CHAPTER 15

Extracting Location Information from Crowd-sourced Social Network Data

Pinar Karagoz*, Halit Oguztuzun, Ruket Cakici,
Ozer Ozdikis, Kezban Dilek Onal and Meryem Sagcan
Middle East Technical University, Computer Eng. Dept., Ankara, Turkey,
*karagoz@ceng.metu.edu.tr

Abstract

With millions of users worldwide, crowd-sourced social media data provide a valuable data source for events happening around the world. More specifically, microblogs, which are social networks that enforce short text messages, have a high popularity due to their availability as a mobile application and the practicality of short messages. Estimating the location of the events detected by following posts in microblogs have been the motivation of numerous recent studies. Extracting the location information and estimating the event location is a challenging task to maintain satisfactory situation awareness, especially for emergency cases such as fire or traffic accidents. Today, Twitter is among the most popular microblogging platforms, and there are recent research efforts aimed at detection of novel events online by following the Tweets. In order to analyze events, researchers generally focus on spatio-temporal features of the posts. Temporal features denote the time and ordering of posts, whereas spatial features are useful for location extraction or estimation. In this work, we present an overview on the process for toponym recognition and location estimation from microblogs.

How to cite this book chapter:
Karagoz, P, Oguztuzun, H, Cakici, R, Ozdikis, O, Onal, K D and Sagcan, M. 2016. Extracting Location Information from Crowd-sourced Social Network Data. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R (eds.) *European Handbook of Crowdsourced Geographic Information*, pp. 195–204. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.o. License: CC-BY 4.0.

Keywords

Location Estimation, Event Detection, Semantic Vector Expansion, Named-Entity Recognition (NER), Toponym Recognition

Introduction

Social media platforms constitute a rich and up-to-date information resource with their rapidly increasing number of users. With its popular features, including the text message limit and easy access on mobile environments, microblogging platforms in particular are intensively used by a large number of users. Among such platforms, Twitter appears as the most popular service with a high number of users. Since most of the users allow public access to posts and user profile information, it provides rich data for research areas including text analysis, text mining, information extraction or social analysis.

Event extraction and location estimation from crowd-sourced data in social networks are important for learning about the events happening. These techniques have applications in various domains involving being aware of what is happening in the vicinity and rapid access to emergency information. As a potential use case, for instance, on the basis of the Twitter posts of users witnessing a major accident in which a chain of cars are involved, it may be possible to estimate the event location, open alternative routes and provide first aid. Similarly, for management of disasters such as earthquakes or floods, accessing such information fast is valuable.

An important task to be fulfilled for location estimation is extracting evidence about location, especially recognizing location names, i.e. toponyms, within a social network message. This problem is a specific sub-problem under named entity recognition (NER) task, which is a well-known natural language processing (NLP) problem aimed at extracting certain types of entities including people, organizations, dates, locations etc. from text. NER techniques proposed in the literature mostly work on long and formal texts, such as newspaper articles. In long and formal texts, literature provides established NER solutions. On the other hand, short and informal texts obtained in microblogs pose important challenges. Short text limits the contextual information that can be obtained from the whole text, whereas informal language makes it difficult to recognize the words. In this paper, we particularly concentrate on toponym recognition in microblog messages under these challenges.

Once toponyms are recognized in microblog posts, they provide useful and rich input for location estimation tasks. However, this step includes several challenges as well. One important challenge is that, in addition to recognized toponyms, there may be other clues to the location of the event, such as the GPS annotation of the message provided by the mobile device. It is necessary to make use of these clues in a complementary manner. Another basic challenge is the existence of several contradictory toponyms or clues, such as having location names that point to different coordinates. It is necessary to devise a mechanism in order to weigh how well conflicting clues contribute to the location estimation.

It is important to note that, although the proposed techniques in the literature are generally demonstrated on Twitter posts, i.e. Tweets, they can be applied on other social media data, especially those having informal use of language.

In this work, we describe these steps in an overview. The organization of this paper is as follows. In the next section, toponym recognition in informal text is described. In Section 3, we describe the location estimation problem and the suggested solutions in the literature, and conclude the paper with a summary and overview in Section 4.

