Information Processing and Management 54 (2018) 129–144

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Location extraction from tweets

Thi Bich Ngoc Hoang[a,b], Josiane Mothe[a,*]

[a] *Université de Toulouse and IRIT, UMR5505 CNRS, France*
[b] *University of Economics, the University of Danang, Vietnam*

ARTICLE INFO

Keywords:
Information systems
Social networks
Location extraction
Location prediction
Tweets analysis
Predictive model
Machine learning
Microblog collections

ABSTRACT

Five hundred million tweets are posted daily, making Twitter a major social media platform from which topical information on events can be extracted. These events are represented by three main dimensions: time, location and entity-related information. The focus of this paper is location, which is an essential dimension for geo-spatial applications, either when helping rescue operations during a disaster or when used for contextual recommendations. While the first type of application needs high recall, the second is more precision-oriented. This paper studies the recall/precision trade-off, combining different methods to extract locations. In the context of short posts, applying tools that have been developed for natural language is not sufficient given the nature of tweets which are generally too short to be linguistically correct. Also bearing in mind the high number of posts that need to be handled, we hypothesize that predicting whether a post contains a location or not could make the location extractors more focused and thus more effective. We introduce a model to predict whether a tweet contains a location or not and show that location prediction is a useful pre-processing step for location extraction. We define a number of new tweet features and we conduct an intensive evaluation. Our findings are that (1) combining existing location extraction tools is effective for precision-oriented or recall-oriented results, (2) enriching tweet representation is effective for predicting whether a tweet contains a location or not, (3) words appearing in a geography gazetteer and the occurrence of a preposition just before a proper noun are the two most important features for predicting the occurrence of a location in tweets, and (4) the accuracy of location extraction improves when it is possible to predict that there is a location in a tweet.

1. Introduction

The power of social networking is demonstrated in the vast number of worldwide social network users. According to Statista,[1] this number is expected to reach about 2.5 billion by 2018. Twitter, which enables users to create short, 140-character messages, is one of the leading social networks. The extensive use, speed and coverage of Twitter makes it a major source for detecting new events and gathering social information on events (Weng & Lee, 2011).

As set out in Message Understanding Conference (MUC) campaigns,[2] events have several dimensions that are equally important and require specific attention. The main dimensions are as follows:

- Location information which indicates where the event takes place;

* Corresponding author at: Université de Toulouse and IRIT, UMR5505 CNRS, France.
E-mail addresses: thi-bich-ngoc.hoang@irit.fr (T.B.N. Hoang), josiane.mothe@irit.fr (J. Mothe).

[1] http://www.statista.com/topics/1164/social-networks/.

[2] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/muc.htm/.

https://doi.org/10.1016/j.ipm.2017.11.001
Received 31 May 2017; Received in revised form 3 September 2017; Accepted 4 November 2017

- Temporal information which indicates when the event takes place;
- Entity-related information which indicates what the event is about or who its participants are.

This paper focuses on the location dimension. More specifically, it focuses on location extraction from tweets, which is vital to geo-spatial applications as well as applications linked with events (Goeuriot, Mothe, Mulhem, Murtagh, & Sanjuan, 2016). One of the first pieces of information transmitted to disaster support systems is where the disaster has occurred (Lingad, Karimi, & Yin, 2013). A location within the text of a crisis message makes the message more valuable than messages that do not contain a location (Munro, 2011). In addition, Twitter users are most likely to pass on tweets with location and situational updates, indicating that Twitter users themselves find location to be very important (Vieweg, Hughes, Starbird, & Palen, 2010).

Name entity recognition in formal texts such as news and long documents has attracted many researchers. However, very little work has been successfully carried out on microblogs. The Stanford NER (Named Entity Recognition)[3] (Finkel, Grenager, & Manning, 2005) achieves an 89% F-measure[4] for entity names on newswire, but only 49% for microblogs (Bontcheva et al., 2013). Similarly, the Gate NLP framework[5] (Bontcheva et al., 2013) achieves a 77% F-measure for long texts but only 60% for short texts. The Ritter tool[6] (Ritter, Clark, & Etzioni, 2011), which is considered to be the state of the art, only achieves a 75% F-measure for Twitter.

