Fine-Grained Geolocalisation of User Generated Short Text based on LBSN

Yao Deng
School of Communication Engineering
Xi’an University of Science and Technology
Xi’an, Shaanxi 710054, China
360676592@qq.com

Wenli Ji
School of Communication Engineering
Xi’an University of Science and Technology
Xi’an, Shaanxi 710054, China
jiwenli@xust.edu.cn

Yongjun Li*
School of Computer Science and Engineering
Northwestern Polytechnical University
Xi’an, Shaanxi 710072, China
lyj@nwpu.edu.cn

Xing Gao
School of Communication Engineering
Xi’an University of Science and Technology
Xi’an, Shaanxi 710054, China
771324116@qq.com

Wei Dong
School of Computer
National University of Defense Technology
Changsha, Hunan 410073, China
wdong@nudt.edu.cn

ABSTRACT

Recently, the fine-grained geolocalisation of User Generated Short Text (UGST) has been attracting much attention from academia. Most of existing methods rarely introduce the semantic information about a location in UGST, and do not prioritize the entities according to their importance. These reduce the performance of existing approaches. To tackle these problems, we propose a Fine-grained Geolocalisation of user-generated Short Text based on LBSN (FGST-L), which consists of three key components: 1) Using Foursquare as a source to build the tight coupling between entity and location, which can address the location-annotated sparseness problem. 2) Filtering out UGST if it does not contain any location-specific entities, which allows us to eliminate the interference of noisy UGSTs at the early stage. 3) Ranking the candidate locations for each remaining UGST based only on its textual data, and selecting the top-ranked location (or top *n* locations) for UGST. The experimental results show the effectiveness of FGST-L.

KEYWORDS

Geolocalisation; Fine-Grained; User Generated Short Text; LBSN

1 INTRODUCTION

With the increasing popularity of mobile social networking, the value of User Generated Short Text (UGST) in social networks is increasingly attracting much attention. Since UGST is spatially fine-grained [3], it could benefit many widespread applications, such as event detection [1], emergency analysis [8], digital health [9], etc.

However, since extremely few UGSTs are geocoded [2, 5, 6], the geolocalisation of UGST has become an important task needs to be solved. In this work, we focus on the fine-grained geolocalisation (i.e. street or special restaurant), which is very different from the works on coarse-grained geolocalisation. Generally, these works on coarse-grained geolocalisation link UGSTs to their originating cities or to time zones [2]. Clearly, the fine-grained geolocalisation of UGSTs is more useful for applications.

In existing works on fine-grained geolocalisation, Sheila et al. [4] geolocated the tweets based on content-similarity. Pavlos et al. [10] improved the above method by considering time-evolution characteristics in matching algorithm. Jorge et al. [3] adopted a weighted majority voting algorithm to the problem of fine-grained geolocalisation of tweets. Chong et al. [2] formulated the fine-grained geolocation as a ranking problem, and then proposed several models that leverage three types of signals from locations, users and peers. Most of existing works largely relied on GPS/human-annotated UGST to infer the location. When users have been less inclined to actively annotate UGSTs [5], the fine-grained geolocalisation is being a very challenging task.

In this paper we propose a Fine-grained Geolocalisation of user generated Short Text based on LBSN (FGST-L). We address the location-annotated sparseness problem by building the probabilistic models for locations using the entities in unstructured UGSTs. Our work differs from existing works in that we build the coupling between entity and location, not word and location. An entity has more semantic information than a word. Besides, we also consider the importance of different entities for geolocalisation. Following the idea of IDF [1], we assign different weights to different entities based on their popularity. To depict the tight coupling between entity and location, we use Foursquare as a source for building these

[*] Yongjun Li is the corresponding author.