Toponym Recognition

As *toponym*, referring to location names, is a type of named entity, for toponym recognition, generally, techniques for Named Entity Recognition (NER) are used. However, the proposed solutions conventionally work on formal texts. Recently, with the increasing amount of potentially rich data from social networks, NER and toponym recognition in web resources has attracted attention. However, solutions on formal texts rely on very basic features such as capitalization of the first letter of a token, existence of an apostrophe character within the token or existence of the token in the gazetteer. Informal and non-standardized language use in social networks poses a challenge in applying the same techniques in this area. Therefore new NER and toponym recognition techniques are being proposed.

In the literature of toponym recognition, the proposed techniques can be categorized as:

- gazetteer-based,
- rule-based, and
- machine learning based approaches.

The first step in all these techniques is tokenizing the messages into words. In addition, *morphological analysis* and Part-of-Speech (POS) tagging may be employed as preprocessing steps if morphemes/suffixes and POS tags are utilized in the toponym recognition process. Morphological analysis enables decomposition of a word into its affixes and the stem. POS tag of a word and suffixes in a word are effective features used in NER systems (StanfordNLP), (Seker 2012). For example, for the word '*happily*', a morphological analyzer for English will show that the stem of the word is '*happy*' and having the suffix '*-ly*'. An English POS tagger will annotate this word as an *adverb*. For stemming and POS tagging, language specific morphological analysis tools are needed. For English, one of the well-known tools is the NLP library of Stanford NLP Group

(StanfordNLP). For example, for Turkish texts, the morphological analysis tool Zemberek (Zemberek 2015) is commonly used.

Another important preprocessing step is *normalization*. Due to the informal use of language, text may include spelling errors or unusual abbreviations. In normalization, such problems are fixed before applying the toponym recognition technique. In Turkish, one of the available tools for normalization is being developed by ITU NLP group (Eryigit 2014). This tool fixes some of the spelling errors and performs capitalization for some proper nouns. Current solutions for normalization mostly include rule-based corrections capturing previously known informal language patterns, such as repetition of characters for emphasizing emotion. For example, *'Soooooo coooooooool !!!!'* is such a message that can be frequently used in microblogs. Normalization process should be able to convert the words to *'So'* and *'cool'*.

Gazetteer-based Approach

In this approach, a predefined list of location names is used as the gazetteer. The recognition process basically relies on checking whether a given token is in the gazetteer. The content and the granularity of the list depend on the context of the toponym recognition application. For general-purpose solutions, the list may contain country, city, town or Point of Interest (POI) names. For a specific geographical region, this list may be more detailed, including hospitals, banks, pharmacies etc. depending on the context of the application. For general-purpose gazetteers, OpenStreetMap (OpenStreetMap 2015) and Wikipedia (Wikipedia 2015) are commonly used resources.

In this approach, toponym recognition is based on checking whether text includes any toponym from the gazetteer. To this end, very simply, each token in the text is looked up in the gazetteer. The ones that are found are marked as toponyms. For example, for the post '*Enjoying good weather in Bebek Park, in Istanbul*', with a general-purpose gazetteer that includes POIs as well as city names, it is possible to recognize the toponyms *Bebek Park* and *Istanbul*. On the other hand, with a more limited gazetteer of city names, only *Istanbul* will be recognized. In this approach, normalization is more crucial, since it is based on matching between the token at hand and the toponym in the gazetteer.

Rule-based Approach

In this approach, certain patterns are defined in the form of the rules in order to recognize toponyms. For instance, if the word *street* follows a token, the token is considered to be a toponym referring to a street name (such as '*Oak Street*'). Some other patterns rely on the morphology or the POS tag of the word. One basic pattern for English is that if a token is preceeded by a pronoun, it is likely to be a toponym (such as '*in Istanbul*'). However, such rules overestimate the toponyms, leading to false positive recognition for the phrases such as *'on the table'*. Therefore, they may have high recall performance, however precision performance, generally, is not high. A recent study conducted on Turkish Tweets has shown that POS tag/morphology based rules achieved 30% precision, whereas with the gazetteer-based approach 67% precision is obtained (Onal 2014).

Machine Learning based Approach

In the machine learning based approach, supervised learning is the most common sub-approach employed for toponym recognition. Given a set of texts in which toponyms are annotated, a toponym model is constructed. More specifically, Conditional Random Field (CRF) is the most commonly used supervised learning technique providing satisfactory recognition ratios for formal texts. The advantage of CRF is that it is possible to capture contextual information in the model, such as having words in the neighborhood of a toponym. For Turkish texts, in Seker (2012), a CRF-based NER solution is presented. However, the success of this approach on informal texts heavily relies on the normalization step.