As mentioned in Bontcheva et al. (2013), each tool has its strengths and limitations. While the Gate NLP framework achieves high recall (83%) and low precision (47%), the Stanford NER achieves the opposite (recall 32%, precision 59%) for the development part of the Ritter dataset (Bontcheva et al., 2013).

Because there are applications that need high recall e.g. what has happened in a given location, and others that need high precision e.g. which locations should we concentrate on first for a given problem, we hypothesized that combining existing location extraction tools could improve the accuracy of location extraction. Moreover, we also hypothesized that filtering out the location using external resources could help the location extraction process. We thus derive our first research question:

RQ1: *How much can we improve precision and recall by combining existing tools to extract the location from microblog posts?*

To answer this question, we have combined various tools, namely, the Ritter tool (Ritter et al., 2011), the Gate NLP framework (Gate) (Bontcheva et al., 2013) and the Stanford NER (Finkel et al., 2005). We also propose to filter the extracted locations using DBpedia[7]. We have used it is as follows: the locations extracted by previous tools are only considered as locations if DBpedia considers them as locations (taking account of the DBpedia endpoint framework). We therefore targeted either recall-oriented or precision-oriented applications.

By associating locations that both Ritter and Gate recognize, we achieved 82% recall (for the Ritter dataset) which is very appropriate for recall-oriented applications. This result can be explained by the fact that these methods use different clues to extract locations from tweets. On the other hand, when using DBPedia to filter out locations that Ritter recognizes, we reached a remarkable precision of 97% (for the Ritter dataset). This high result was obtained because imprecise recognized location names were discarded.

As mentioned earlier, social network and microblogs are widely used media of communication. As a result, a huge number of posts and tweets are posted daily, but only a very small proportion contains locations. For instance, in the Ritter dataset (Ritter et al., 2011), which was collected during September 2010, only about 9% of the tweets contain a location. It is thus time consuming to try to extract locations from texts where no location occurs. If we could filter out tweets that do not contain locations, *prior* to extracting locations, then efficiency would be improved. This leads us to our second research question:

RQ2: *Is it possible to predict whether a tweet contains a location or not?*

We conducted a preliminary study by using location extraction tools only on tweets that contain locations; we achieved significantly higher accuracy than when implementing them on the entire datasets. This first result shows that if we could predict the fact that the text contains a location, it would be easier to extract this location.

One main contribution of this paper is that we define a number of new tweet features and use them as location predictors. Another contribution is that we evaluate the tweets using machine learning classifier algorithms with various parameters. In the experimental section, we show that the precision of NER tools for the tweets we predict to contain a location is significantly improved: from 85% to 96% for Ritter collection and from 80% to 89% for MSM2013 collection. This increase in precision is meaningful and crucial in systems where the location extraction needs to be very precise such as disaster supporting systems and rescues systems.

The rest of the paper is organized as follows: Section 2 presents the related work; Section 3 details the location extraction method we promote and its evaluation. In Section 4, we explain our original method to predict location occurrence in tweets and show its usefulness and effectiveness. Finally, Section 5 is the discussions and conclusion.

2. Related work

With the rising popularity of social media, many studies propose different ways to extract information from this resource. Previous similar studies can be grouped into two categories: location extraction and location prediction.

[3] http://nlp.stanford.edu/software/CRF-NER.shtml.

[4] F-measure is approximately the average (harmonic mean) of the precision and recall.

[5] https://gate.ac.uk/family/developer.html.

[6] https://github.com/aritter/twitter_nlp.

[7] http://dbpedia.org/snorql/ BDpedia structures the information from Wikipedia pages; it can be queried using SPARQL to extract structured information locally stored in DBpedia or through an endpoint framework.