Automatic Discovery of Language Dialects via Explainable Machine Learning

Amir Feghahati
University of California, Riverside
Riverside, CA 92521
sfegh001@ucr.edu

Mike Izbicki
Claremont McKenna College
Claremont, CA 91711
mike@izbicki.me

1 Introduction

Languages are constantly evolving, and they evolve differently in different parts of the world. Consider the Spanish language. In Mexico, the Spanish word `computer` is `computadora`. Mexicans have a lot of contact with the United States, and so their word for `computer` was adopted from the English word. In Spain, however, the Spanish word for `computer` is `ordenador`. Spaniards have more contact with France than the United States, and so their word for `computer` was adopted from the French word `ordinateur`. Our goal is to map these geographic variations in language usage. Specifically, we want to answer the following two questions:

1. Which linguistic features characterize language dialects?
2. Which dialects are used in a given geographical area?

We propose to answer these questions by combining techniques from twitter geolocation and explainable machine learning. The UnicodeCNN (Izbicki et al., 2019b;a) is a state of the art model for tweet geolocation. It was trained on 900 million tweets written in more than 100 languages. This model has two advantages for our purposes: First, the UnicodeCNN generates features from the input text directly from Unicode code points. Because all languages can be represented in Unicode, the UnicodeCNN can generate features for all languages, and we will be able to identify dialects of all languages. Second, the UnicodeCNN works at the character level instead of the more popular word level. Therefore, we will be able to identify sub-word dialectical differences, which are characteristics of morphologically complex languages like Spanish and Arabic. Unfortunately, the UnicodeCNN is highly complex, and therefore it is not-obvious what properties of the text the model is using to perform geolocation.

2 Explaining a Tweet's Location

We use a sliding window strategy to identify which characters in the input text are most important for geolocation. The sliding window method was introduced by Zeiler and Fergus (2014) to explain image classification problems, and we adapt their method to explaining the UnicodeCNN text geolocation model.

The idea of the method is as follows. Given a text x with associated GPS coordinate y, we first calculate the loss $\ell(x, y)$. Then, for each character i in the text, we create a modified text x_i by deleting the ith character, and we calculate the loss $\ell(x_i, y)$. The *value* of character i is then defined to be

$$v_i = \ell(x_i, y) - \ell(x, y). \tag{1}$$

When character i is important to geolocation, then removing character i from the text will cause the loss to increase. Thus, v_i should be large for characters that are important. Sometimes, characters can be misleading, and v_i will be negative in these cases. Figure 1 shows example texts with the value of each character plotted.

SoCal NLP Symposium (2019), Los Angeles, CA, USA.

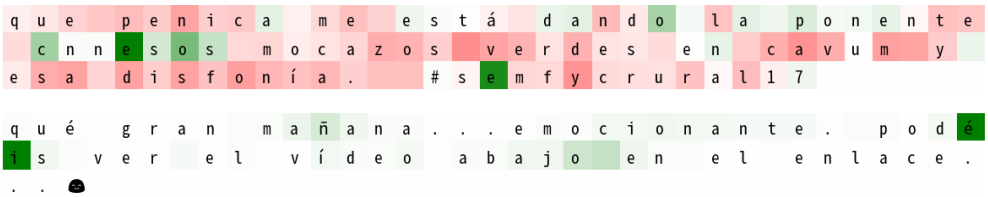

Figure 1: This figure shows two Spanish language tweets that our geolocation model correctly predicts as being sent from Spain. Letters with a positive value (v_i) are highlighted in green, and letters with a negative value are highlighted in red. The explanation of the top tweet is highly complex, with many letters highlighted in both red and green. This makes it difficult to understand exactly why the tweet was geolocated in Spain. The bottom tweet, in contrast, has a simple explanation with a small number of letters highlighted. These highlighted letters correspond to a grammatical construct called the `vosotros` verb conjugation that is only used in the Castillian Spanish dialect of Spain (and not used in Latin American Spanish). This bottom tweet's explanation therefore demonstrates that our geolocation model understands that the `vosotros` verb conjugations are associated with the Castillian Spanish dialect. Unfortunately, most tweets have complicated explanations like the top tweet, making it difficult to find tweets like the bottom tweet that concisely identify properties of different language dialects. The goal of this work is to find these tweets.

3 Finding the Best Explanations

We used two strategies for the sliding window: fixed size and variable size. For the fixed size strategy, we fixed the window size to 7 and compute maximum, average and maximum minus average of the losses for each window. For the variable size method, first, we compute the average loss over the tweet. Then, start of each window determined by a character which has a loss greater than average. We continue sliding over the characters till the loss of the character is less than average. We compute the average loss of this window as the output of that potion of text.

References

Izbicki, M., Papalexakis, V., and Tsotras, V. (2019a). Exploiting the earth's spherical geometry to geolocate images. *ECML-PKDD*.

Izbicki, M., Papalexakis, V., and Tsotras, V. (2019b). Geolocating tweets in any language at any location. *Under Review*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.