natural Language Processing. We propose a novel method to identify regional words and provide their meaning based on a large corpus of geolocated 'tweets'. The method combines the notions of specificity (tf-idf), space correlation (HSIC) and neural word embedding (word2vec) to produce a list of words ranked by their degree of regionalism along with their meaning represented by a set of words semantically related and examples of use. The method was evaluated against lists of regional words taken from regional dictionaries produced by lexicographers and from collaborative websites where users contribute freely with regional words. We tested the effectiveness of the proposed method and produced a new resource for 21 Spanish-speaking countries composed of 5,000 regional words per country along with similar words and example 'tweets'.

Keywords: Spanish regionalisms · automatic regional words detection · regionalisms meaning · HSIC · TF-IDF · Word2vec

1 Introduction

Most NLP applications for a particular language assume that the language is homogeneous in the geographical territory where it is used. This situation worsens when that language, like Spanish, is spoken in a very extensive and diverse territory. This situation is probably due to the lack of an updated, representative and machine-readable source of linguistic differences in the language's territory. In corpus linguistics, a regionalism can be considered as a word that meets two

* Supported by Asociación de Amigos del Instituto Caro y Cuervo. S. Mancera was supported by a scholarship given by CONACYT, Mexico.

conditions: comparatively, it is more used in a region than in the rest of the territory where the same language is spoken (specificity) and its frequency of use in different places is related to a continuous geographical distribution (spatial correlation). The automatic detection of such regionalisms and their meanings could improve the performance of other NLP tasks, such as word sense disambiguation, language understanding, automatic translation, among others. We can expect that improvement at least in these scenarios: a) when a word has different meanings according to the region where it is used[3] ,b) when a word refers to concepts that varies from one region to another[4], and c) when a word was created in a specific region to define a local concept that is unknown in the rest of the language's territory.

As an antecedent to this problem, Calvo (2014) [2] implemented the concept of specificity using the measure *tf-idf* to classify regionalisms in Spanish. He used the snippets returned by the Google search results to expand an initial list of regionalisms using the country code in the URL to determine the country of origin of the text snippets. This approach has two drawbacks. First, the country code in the URL does not guarantee that the text was produced in that country. Second, the geographical distribution of the countries was ignored. This work aims to remedy these problems by collecting a geolocated Twitter corpus, following the trend of current practices in the field of dialectometry [7,3,11]. Unlike the Google search results used by Calvo, Twitter data can be considered as an information source updated, representative, close to oral language, geographically accurate, and large enough to detect regionalisms.

The proposed approach extends Calvo's work by combining the measure *tf-idf* with the HSIC spatial autocorrelation test [6,10] avoiding the need for an initial list of regionalisms. In addition to the detection of regional words, we provide their meanings by associating them with other regional and non-regional words, and also providing a set of sample tweets for each regionalism found. The proposed method was evaluated with different dictionaries of regionalisms and lists of regional words compiled collaboratively in `asihablamos.com` and `diccionariolibre.com`. The proposed method produced a new resource that includes the regional words detected together with their semantically closed words, and examples of use, for 21 countries in the Pan-Hispanic world.

The rest of the paper is organized as follows. In Section 2, we present the concepts used to construct the proposed approach. Section 3 provides a detailed description of the methods and data used. In Section 4, we present the experimental validation. Finally, in Section 5 and Section 6, the results are discussed and some concluding remarks are provided.

[3] For example, the common word for bus in peruvian spanish is *micro*, in colombian spanish *bus* for urban ones and *flota* for long distances ones, argentinian spanish uses *bondi* or *colectivo*, etc...

[4] For example, in some regions of Colombia, the word *galleria* refers to a marketplace, but in general Spanish, that word means an art gallery or a covered path.