In a more recent study (Sagcan 2014), a CRF-based solution that focuses on toponym recognition in microblogging messages is proposed. The main motivation behind the work is performing toponym recognition without using any gazetteer and with less preprocessing effort for normalization. The main architecture of this approach is given in Figure 1.

As seen in the figure, the architecture is composed of two parts. In the first part, data preparation is performed. The second part contains CRF-based learning modules.

In the data preparation phase, initially, conventional tokenization and morphological analysis operations are applied. Afterwards, two simple normalization steps are applied. In the first normalization step, repeated characters are eliminated. For instance, 'çoooook güzeel' (çok güzel (Turkish) = very good (English)), 'gooooool' (gol (Turkish) = goal (English)) are commonly seen misspellings in Turkish Tweets. Such character repetitions occur in all languages in Twitter to denote an emphasis or exaggeration of emotion. The second normalization step is for the cases in which the English alphabet is used instead of the Turkish alphabet characters. Most of the misspellings in Turkish Tweets originate from replacement of Turkish characters with diacritics *(ç, g, ı, ö, ş, ü)* with non-accentuated characters *(c, g, i, o, s, u)*. Such usage is also very common in other non-English microblogging posts. This step can be customized according to the language alphabet under interest. In the literature, the proposed normalization step involves intensive cleaning and pre-processing on the text. In Sagcan (2014), only two simple normalization steps are applied and the other informal language problems are expected to be resolved during the learning phase.

In the second part of the architecture, for CRF-based learning, an annotated training data set is prepared in order to be used for model construction. In this

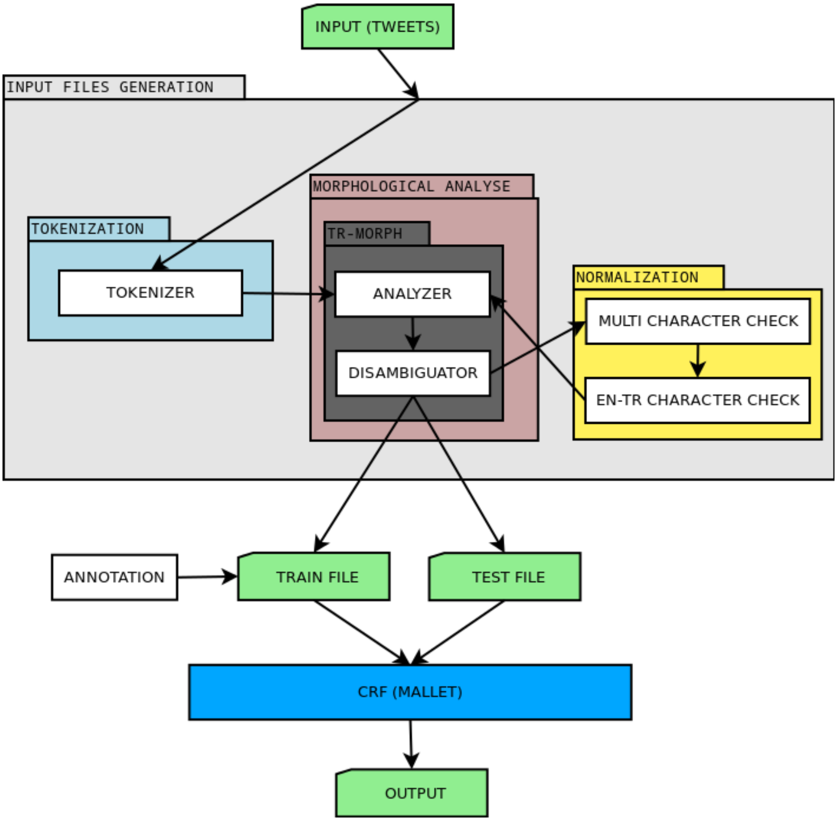

Figure 1: The architecture for CRF-based Toponym Recognition (Sagcan 2014).

step, the important issue is which attributes to use in the term vector representing the text message. Within the study, toponym recognition performance under several attribute compositions is analyzed. Under the best attribute composition, the precision performance is reported as 88%.