[1] https://en.wikipedia.org/wiki/Tf-idf

RecNLP 2019, Jan 28, 2019, Honolulu, Hawaii, USA
© 2019 Association for Computing Machinery.
ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

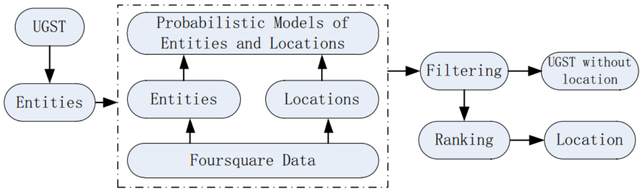

Figure 1: the Framework of FGST-L

probabilistic models. Based on the coupling models, we sketch our FGST-L. Given a non-geotagged UGST, we firstly extract the entities from UGST, and then determine whether this UGST contains the location-related information. We use the naive Bayes model to rank the candidate locations for the UGST with location-specific entities. Finally, we assign the top-ranked location to UGST. Experiment results on real datasets illustrate the superiority of FGST-L compared with the-state-of-art methods.

2 THE PROPOSED APPROACH

2.1 Overview

Figure 1 illustrates the high level data flow of our FGST-L. Our proposed method consists of four steps. (I) **Extracting the entities in UGST** : we extract all entities from each UGST based on Microsoft Probase[2]. (II) **Building the probabilistic model of entity and location** : we use Foursquare as source to build the tight coupling between entity and location, which can address the location sparseness problem. (III) **Filtering the UGSTs** : we filter out UGST if it does not contain any location-specific entities, which allows us eliminate the interference of noisy UGSTs at early stage. (IV) **Ranking the candidate locations** : we rank the candidate locations for each remaining UGST based only on its textual data, and select the top-ranked location for UGST. We will explain each step in detail in the following subsections.

2.2 Extracting Entities in UGST

For the sake of text analysis, we firstly preprocess each UGST by breaking the UGST into tokens, stemming them, and removing stop words. After that, we formally model each UGST t as a set of words, denoted by $t' = \{w_1, w_2, ..., w_i, ..., w_n\}$, where w_i is the i^{th} word of UGST t. Intuitively, an entity contains more sematic information than a word. For example, the entity *New York* is more indicative than word *New* or word *York*. Thus, we further denote t as a set of entities $\{e_1, e_2, ..., e_i, ..., e_m\}$, where $e_i = \{w_k, w_{k+1}, ..., w_l | 1 \leq k \leq l \leq n\}$.

Given a UGST t, we expect to obtain all entities in t. We firstly use the Stanford natural language processing tool [7] to get the bag-of-words in t, then find all possible entities based on Microsoft Probase. As a result, we obtain $t = \{e_1, e_2, ..., e_m\}$. In this situation, entity e_i and entity e_j may contain the common words. For example, both *Washington square park* and *square park* are permissible entities in UGST *amazing views of Washington square park*. This is a reasonable approach because some users refer to Washington Square Park as *Washington square park*, and others refer as *square park*. Surely the entities we obtained are restricted to the selected repository. The reason why we select Probase is that Probase provides a tremendous concept space and concept clusters.

2.3 Building Probabilistic Model of Entity and Location

Foursquare, as a location-based social network, has a collection of PoIs (Point of Interest), and each tip is associated with a UGST message and a PoI. This makes Foursquare a valuable resource for building high quality models for locations [5].

Assume the set of locations in Foursquare is $L = \{l_1, l_2, ..., l_n\}$. To depict the coupling between entity and location, we first build a conditional probability model for each PoI based on a set of tips associated to that PoI. We denote the set of tips associated to PoI l_i as $T(l_i) = \{t_1, t_2, ..., t_m\}$. Obviously, different PoIs have different numbers of tips. These popular PoIs may have more tips, so their models are higher quality. Let $tf(e, t)$ denote the number of occurrences of entity e in the UGST t, $c(e, l)$ denote the number of occurrences of entity e in all tips associated to PoI l. We use the maximum likelihood estimation method to calculate the probability of entity e in PoI l as follows.

$$\begin{aligned} p(e|l) &= \frac{c(e,l)}{\sum_{e_k \in E(l)} c(e_k, l)} \\ c(e,l) &= \sum_{t \in T(l)} tf(e,t) \\ E(l) &= \{e | e \in t, t \in T(l)\} \end{aligned} \tag{1}$$

where $E(l)$ denotes the set of entities associated to PoI l.

