Application of Hilbert-Schmidt Independence Criterion to Lexical Geographical Variation in Lyon, France

Taha Merghani
Georgia Institute of Technology
taha@gatech.edu

June 24, 2025

Abstract

With the advent of social media platforms, language variation across geographical and demographic factors has been the subject of several data driven dialectal and sociolinguistic studies. The first step is typically identifying potential geospatially dependent linguistic variables. Most of the previous works, however, relied on the scholar's intuition about those variables; were sensitive to the type of linguistic variables (binary, frequency or categorical); made parametric assumptions about the data; and relied on predefined discretized geographical bins. To overcome these limitations, (Nguyen and Eisenstein, 2017) introduced a consistent nonparametric independence test based on Reproducing Kernel Hilbert Space (RKHS) representations which applies to both linguistic and geographical data. While the test (Hilbert Schmidt Independence Criterion, HSIC henceforth) has been shown to be applicable to determining geolinguistic dependence in geotagged corpra, it has not been well studied as an exploratory tool for mining dialectal variables within cities. In this paper, we describe our attempts to mine geographically dependent words in a geotagged corpus of tweets in Lyon, France using the HSIC test, retrieving a number of words that are more meaningfully associated within Lyon than chance.

1 Introduction

The massive amounts of social media data invite statistical approaches to analyzing socio-linguistic and geographical variation [EOSX14]. A common approach is to first identify the candidate linguistic *variables* for geographical variation (e.g., *soda*, *pop*, or *coke* to refer to a soft drink) and identifying geographical units of variation (municipalities, states, etc.) and applying classical statistics methods such as Moran's I [GSG11], Join Count Analysis [LKJ93], and the Mantel test [Sch12] . Other computational approaches build probabilistic models over predefined geographical bins [HJS15]. While the latter approaches provide insight about the interaction between geography and language, they do not directly measure that dependence.

Recently, Nguyen and Eisenstein[NE17] proposed HSIC as a broadly applicable and consistent test statistic to directly quantify the extent of the dependence between linguistic variables and geography. Here, we use HSIC as an exploratory tool to retrieve geographically dependent words in a corpus of geotagged tweets in Lyon. We describe how the method is applied on the twitter corpus and discuss the top 40 retrieved geographically dependent words according to the method. We find that the retrieved top words do exhibit regional dependence.

2 Data Preprocessing

We use a corpus of 650,161 automatically geotagged tweets in Lyon, France from 4081 unique users. The tweets were tokenized using NLTK, and (@ mentions) were normalized. This resulted in a total of 418,516 unique tokens. To focus on tokens that are more likely to be dialectal, we consider words that appear on a minimum of 100 tweets, resulting in a total of 5008 candidate linguistic variables.

3 Hilbert-Schmidt Independence Criterion (HSIC)

The HSIC non-parametric test [GBSS05] works by approximating a measure of the discrepancy between the joint geolinguistic distribution P_{XY} and the product of independent distributions P_X P_Y. Since the forms of the distributions (P_X and P_Y) are unknown, the maximum mean discrepancy (MMD ; a scalar function of the distributions) could be used to estimate that discrepancy between the two distributions. When linguistic similarity tend to co-occur with geographical similarity, the MMD will be high. The key insight is that it is possible to approximate the MMD from a finite sample of observations. This could be could be written as a sum of appropriately chosen kernel similarity functions, reaches the exact MMD if given infinite data.

For further details, the reader is highly encouraged to refer to [NE17].

3.1 Application on Corpus

Approach Our goal is to use the HSIC value as a tool to discover words that are statistically dependent on certain geolocations within the city of Lyon. Once the top linguistic have been identified, it is possible to study the properties of the networks of the authors of these words to answer downstream sociolinguistic questions. HSIC [NE17] HSIC has been shown to be effective in testing geographical association of linguistic variables. However, it has not been extensively studied as a tool for retrieving geographically dependent words, especially within cities. To do this, we compute the HSIC values and their corresponding p values for each linguistic variable. We then consider the words with the significantly highest HSIC values ($p < 0.05$) to be words that are geographically dependent.

Let x_i represent a scalar linguistic observation for unit (or tweet) $i \in \{1, ...,n\}$ representing the presence of token x_i in vocabulary V in tweet i. If we have n observation in a geotagged corpus D = { (x_1, y_1), (x_2, y_2), ... , (x_n, y_n)}

Sampling process Since most lexical variables do not appear in most tweets as well as the computational cost of computing the HSIC value for each variable, we sample without replacement from the subset of tweets $S_i \subset D$ an equal number of tweets in which x_i appears and where it does not. We operationalize the maximum number of occurrences of x_i considered to be 3000. Hence, the maximum $\mid X \mid = \mid Y \mid = \mid S_i \mid = 6000$. Then we apply HSIC on (X, Y) with a Gaussian kernel on the geolocations Y and a Delta kernel on the linguistic variables X. Here, X is the a binary vector of a lexical variable appearances, and Y is the corresponding vector of latitude and longitude pairs of the sampled units S_i. The p value is then estimated from a number of bootstraps N in two 2 ways we describe below.

4 Results and Discussion

Below are the top 40 ranked words according to the method ranked first by HSIC values (from highest to lowest) and p values(from lowest to highest):

atp, brocante, iot , temp, federer, actuelle, /2, maxi, turf, 9/10, tudtr, [<3<3 EMOJI][1], *tpcconseil, turfistes, 8/10 nintendo, météo, ibn, =p, //t.co/wpc5dxlnon, quinte+, quinte, open, enghien, h/f, rmclive, more, reunions, découvrez, pou, pronos, wai, qi, wech, résultats, c3, c1, pronostic, islam*[2]*, sélection.*

Besides url and emoji preprocessing errors, the word *wech* is an internet slang for yes, and the other words in this list mostly include topical words that are meaningfully associatted with Lyon. The words *atp* (Association for Tennis Players), *federer* (tennis player), *open* (Lyon Open; a tennis tournament in Lyon 2017) are related to tennis events. Other words are related to horse racing, such as *quinte (top5 horses), quinte+ (a bet on the top 5 horses), pronosotic, turf. iot* refers to a major technology festival in Lyon[3].

[1] difficulties inserting the true emojis

[2] words borrowed from Arabic such as *ibn, tunisia, wlh, inchallah, and wallah* were all ranked within the top 200 ranked by the method

[3] Lyon is a major technology hub and words such as *data, entrepreneurship, cdi (job contract type), ste (company), adp (technology company), avk (technology company) all ranked within the top 70*

5 Acknowledgement

We would like to thank Georgia Tech's Computational Linguistics Lab for access to the data as well as Jacob Eisenstein, Sandeep Soni, and Brain Jany for helpful discussions.

References

[EOSX14] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114, 2014.

[GBSS05] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[GSG11] Jack Grieve, Dirk Speelman, and Dirk Geeraerts. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(2):193–221, 2011.

[HJS15] Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, pages 452–461. International World Wide Web Conferences Steering Committee, 2015.

[LKJ93] Jay Lee and William A Kretzschmar Jr. Spatial analysis of linguistic data with gis functions. *International Journal of Geographical Information Science*, 7(6):541–560, 1993.

[NE17] Dong Nguyen and Jacob Eisenstein. A kernel independence test for geographical language variation. *Computational Linguistics*, 43:567–592, 2017.

[Sch12] Yves Scherrer. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, EACL 2012, pages 63–71, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.