Chapter 12
Towards an Embedding-Based Approach for the Geolocation of Texts and Users on Social Networks

Sarra Hasni
LTSIRS Laboratory, National Engineering School, Tunis, Tunisia

ABSTRACT

The geolocation task of textual data shared on social networks like Twitter attracts a progressive attention. Since those data are supported by advanced geographic information systems for multipurpose spatial analysis, new trends to extend the paradigm of geolocated data become more emergent. Differently from statistical language models that are widely adopted in prior works, the authors propose a new approach that is adopted to the geolocation of both tweets and users through the application of embedding models. The authors boost the geolocation strategy with a sequential modelling using recurrent neural networks to delimit the importance of words in tweets with respect to contextual information. They evaluate the power of this strategy in order to determine locations of unstructured texts that reflect unlimited user's writing styles. Especially, the authors demonstrate that semantic proprieties and word forms can be effective to geolocate texts without specifying local words or topics' descriptions per region.

DOI: 10.4018/978-1-7998-1954-7.ch012

INTRODUCTION

Nowadays, a radical transformation in the paradigm of geospatial data is manifested through the evolution of related technologies. For example, the use of Geographic Information Systems (GIS) becomes wider by reinforcing their storage and management capacities. Such advantage makes the support of data from even unofficial sources more possible (Sui, 2011). Among those data, geotagged messages (tweets) that are published in the location-based social network (LBSN) Twitter constitute a considerable part of Big GeoData and proven to be useful for many purposes. For example, such messages report on human practices and daily lives which are in turn valuable for epidemiological monitoring (Allen, 2016), analysis of geolocated sentiments (Yaqot, 2018), prevention and resolution of crimes (Corso, 2017), etc.

Despite their effectiveness, the ability of geotagged tweets to bridge the gap between the physical world and the virtual one is still limited. In fact, previous studies demonstrate that the rate of geolocated tweets is less than 0.85%. This limitation promotes the development of several works for user/tweet geolocation based on textual content analysis in order to concretize the relationship between texts and space (Han, 2012). Through these works, a particular attention was accorded to statistical language models. For example, a given word may be a representative of a region if its use is more frequent compared to other words (Cheng, 2010). Otherwise, frequent terms and topics are marked by a set of relevant geospatial features making them useful to distinguish between different regions.

From a deep study of the proposed geolocation strategies, we think that the employment of statistical language models limits their performance. Precisely, we assume that the propagation of topics on social networks makes the estimation of their distribution more complex. For example, a user may report an event that occurs in a different location to that where it is located. A second limit that we consider the most critical is the rigidity of these models. They particularly seem to be less effective to treat new tweets which often contain out-of-vocabulary (OOV) words. Otherwise, the selective choice of local words limits their performance.

Given these problems, we consider that the geolocation task must be approached by paying more attention to word proprieties. In other words, we have to foster the relationship between inherent textual proprieties and geospatial dimensions. Particularly, we think that measuring the geo-semantic distribution may be efficient similarly to (Ballatore et al., 2013) and (Hu et al.,2017). Starting from the assumption that similar meanings occur in similar contexts, we consider that word embedding models can be effective solutions to measure the distribution of words / topics in space. Nevertheless, this utility is conditioned by their ability to determine the geographical belonging of a context that occurs in several regions at the same time. In addition, we enable our geolocation strategy to treat new tweets that contain OOVs