On Language Models for Places from User-Generated Content

VANESSA MURDOCK
Principal Applied Researcher
Microsoft
Email: vanmur@microsoft.com

We inhabit two worlds. We have offline interactions, which we have been developing and fine-tuning since the beginning of humanity, and we have online interactions, in which we employ our tremendous human capacity for social behavior. There are many points of overlap between the two spheres, such as friends we cultivate both online and offline, events that happen offline that we experience primarily online, or online interactions that have offline consequences. Location is one point of contact between the online and the offline experience.

Because GPS-enabled smart phones are widely used, we have an unprecedented amount of data about where people are, and what they were doing and thinking at the time. We can leverage this data directly for specific tasks, such as identifying the location a photo was taken, or more obliquely as a background model of a user’s geographic context, to be used in ranking, recommendation or prediction.

One approach to harnessing geo-tagged social media is to turn it into an information retrieval problem. We propose to segment the globe into “cells” and populate each cell with the social media artifacts emanating from that cell. We treat each cell as a “document” and estimate a term distribution from it. This document representation of places allows us to rank places given a query, or compute a similarity function between a short text and a place.

In our work we investigated several segmentation schemes (based on fixed grid cells, on zip code boundaries, and on dynamically-sized grid cells). We have investigated several approaches to estimating the term distribution associated with a place, and several different ranking functions. We discovered that if the boundary of a place is not important, dynamically sized grid cells produce the most accurate language models (Murdock, 2014). Regardless of the boundaries, if the term distributions are estimated from user-generated content, they are better estimated with the user frequency than the term frequency (O’Hare and Murdock, 2013). Dirichlet smoothing produces better results than Jelinek-Mercer smoothing for fixed-sized cells, because fixed-sized cells have a greater variance in the number of terms used to describe the place (O’Hare and Murdock, 2013). This does not hold for dynamically sized cells, because the cell size is adjusted according to the vocabulary so there is less variance in the document length (Murdock 2014).

---

1 An example would be experiencing a sporting, or political event through live Twitter feeds of people who are there.
2 An example is #gamergate in which female games developers were chased from their homes, and canceled speaking events because of online harassment. [Rawlinson, K. (2 September 2014) Gamers take a stand against misogyny after death threats. BBC News. Retrieved from http://www.bbc.com/news/technology/ November 2014.]
Each data source is useful, but they differ in the type of information they convey. For example, Flickr is especially useful for location modeling because the user is very often describing the location of the photo. A caveat of Flickr is that the photos are often associated with a particular user context, and the text associated with the photographs may not be meaningful outside of that context (van House, 2007). Users frequently apply the same, or nearly the same, tag set to large numbers of uploaded photos, although this is easily remedied by removing exact duplicates, and estimating term distributions with the user frequency, which counts one instance of a term per user per location, to avoid a single user dominating the text representation of a place.

Foursquare allows users to check in at a pre-existing place, or create a new place. They can also enter a short description or recommendation of a place (a “tip”) that other users can see. Foursquare focuses on business venues, although it is not limited to them.

While Flickr and Foursquare data link text to a place very directly, Twitter does so to a lesser degree. People often tweet about events or places around them, but they frequently tweet about topics unrelated to their current location. Mobile queries are also less often related to the location of the user. However, both Twitter and mobile queries allow researchers to infer the places people habitually visit, which is helpful for determining, among other things, which places are popular (Liu, 2013), and the user’s likely itinerary (de Choudhury et al. 2010).

A criticism of social media is that it has a low signal to noise ratio. For every gem in a Twitter feed, there are thousands of status updates that are not informative to users outside of the sender’s social context (Naaman et al. 2010). The sheer volume of user-generated content offsets this. For example, while in Twitter a very small percentage of tweets are associated with geographical coordinates, it is still a vast number of tweets. Flickr currently has hundreds of millions geo-tagged images. Foursquare users enable location by default, and so the majority of the Foursquare data is geo-tagged. Even queries from search engines on mobile devices are annotated with the coordinates of the users, in enormous volumes.

Social media and other user-generated content have been shown to be very useful for a variety of prediction, ranking and recommendation tasks. Among them, identifying the user’s location (Kinsella 2010 et al., Cheng et al. 2010), disambiguating points of interest (Rae et al. 2012), improving the ranking of search engine results (Bennett et al. 2011), inferring tourist itineraries (de Choudhury et al. 2010), and event recommendation (Quercia et al. 2010). It is a vast source of data, and growing. The challenge is to find creative ways to use it while respecting the user’s privacy.

References

1 Flickr, Twitter and Foursquare have public APIs granting limited access to their data.