2 Background

2.1 Inverse Document Frequency, IDF

IDF is a measure of the specificity of a word in a document collection [13]. In this way, a word that appears in all the documents obtains the minimum value of the measure (that is, zero). In contrast, a word that appears in a single document among n gets the maximum (that is, $\log n$). In our scenario, a document means the corpus compiled for each geographic location. Thus, the IDF of a word w is defined by:

$$IDF(L_w) = \log \tfrac{n}{df(L_w)}$$

Where, $L_w = \{l_1, \cdots l_n\}$ is the list of frequencies of w in the n locations and $df(L_w)$ is the number of locations where w occurred. In an attempt to make the measure robust against noise and given that the number of words in the corpus of each location is considerable, we consider that w occurred in the i-th location only if $l_i > 3$.

2.2 Hilbert-Schmidt Independence Criterion, HSIC

HSIC is a measure capable of evaluating the correlation between two multidimensional and non-linear variables [4]. These non-linearities of the variables are admissible because HSIC is a kernel-based method [6]. Since the representation of each variable in the kernel methods is a square matrix that contains pairwise distances between samples (the Gram matrix), the dimensionality and the original coordinates of the variables become irrelevant. Support for nonlinearities is provided by the so-called “kernel trick”, which projects the data from the original input space to a feature space with a greater dimensionality by applying a simple transformation to the inputs of the Gram matrix (for example, raising its inputs to a power). Once the two input variables to be tested are converted into Gram matrices and projected to a feature space, the covariance between the two resulting matrices is the value of the HSIC measure.

Recently, Nguyen and Eisenstein [10] used HSIC to measure the auto-spatial correlation of the geographical coordinates of a set of locations and a linguistic variable associated with each location (for example, the frequency of a word). HSIC proved to be a better alternative for this task compared to traditional approaches such as Moran’s I [5], join count analysis [8], and the Mantel Test [12]. A linguistic variable that obtains a high value of HSIC with a set of geographic locations means that it exhibits a regional pattern. It also means that the linguistic variable is a good predictor of the geographical location, that is, a regional word. In practice, the measurement of HSIC between a geographic variable G ($g_1 \dots g_n$ longitude-latitude pairs) and a linguistic variable L ($l_1 \dots l_n$ word frequencies paired to the g_i coordinates) is calculated by:

$$\mathrm{HSIC}(G, L) = \tfrac{\mathbf{tr}(K_G \times H \times K_L \times H)}{n^2}$$

Where K_G and K_L are, respectively, the Gram matrices for G and L. H is a centering matrix defined by $H = \mathbb{I}_n - \frac{1}{n}\mathbf{11}_n$, where $\mathbb{I}_n$ is the identity matrix and $\mathbf{11}_n$ is a matrix filled with ones, both of dimensions $n \times n$. Finally, $\mathbf{tr}(\cdot)$ is the trace of the resulting matrix, that is. the sum of the elements in the diagonal (covariance). The Gram matrix K_G is obtained by projecting the pairwise Euclidean distances, $dist(\cdot,\cdot)$, between the n locations with a Gaussian transformation. K_L is obtained analogously. The expressions are:

$$K_G(g_i, g_j) = e^{-\gamma_g dist(g_i,g_j)^2};\ K_L(l_i, l_j) = e^{-\gamma_l (l_i - l_j)^2};\ i, j \in 1 \ldots n.$$

In essence, HSIC is a nonparametric test that does not require assumptions about the data. The only parameter is γ_g that can be determined heuristically by the median of the squared pairwise distances, $dist(g_i, g_j)^2$. Similarly, γ_l is the median of the squared differences $(l_i - l_j)^2$. The schematic process for calculating HSIC is depicted in Figure 1.

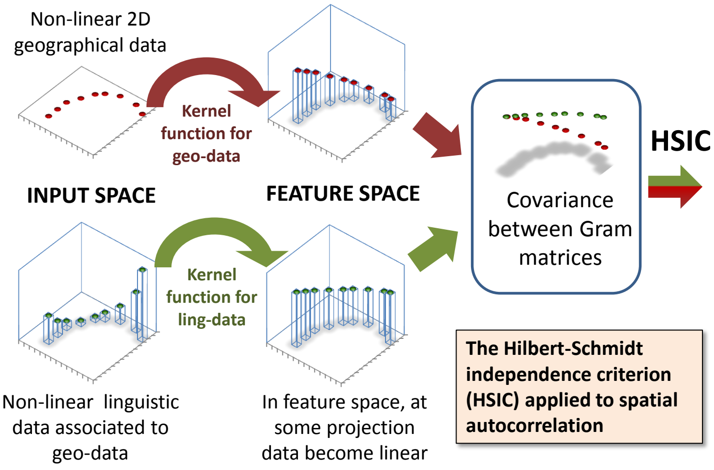
Non-linear 2D geographical data
Kernel function for geo-data
INPUT SPACE
FEATURE SPACE
Kernel function for ling-data
Non-linear linguistic data associated to geo-data
In feature space, at some projection data become linear
Covariance between Gram matrices
HSIC
The Hilbert-Schmidt independence criterion (HSIC) applied to spatial autocorrelation