Location Estimation

Studies on location estimation are commonly related with event detection efforts. Recently, numerous studies have been conducted to detect real-world events by collecting public Tweets in Twitter. These include methods to identify various types of events including earthquakes (Sakaki 2010, 2013), disasters and crises (Yin 2012), and accidents (Rui 2012). The location estimation techniques proposed in these studies use different data and clues extracted from Tweets. In this section, we revise the most basic location estimation techniques according to the data used.

Geo-Annotation based Approach

In most of these event detection studies, GPS metadata of geo-tagged Tweets are the main resources for estimating the locations of events. A very simple procedure is followed. Once the location is estimated, location information about the detected events is presented on the map by plotting the geo-tagged Tweets about the event.

User Profile based Approach

In several other studies, the textual content in the location attribute of user profiles has been utilized in the place of GPS data for non-geo-tagged Tweets in, for example, Sakaki (2010, 2013) and Yin (2012). However, an in-depth study of user profiles show that since this attribute is a free-text field limited to 30 characters, it may contain multiple location names or even non-existing locations. According to the observation of Rui (2012), only 12% of users specified a location in their profile. Hence the authors tried to predict the user locations by analyzing their previous Tweets and locations of their friends.

Toponym based Approach

Due to the scarcity of geo-tagged Tweets and location information in user profiles, in Middleton et al. (2014), researchers focused on analyzing the Tweet content for location references by implementing a geo-parser. Location estimation studies concentrating on Tweet content make use of the toponym recognition techniques as discussed in Section 2. For example, in Sankaranarayanan et al. (2009), the authors used a gazetteer for toponym recognition and described a heuristic for toponym resolution in Tweets. Within the scope of event detection, some of the studies concentrate on estimating the location by using the whole cluster of Tweets that correspond to an event. The most frequently mentioned toponym in Tweet contents and user profiles is designated as the event location.

Evidence Combination based Approach

In Ozdikis (2013) a Bayesian approach is employed that uses evidences from several resources for location estimation. To this aim, three resources are used:

- Toponyms within the Tweet text
- GPS annotations on Tweets
- Toponyms in user profiles

As a basic difference from previous studies, for location estimation, Dempster-Shafer Theory (DST) (Dempster 1967), which can combine evidence from various resources in a single model, is used. DST is a generalization of Bayesian inference technique. DST is a technique under cases in which evidences are limited or they provide conflicting data. In addition, estimation result can be given in a belief interval such as [X%, Y%] instead of a single probability value.

In social networks, the number of postings sent from populated regions may distort the result, since the average number of messages posted from such locations are already higher than the other places and this leads to a bias towards such places. Hence, in Ozdikis et al. (2013) a normalization operation is devised and applied on evidence probabilities to prevent this bias.

As the evidence from the toponyms within the Tweet texts, the authors initially employed a gazetteer-based approach, however, other toponym recognition techniques such as machine learning-based approaches can be incorporated into the system.

In Ozdikis (2013), experiments are conducted on a set of Tweets posted about minor two earthquakes in Turkey, which occurred on May 17, 2013 near the city of Mugla, and on July 30, 2013 on an island near the city of Canakkale. They were not strong quakes, and did not cause any damage, but they were felt by people in these regions and triggered a reaction in Twitter traffic. For the first event, the belief interval found for Mugla is [0.77, 0.97], which can be considered as a confident decision. For the second event, the highest belief interval is found for Canakkale, as [0.24, 0.42]. The next highest belief intervals are [0.16, 0.34] for Edirne and [0.14, 0.31] for Tekirdag. Although Istanbul is a metropole from which many messages are posted, the belief interval as the location of the event is calculated only as [0.03, 0.21]. Hence, the results show the applicability of the method for location estimation.

Conclusion

Social networks, especially microblogs, contain valuable data including geographical footprints of users. A message becomes geo-annotated in terms of latitude and longitude if the user posts it using a GPS-enabled device and allows the sharing of this geographic information. In addition, users tell their location in their social network profile. Moreover, users may talk about places in their messages. Such attributes of Tweets are valuable resources for spatial analysis. There are various studies on location estimation that use these attributes. Most of them rely on a single attribute. However, techniques that combine several attributes in a single model for location estimation have more potential for accurate estimation.