In some cases, if $c(e, l) = 0$, there is $p(e|l) = 0$. This leads to zero probability problem. We use Laplace smoothing method to address this problem. $p(e|l)$ is further defined as follows.

$$p(e|l) = \frac{c(e,l) + 1}{\sum_{e_k \in E(l)} (c(e_k, l) + 1)} \tag{2}$$

Then we can calculate the probability of UGST t in PoI l as follows.

$$\begin{aligned} p(t|l) &= \prod_{e_i \in t} p(e_i|l) \\ &= \prod_{e_i \in t} \frac{c(e_i,l) + 1}{\sum_{e_k \in E(l)} (c(e_k, l) + 1)} \end{aligned} \tag{3}$$

2.4 Filtering UGSTs

Geocoding of UGSTs is largely dependent on the inferability of location from textual information. For example, we can not infer the location of UGST *It is a good day*. Before geocoding a UGST, we determine whether this UGST has some hints about location. We filter this UGST out if it has no any indicative information.

In some cases, the hint about location occurs in UGST explicitly. For example, entity *Washington square park* is an explicit hint in UGST *I am at Washington square park now*. Given an entity $e_i \in t$, we define the following indicator function to illustrate whether e is a location.

$$1^{ex}(e_i, L) = \begin{cases} 1, & e_i \in L \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

We further illustrate whether UGST t has hint about location by the following function, where $\bigvee$ is *OR* operation.

$$1^{1}(t, L) = \bigvee_{e_i \in t} 1^{ex}(e_i, L) \tag{5}$$

[2] https://concept.research.microsoft.com/

In other cases, the hint about location appears implicitly. For example, entity *Big Apple* is the implicit hint in some UGSTs. Inspired by TFIDF's successful application on identifying local words [5], we define the following function.

$$f_{tfidf}(e,l) = \frac{c(e,l)+1}{\sum_{e_k \in E(l)}(c(e_k,l)+1)} \times \left[\ln \frac{|L|}{df(e)+1} + 1\right] \tag{6}$$

where $df(e)$ is the number of locations having e in the UGSTs.

If $f_{tfidf}(e,l) \geq \theta$, we consider the entity e as a local entity with respect to l. θ is a predefined threshold. The greater θ is, the smaller the number of identified local entities is. θ is a tuning parameter. The indicator function is defined over $f_{tfidf}(e,l)$ as:

$$\mathbb{1}^{im}(e_i,l) = \begin{cases} 1, & f_{tfidf}(e_i,l) \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

We further illustrate whether UGST t has any implicit hint by the following function.

$$\mathbb{1}^{2}(t,L) = \bigvee_{l \in L} \bigvee_{e_i \in t} \mathbb{1}^{im}(e_i,l) \tag{8}$$

We define the following indicator function to express whether t has hint about location.

$$\mathbb{1}(t,L) = \mathbb{1}^{1}(t,L) \bigvee \mathbb{1}^{2}(t,L) \tag{9}$$

If $\mathbb{1}(t,L) = 1$, we consider that t has hint about location. Otherwise, we filter t out.

2.5 Ranking Candidate Locations

After filtering out those UGSTs without any hint about location, we need to rank the candidate locations for each UGST, and select the top-ranked location (or top n locations) as the location of UGST. Given a UGST $t = \{e_1, e_2, ..., e_m\}$ and its candidate locations $L = \{l_1, l_2, ..., l_m\}$. We use simple naive Bayes probabilistic model to rank the candidate locations. The probability that the location of t is l can be defined as follows.

$$\begin{aligned} p(l|t) &\propto p(t|l) \times p(l) = \left(\prod_{e_i \in t} p(e_i|l)\right) \times p(l) \\ p(l) &= \frac{N(l)}{\sum_{l_i \in L} N(l_i)} \end{aligned} \tag{10}$$

where $N(l)$ is the occurrences of l.