Fig. 1. The Hilbert-Schmidt independence criterion applied to spatial autocorrelation with frequencies of words obtained from a corpus.

2.3 Word2vec Word Embedding

Word2vec is a popular method based on a neural network to obtain a geometric model for the meaning of words from a large corpus [9]. In that model, words are represented as points in a high-dimensional space, usually from 100 to 1000 dimensions. There, the distances between pairs of words reflect their semantic similarity and those distances combined with the direction of the differences reflect semantic relations. Another property, which is still not fully understood, is that the relative positions of the words represent semantic relationships that give the model the ability to make compositional and analogical reasoning.

3 Data and Proposed Method

3.1 Data

The data for this study was collected semi-automatically from the web search interface of Twitter respectfully of their access quotas. We selected 333 cities with more than 100,000 inhabitants in the Pan-Hispanic world. In those locations, a query by the official geographical coordinates of each city specifying a 15 miles radius and Spanish language. When two cities overlapped their query areas, the small one was discarded. The collected tweets were preprocessed by removing URLs, hashtags, references to user names and non-alphabetical words. The size and other features of collected corpus are presented in Table 1. From that corpus we obtained a large database with the number of occurrences of each word for each city. Although by Twitter policies, the corpus can not be published publicly, we made available the word frequencies database[5]. In addition, for the analysis of regionalisms, we ignored any word containing three or more consecutive repeated letters (for example 'holaaaa'), words that appear predominantly with initial capital letters (proper names), and sequences that represent laughter in Spanish (for example 'jajajajjaa').

Table 1. Statistics of the Spanish corpus collected from Twitter

Country	ISO	Cities	Words	Tweets	Vocabulary	Users
Argentina	ar	26	254,982,258	26,933,107	5,264,160	859,197
Bolivia	bo	8	3,136,167	289,683	206,944	24,508
Chile	cl	24	155,791,513	15,291,490	3,679,096	599,059
Colombia	co	31	209,085,865	19,875,419	4,575,636	871,247
Costa Rica	cr	5	43,905,034	4,272,517	674,130	97,211
Cuba	cu	1	122,595	13,246	14,044	5,354
Ecuador	ec	10	49,016,999	4,483,875	1,257,676	197,544
El Salvador	sv	3	19,898,193	1,835,850	453,030	65,543
Guatemala	gt	7	31,753,056	3,131,936	827,927	147,460
Honduras	hn	7	18,282,159	1,710,399	579,025	65,786
Mexico	mx	74	453,724,537	43,544,549	10,187,200	1,983,207
Nicaragua	ni	4	10,982,904	1,222,135	321,567	25,862
Panama	pa	5	33,237,123	3,078,389	855,235	114,062
Paraguay	py	6	39,753,880	3,968,928	765,886	113,243
Peru	pe	14	35,355,182	3,329,937	973,957	181,880
Puerto Rico	pr	3	35,230,113	3,863,552	666,343	94,589
Dominican Rep.	do	5	86,657,210	8,608,484	1,603,572	245,348
Spain	es	36	499,630,471	45,276,446	10,771,631	1,646,083
USA	us	35	59,974,018	6,172,521	2,759,849	956,255
Uruguay	uy	7	37,121,241	4,252,022	896,557	102,350
Venezuela	ve	22	194,073,318	16,773,933	4,343,584	764,215
TOTAL	21	333	2,271,713,836	217,928,418	51,677,049	9,160,003