Each of these attributes cause different challenges. In geo-annotated messages, GPS coordinates provide precise geographic position in terms of latitude and longitude. However, the number of geo-tagged messages is very limited. In addition, GPS coordinate and the location of the event mentioned in the Tweet may not be the same.

Similar issues appear for location attribute of user profiles. The number of profiles including valid location information is very limited. In addition, having valid location information in the profile does not guarantee that the user is actually at that location at the time of message posting, or the user is mentioning that location in the Tweet.

Challenges with the Tweet content and location attributes of user profiles mostly concern text processing and toponym recognition. The quality of content is not as good as that in news articles due to misspellings, extraordinary writing conventions and abbreviations. Therefore, state-of-the-art NLP parsers do not perform as accurately on informal social media message texts. The current efforts mostly rely on applying a normalization process before toponym recognition. The research trend is towards minimizing the normalization effort and proving a more general adaptive approach.

Toponym recognition and location estimation on social media research in the literature are mostly applied on Twitter posts. This is due to the fact that Twitter has a high number of users and most of the data in Twitter is publicly available through its application-programming interface. However the described techniques are also applicable to other crowd-sourced social media.

As a future work, several other research dimensions can be pursed for toponym recognition and location estimation. For toponym recognition, various other features can be analyzed. Hashtags appear to be valuable resources. In addition, links or photographs attached in the messages can be further investigated. Meta features or tags of the pictures and videos, too, may provide toponyms. For location estimation, in addition to combination of features from a single data source, evidence from several complementary data resources, such as Twitter and Foursquare, may be combined for increasing the precision of the location prediction.

References

Dempster, A. 1967. *Upper and Lower Probabilities Induced by a Multi-valued Mapping. Annals of Mathematical Statistics, 38*: 325–339.

Eryigit, G. 2014. ITU Turkish NLP web service. In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014.

Middleton, S. E., Middleton, L., & Modafferi, S. 2014. Real-Time Cri- sis Mapping of Natural Disasters Using Social Media. *Intelligent Systems, IEEE, 29*(2): 9–17.

Onal, K. D., Karagoz, P., & Cakici, R. 2014 (April). Turkce Twitter Gonderilerinde Lokasyon Tanima (Toponym Recognition on Turkish Tweets), SIU 2014, Trabzon, Turkey, pp. 1758–1761.

OpenStreetMap. Avaialble at: https://www.openstreetmap.org (Last accessed 15.2.2015).

Ozdikis, O., Oguztuzun, H., & Karagoz, P. 2013 (November). Evidential Location Estimation for Events Detected in Twitter. In: *Proceedings of ACM SIG Spatial/GIS Workshop on Geographic Information Retrieval (GIR)*. Orlando, USA, pp. 9–16.

Rui, L., Lei, K. H., Khadiwala, R., & Chang, K. C. C. 2012. TEDAS: A Twitter-based Event Detection and Analysis System." In: *Proc. 28th Int'l Conference on Data Engineering (ICDE '12)*, pp. 1273–1276.

Sakaki, T., Okazaki, M., & Matsuo, Y. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In Proc. of the 19th Int'l Conference on World Wide Web (WWW '10), pp. 851-860, 2010.

Sakaki, T., Okazaki, M., & Matsuo, Y. 2013. Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development. *IEEE Transactions on Knowledge and Data Engineering, 25*(4): 919–931.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. 2009. TwitterStand: News in Tweets. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, November 04–06, 2009. Seattle, Washington.

Seker, G. A., & Eryigit, G. 2012. Initial explorations on using CRFs for Turkish named entity recognition. *COLING*: 2459–2474.

Sagcan, M. A. 2014. Hybrid Method for Toponym Recognition on Informal Turkish Texts (M.Sc. Thesis), METU Computer Eng. Dept.

Stanford NLP Library. 2015. Available at: http://nlp.stanford.edu/software/index.shtml.

Wikipedia. Available at: http://www.wikipedia.org (Last accessed 15.2.2015).

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. 2012. Using Social Media to Enhance Emergency Situation Awareness. *Intelligent Systems, IEEE, 27*(6): 52–59.

Zemberek. Available at: https://github.com/ahmetaa/zemberek-nlp-distributions (Last accessed 15.2.2015).