UGSTs often contain general entities such as *delicious food* and *park*. Those entities have lower discriminability. Instead, there are also some local entities, such as *Big Apple*, which are tight coupling with special locations. These entities have higher discriminability. In other words, the importance of an entity is weighted based on its occurrence as a location identifier. Given entity e_i, its weight is defined as follows.

$$w(e_i) = 1 + \ln \frac{|L|}{df(e_i)+1} \tag{11}$$

The probability that the location of t is l can be further defined as follows.

$$p(l|t) \propto \left(\prod_{e_i \in t} (w(e_i) \times p(e_i|l))\right) \times p(l) \tag{12}$$

$$\begin{aligned} \ln p(l|t) &\propto \sum_{e_i \in t} (\ln w(e_i) + \ln p(e_i|l)) + \ln p(l) \\ &= \sum_{e_i \in t} \ln w(e_i) + \sum_{e_i \in t} \ln p(e_i|l) + \ln p(l) \end{aligned} \tag{13}$$

3 EXPERIMENTS

3.1 Datasets

We crawl the Foursquare locations (called PoI) from *New York* spanning from Jan. 2018 to Apr. 2018, and their tips by Foursquare API. A dataset consiting 498722 tips and 74942 PoIs was obtained. The distribution of tips discernibly varied over all PoIs, see Table 1.

Table 1: Relationship between PoIs and their Tips

Tips	≥ 1	≥ 10	≥ 20	≥ 30	≥ 40	≥ 50
PoIs	74942	9208	4939	3790	2930	2132

The numbers of tips for about 87.7% of PoIs are less than 10. Only 3790 PoIs have more than 30 tips. Intuitively, the number of tips has some effect on the performace of FGST-L. We conduct the experiment to validate our prediction.

To prove effectiveness and generalization of FGST-L, we also gather a set of tweets from Twitter and a set of posts from Facebook. We only select the tweets or posts from *New York*. After selection, we obtain 19231 tweets and 6699 posts. 16.7% of posts and 32.4% of tweets are geocoded the fine-grained location, which are used in our evaluation. Besides, we also select 12000 tips from Foursquare, and 6000 tips containing hints about location. These three datasets are our test datasets and named as FB, TW and FS, respectively.

3.2 Experimental Setting

We compare FGST-L with the following representative methods.

- FRV [5]: a 3-step technique (Filtering-Ranking-Validating) for fine-grained tweet location prediction, which uses Foursquare as resource for building the tight coupling between words and locations.
- LW [2]: a Location-indicative Weighting scheme which assigns more weights to location-indicative words, and is easily incorporated into the naive Bayes model.
- WMV [3]: a Weighted Majority Voting algorithm to the problem of fine-grained geolocalisation of tweets, which estimates the geographical location of tweet by collecting the geo-location votes of the geo-tagged tweets that are most similar regarding their contents to that tweet.

For all methods, we conduct all experiments on the same datasets. We employ two widely used *Accuracy@1km*, *Average Error Distance (km)* [3] to evaluate the performance of all algorithms.

Average Error Distance (km): we compute the distance on Earth between the predicted location and the real coordinates of the UGST in our ground truth.

Accuracy@1km: the accuracy of the model is measured by determining whether a UGST t has hint about location, and whether a predicted location lies within a radius of 1km from the real location.

3.3 Performance of FGST-L w.r.t θ

As shown in Eq. (7) and Eq. (8), given a UGST t, t is filtered out if all $f_{tfidf}(e,l) < \theta$. Naturally, the predefned threshold, θ, has a strong

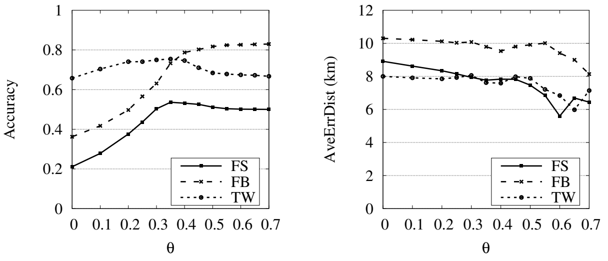

Figure 2: Results w.r.t. θ

influence on the results. To study the effect of θ on the results, we conduct the experiments according to the different values of θ. The results are shown in Figure 2.