[5] `https://www.datos.gov.co/browse?q=F-TWITTER`

3.2 Rank Functions for Lexical Regionalism

The IDF and HSIC functions presented in Section 2 provide a general measure of the degree of specificity and regionalism of each word in the corpus. The IDF is usually combined with TF, the term frequency in a document (the corpus of a city in our scenario), producing the well-known term weighting scheme TF.IDF. Let $TF(w, i)$ be the number of times the word w occurred at the i-th location. In this way, the multiplicative combination of TF and IDF produces a measure for each word in each city that yields high scores only when a word is frequent in a city and is used in a few cities. As Calvo [2] observed in an intuitive and empirical way, this property coincided with the notion of a regional word. However, IDF can not discriminate regional words in some cases, depending on the geographical distribution of occurrences. For example, a word that appears in half of cities can be considered regional only if those cities are grouped in a region. In the event that these cities were sparsely distributed throughout the geographical area, the word could not be considered a regionalism. In contrast, HSIC can effectively discriminate these geographic patterns. To exemplify HSIC, Table 2 shows the words with the highest HSIC values in the corpus. The majority of these words are Mexican regionalisms that occur almost in all 74 Mexican cities among all 333 cities in the corpus (i.e. low specificity). However, HSIC does not identify regionalisms with maximum specificity. That is, when a word occurs only in a city, the HSIC measure gets its minimum score.

Given that neither IDF nor HSIC seem to adequately model our notion of regionalism based on specificity and geographic association, we propose several multiplicative combinations of the TF, IDF and HSIC factors. A fourth factor identified as HSIC1 is equivalent to HSIC but it filters small values of the measure, which could be produced by the effect of randomness. In our experiments, we observed that a convenient value for the filtering threshold is $\theta = 0.009$. Therefore, the four classification functions used to determine the degree of regionalism of a word w in the i-th location are:

$$\mathsf{TF.HSIC}(w, i) = TF(w, i) \times (HSIC(G, L_w) + 1)$$
$$\mathsf{TF.IDF}(w, i) = TF(w, i) \times IDF(L_w)$$
$$\mathsf{TF.IDF.HSIC}(w, i) = TF(w, i) \times IDF(L_w) \times (HSIC(G, L_w) + 1)$$
$$\mathsf{TF.IDF.HSIC1}(w, i) = TF(w, i) \times IDF(L_w) \times (HSIC1(G, L_w) + 1)$$

Table 2. The 24 words with the highest scores of the HSIC measure for the corpus.

word	HSIC	word	HSIC	word	HSIC	word	HSIC
mexicanos	0.0464	chingar	0.0426	cabrona	0.0417	pelu	0.0404
tamales	0.0451	mexico	0.0426	pinche	0.0417	chivas	0.0403
frijoles	0.0450	chicharito	0.0425	mero	0.0416	yolo	0.0401
cabron	0.0448	orale	0.0419	culero	0.0408	impresentable	0.0400
cabrones	0.0440	fam	0.0419	chingaderas	0.0406	tortillas	0.0399
corridos	0.0430	chingada	0.0418	chicharo	0.0404	azteca	0.0397

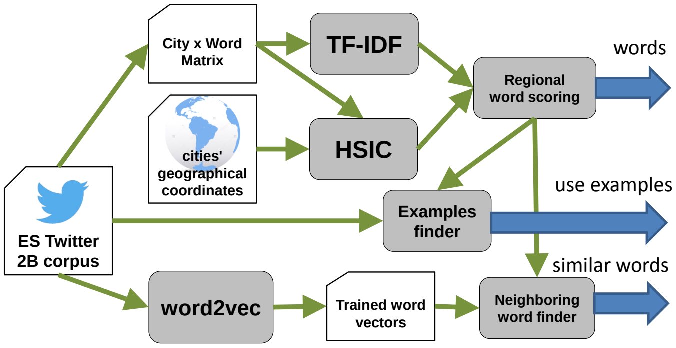

Fig. 2. Process for regional word detection, exemplification and meaning determination from a large corpus

3.3 Determining the Meaning of Regional Words

Once the k words with the highest regional score for each location are determined using one of the proposed ranking functions, the meaning of these words must be determined to make that list useful. For that, we provided two mechanisms to determine the meaning of the regionalisms. First, we trained a *word2vec*[6] model with the corpus using the following parameters: CBOW algorithm, 100 dimensions, window size of 5 words, and a learning rate of 0.025. Next, we obtained the nearest c neighbors to each regionalism. Second, we look for tweets where each regional word is used in context. Then, we calculate the average regionalism score for each example tweet and report the best t tweets with the lowest regionalism score for each regional word. In this way, the selected tweets illustrate the regional word surrounded by non-regional words thus facilitating the inference of its meaning. Finally, the 333 cities were added in their 21 corresponding countries and the lists were produced by establishing $k = 5,000$, $c = 30$, and $t = 30$. We published the 21 list of regional words with their nearest neighbors[7] and with their example tweets[8]. Table 3 contains a small sample of the created resource.
Figure 2 shows a summary in a block diagram of the architecture of the proposed method.

4 Experimental Validation

The experimental validation proposed in this section is aimed to determine to what extent the proposed ranking functions for the detection of regionalisms coincide with the notion and definition of "regional word" given by the Spanish speakers and by professional lexicographers.

[6] https://www.datos.gov.co/browse?q=word2vec
[7] https://www.datos.gov.co/browse?q=regionalismos%20cercanas
[8] https://www.datos.gov.co/browse?q=regionalismos%20ejemplos

Table 3. Examples for four countries of their top-regional words found (in bold face) along with their closest neighboring words in meaning and a sample tweet.

ARGENTINA	
lpm	lrpm, lcdsm, lptm, lpmmm, lcdll, lpmm, ptm, lpmmmm, laputamadre, jodeeeer,csm, lpmqlp, jodeer, lpmqlp, puff, jodeeeer, lpmqlrp "*me duelen los cortes en la mano **lpm** no doy mas*"
pelotuda	boluda, pajera, pendeja, tarada, forra, boba, tonta, mogolica, gila, payasa, weona, retrasada, estupida, cabrona, maricona, conchuda "*Me tendrían que regalar un premio por ser tan **pelotuda***"
chabon	pibito, vato, pibe, bato, waso, chaval, muchacho, chamo, weon, wn, tipito, wey, pive, vaguito, maje, chaboncito, chavalo, chabón "*Ya me pone de mal humor este **chabon** loco ...*"
COLOMBIA	
hpta	hpt, hp, hijueputa, hijuemadre, ijueputa, hptaa, hptaaa, hijueputaa, hijodeputa, wn, csm, hijoeputa, hijueputaaa, weon, conchesumadre "*Por mas **hpta** que sea no lo voy a dejar.*"
vallenato	rap, folclore, ballenato, regueton, malianteo, reggae, folklor,flocklore, regaeton, mariachi, reggeaton, reggueton, reggeton, folklore, mayimbe "*Que vivan las mujeres hermosas que interpretan el **vallenato***”
chimba	gonorrea, chimbita, chimbaa, chimbaaaaa, chimbaa, chiva, depinga, pinga, chimbo, chingon, tuanis, bacan, chevere, guay, chido, bacano "*Que **chimba** es ir a la nevera y encontrar algo de comer.*"
nojoda	njd, njda, nojodas, nojodaaaa, nojodaa, nojodaaaaa, nojodaaa, coñoooo, coñooo, coñooooo, coñoooooo, vergacion, nojodaaaaaa "*Hoy tengo más ganas de beber que de vivir, **nojoda***"
MEXICO	
neta	enserio, vdd, encerio, esque, alchile, sinceramente, acho, env, puñeta, netaaa, pucha, verga, verdad, posta, marico, netaaaaa, netaaaa "*Ganamos y **neta** a como jugamos no merecíamos ganar*"
peda	borrachera, farra, pedota, fiesta, juerga, pedita, pisteada, pedocha, guarapeta, peduki, parranda, bebeta, fiestota, pedaaaa, verguera "*Un brindis por esos amigos que te cuidan en la **peda***"
hueva	weba, flojera, weva, pereza, paja, flojerita, ladilla, wueba, wueva, caligueva, arrechera, flogera, fiaca, webita, bronca, hueba, jartera "*A mi mamá y a mi nos dio **hueva** cocinar*"
PERU	
chamba	ofi, oficina, faena, peguita, uni, mudanza, facu, talacha, ofis, chambita, uní, biblio, vagancia, vacavión, pachanga, pelu, pegita, farra, maleta "*Un día más de **chamba** para cerrar una buena semana! xD*"
csm	ctm, ptm, conchesumadre, conchesumare, conchasumadre, hpta, csmmmm, csmr, cdsm, jueputa, ptmr, hpt, conchetumare, hp, ptmre "*Quiero llegar temprano y hay un trafico de la **csm***"
huevadas	webadas, babosadas, tonteras, giladas, pelotudeces, pavadas, chorradas, idioteces, pendejadas, wevadas, muladas, boludeces "*Ya, mejor me voy a dormir antes de seguir pensando **huevadas**.*"

4.1 Benchmarks

The benchmarks used for evaluation were two collaborative edited websites in which users contribute freely with regional words and expressions of their countries of origin, namely ‘AsiHablamos’ and ‘DiccionarioLibre’. A third benchmark (‘Diccionarios’) was built by merging sources such as the “Diccionario de Colombianismos” (2018) from the Instituto Caro y Cuervo, the “Diccionario breve de mexicanismos” de Guido Gómez de Silva (2001), and others. From all these sources we removed all multi-word expressions and definitions[9]. The number of words included on each benchmark for each country is reported in Table 4.

Table 4. Number of regional words on each one of the evaluation benchmarks for each country.

Benchmark	ar	bo	cl	co	cr	cu	ec	sv	gt	hn	mx	ni	pa	py	pe	pr	do	es	us	uy	ve	Total
AsiHablamos	309	27	150	226	106	12	98	51	77	47	245	40	75	36	38	102	69	109	94	-	173	2,084
DiccionarioLibre	1,321	860	320	1,042	86	94	158	70	105	80	-	77	909	33	1,219	339	2,667	1,777	-	124	1,853	13,134
Diccionarios	905	529	672	5,893	347	-	228	52	-	2,407	6,153	699	225	-	332	-	469	-	-	-	-	18,911

4.2 Evaluation Measures

The objective of the evaluation measures is to quantitatively assess the degree of agreement between a list of regionalisms obtained from one of the ranking functions proposed in subsection 3.2 and a benchmark list. For that, we used two popular measures of the *Information Retrieval* field [1], that is *Mean Average Precision* (MAP) and *Precision at 100* (P@100). P@100 measures the percentage of common words between a ranked list and a benchmark list in the first 100 positions. MAP measures the average of the P@n only for positions n in the ranked list that contains a word in the benchmark. Figure 3 illustrates four examples of calculation of these measures in our particular setting.

4.3 Experimental Setup

The procedure to obtain a list of words ranked by their degree of regionalism is as follows. First, we processed the complete corpus by collecting word occurrences for each of the 333 cities and the global word frequency. Then, the IDFs and HSICs scores were obtained for the 100,000 most frequent words in the corpus. Next, these scores were used to calculate each of the proposed ranking functions

[9] `https://github.com/sgjimenezv/spanish_regional_words_benchmark`

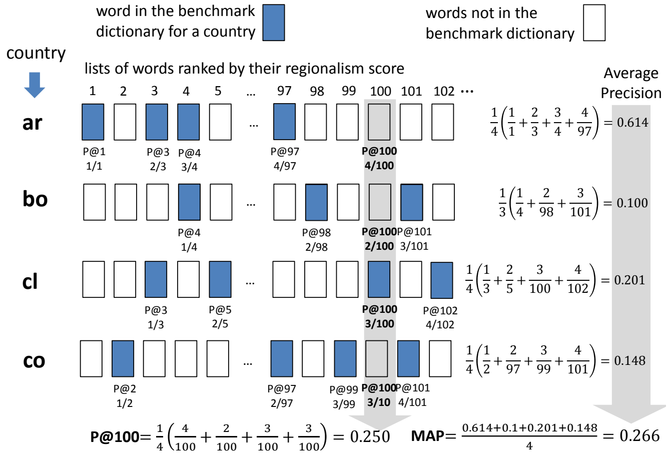
word in the benchmark dictionary for a country
words not in the benchmark dictionary
country
lists of words ranked by their regionalism score
1 2 3 4 5 ... 97 98 99 100 101 102 ...
Average Precision
ar
P@1 1/1
P@3 2/3
P@4 3/4
P@97 4/97
P@100 4/100
$\frac{1}{4}\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{4}+\frac{4}{97}\right)=0.614$
bo
P@4 1/4
P@98 2/98
P@100 2/100
P@101 3/101
$\frac{1}{3}\left(\frac{1}{4}+\frac{2}{98}+\frac{3}{101}\right)=0.100$
cl
P@3 1/3
P@5 2/5
P@100 3/100
P@102 4/102
$\frac{1}{4}\left(\frac{1}{3}+\frac{2}{5}+\frac{3}{100}+\frac{4}{102}\right)=0.201$
co
P@2 1/2
P@97 2/97
P@99 3/99
P@100 3/10
P@101 4/101
$\frac{1}{4}\left(\frac{1}{2}+\frac{2}{97}+\frac{3}{99}+\frac{4}{101}\right)=0.148$
P@100= $\frac{1}{4}\left(\frac{4}{100}+\frac{2}{100}+\frac{3}{100}+\frac{3}{100}\right)=0.250$
MAP= $\frac{0.614+0.1+0.201+0.148}{4}=0.266$

Fig. 3. Calculation examples of the evaluation measures MAP and P@100.

for each city (see subsection 3.2). Then, the rankings were merged to produce a ranking for each one of the 21 countries and for each function taking the top-5,000 words. Finally, the country rankings were compared with the three benchmarks by measuring MAP and P@100 for each possible combination of ranking function, country and benchmark.

4.4 Results

Figure 4 shows the results of the averages obtained in all countries for each benchmark for MAP and P@100 measures. That figure clearly shows that the TF.HSIC measure performed considerably worse than the other measures. The second observation is that, both in MAP and in P@100, TF.IDF performed practically identically as TF.IDF.HSIC. However, there is a difference in the performance between TF.IDF and TF.IDF.HSIC1. The average for the two measures throughout the 52 evaluations (that is, 20 countries for AsiHablamos, 19 for DiccionarioLibre, and 13 for Diccionarios) reveals a difference of 7.46% in MAP and of 1.36% in P@100. To evaluate the statistical significance of these differences, we used the Wilcoxon signed rank test obtaining $p = 0.0025$ for MAP (highly significant) and $p = 0.5656$ for P@10 (non-significant).

5 Discussion

The results clearly indicate that spatial correlation is a weaker signal for the detection of regionalisms compared to specificity. In addition, the HSIC measure does not seem to contribute in the top-100 regionalisms compared to any of the three benchmarks. However, the HSIC1 factor managed to improve the results for the MAP measure. This lead us to the conclusion that the HSIC scores lower

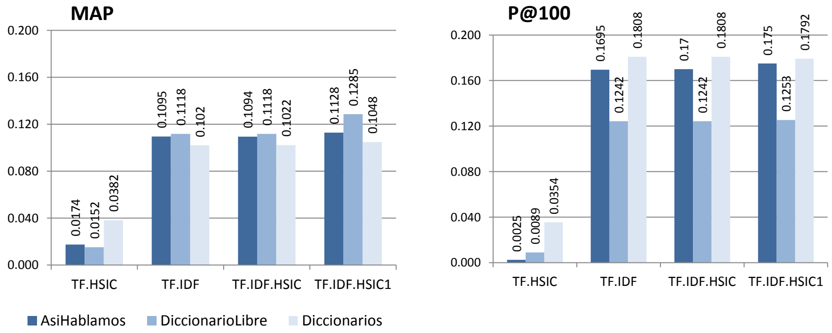

Fig. 4. Results of the agreement between the four proposed functions of ranking of regionalisms with the three benchmark lists.

than the threshold established in HSIC1 ($\theta = 0.009$) seems to be noisy producing a decrease in performance equivalent to the benefit of using the scores above that threshold, yielding zero effect consolidation when comparing TF.IDF versus TF.IDF.HSIC. In fact, the value for the threshold θ was obtained by optimizing the MAP measurement. In addition, the results reveal that the HSIC1 factor benefits mainly in positions beyond the word 100th in the ranked list of regionalisms. Given this result, the 21 datasets were obtained using the TF-IDF.HSIC1 ranking function. Regarding the performance differences between benchmarks, it is clear the variations are considerable for the leadership positions in the ranking (P@100), while the performance becomes a tie when the full ranking is evaluated (MAP).

Being the authors of this paper native speakers of Spanish, we manually evaluated the first 100 regionalism produced for our country of origin, Colombia. In that list we recognized 71 regional words associated to global concepts, 13 names of regional entities, 3 names for local concepts, 11 standard Spanish words with a noticeable increase in use in our country, and 2 errors. This result contrasts with the fact that, on average, only 16 of every 100 words in the top positions of the ranking were also included in one of the benchmark lists. This comparison reveals that the proposed method is effective for the identification of regionalisms and that the benchmarks obtained from compilations made by speakers or professional lexicographers have a very low coverage of the real regional patterns of the Spanish language.

6 Conclusion

A corpus-based and language-independent method was proposed to build a new resource containing the most representative regional words and their meanings for 21 countries in the Pan-Hispanic world. This resource has the potential to benefit NLP applications that deal with utterances produced in informal environments, where the use of regional words is frequent. The constructed resource was evaluated in comparison with a benchmark composed of contributions of speakers and professional lexicographers. This evaluation leads us to conclude that, for the detection of regional words, the specificity of a word in a corpus (measured by tf-idf) is a stronger signal than the geographic correlation of its use (measured by HSIC). However, the combination of tf-idf and HSIC produce the best results. In addition, a manual inspection of the results for a country showed that the proposed benchmarks suffer from lack of representativeness and that the list produced by our methods reflects the regional jargon quite well. As future work, we hope to extend this work by addressing the more challenging task of identifying regional expressions of multiple words and their meanings.

References

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
2. Calvo, H.: Simple tf· idf is not the best you can get for regionalism classification. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 92–101. Springer (2014)
3. Donoso, G., Sanchez, D.: Dialectometric analysis of language variation in twitter. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). pp. 16–25. Association for Computational Linguistics, Valencia, Spain (April 2017)
4. Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.J.: A kernel statistical test of independence. In: Advances in neural information processing systems. pp. 585–592 (2008)
5. Grieve, J., Speelman, D., Geeraerts, D.: A statistical method for the identification and aggregation of regional linguistic variation. Language Variation and Change **23**(2), 193–221 (2011)
6. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. The annals of statistics pp. 1171–1220 (2008)
7. Huang, Y., Guo, D., Kasakoff, A., Grieve, J.: Understanding us regional linguistic variation with twitter data analysis. Computers, Environment and Urban Systems **59**, 244–255 (2016)
8. Lee, J., Kretzschmar Jr, W.A.: Spatial analysis of linguistic data with gis functions. International Journal of Geographical Information Science **7**(6), 541–560 (1993)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
10. Nguyen, D., Eisenstein, J.: A kernel independence test for geographical language variation. Computational Linguistics **43**(3), 567–592 (2017)
11. Rodriguez-Diaz, C.A., Jimenez, S., Dueñas, G., Bonilla, J.E., Gelbukh, A.: Dialectones: Finding statistically significant dialectal boundaries using twitter data. In: International Conference on Intelligent Text Processing and Computational Linguistics (in press). Springer (2018)
12. Scherrer, Y.: Recovering dialect geography from an unaligned comparable corpus. In: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. pp. 63–71. Association for Computational Linguistics (2012)
13. Spärck Jones, K.: Idf term weighting and ir research lessons. Journal of documentation **60**(5), 521–523 (2004)