The left figure of Figure 2 shows the *Accuracy* w.r.t. θ. On FS and TW, as θ increases, the accuracy of FGST-L rises firstly and then drops. The accuracy reaches peak when $0.3 < \theta < 0.4$. Generally, many UGSTs do not have any hint information about location. When θ is relatively small, many UGSTs without location information are mistaken for UGSTs with location information. This leads to reduce the accuracy of FGST-L. Similarly, when θ is relatively large, FGST-L misjudges the UGSTs with location information as UGSTs without location information, which also reduce the accuracy. On FB, the accuracy curve shows different trend. The curve rises quickly at the beginning then rises slightly later. Because most of UGSTs in FB are location-independent, as θ increases, more and more UGSTs are correctly classified as location-independent UGSTs. When $\theta > 0.4$, most of UGSTs without location information have been correctly filtered out. This leads to the increased trend slows. In FB, the severe data skew is the main reason that the accuracy keeps increasing. When $\theta = 0.7$, all UGSTs are filtered out, and the accuracy reaches peak at 83.3%.

The right figure of Figure 2 shows the *Average Error Distance* w.r.t. θ. When $0.3 < \theta < 0.4$, the average error distance of inferred location is optimal. *Average Error Distance* has minimum value when $\theta > 0.6$. However, most of UGSTs are filtered out in this situation, which is not our expectation.

3.4 Performance Comparison between FGST-L and Existing Works

Figure 3 illustrates the performance comparsion between FGST-L and existing works on three datasets. From the results, we easily reach the following conclusions.

FGST-L outperforms all baselines which validates the effectiveness of FGST-L. This is because we 1) build the tight coupling between entity and location, not word and location. The entity has more semantic information than word; 2) assign the entities different weights according to their occurrences; 3) filter out the location-independent UGSTs.

FRV also performs better than WMV and LW. This indicates that filtering out the UGSTs without location information is an effective step. When some of UGSTs have no hint about location, the filtering process can reduce the impact of noisy data.

FGST-L and FRV perform better on FB than on FS and TW. As described above, about 83.3% of UGSTs on FB are location-independent. The filtering process filters these UGSTs out, which increases the accuracy. Besides, the accuracies of all methods on FS are worse than results on FB and TW. It is because the percentage of location-dependent UGSTs in FS is the largest. This futher indicates that determining whether a UGST is location-independent is easier than inferring the fine-grained location of UGST.

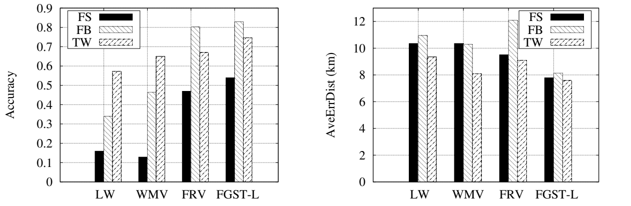

Figure 3: Results on FS, FB and TW

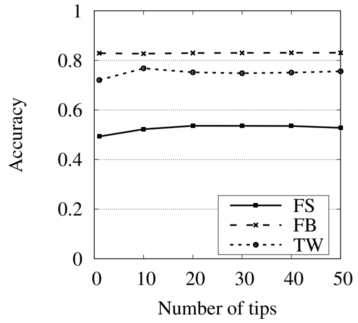

Figure 4: Results w.r.t. *Number of Tips*

3.5 Performance of FGST-L w.r.t *Number of Tips*

As shown in Table 1, there is somewhat difference on the number of tips for different PoIs. Intuitively, the more the number of tips is, the more accurate the probabilistic model for PoI is. In this subsection, we futher conduct the experiments to validate whether the number of tips has impact on the probabilistic models for locations. We train the probabilistic models with different numbers of tips, and calculate the accuracy of each model. The results are shown in Figure 4.

We find that the number of tips has not much impact on the probabilistic models for locations. This result is contrary to our intuition. Especially, when it is larger than 20, the number of tips has little impact on the model. This result may be due to two reasons: 1) in our three test datasets, more than half of UGSTs are location-independent; 2) the futher study shows that the entities in 20 UGSTs cover almost entities in all UGSTs for a PoI. It is appropriate that we build the tight coupling between entities and locations based only on 20 UGSTs. This could reduce the need for computing resources.

4 CONCLUSIONS

In this paper, we consider both the tight coupling between entities and locations and the importance of different entities for geolocalisation, and then propose a novel Fine-grained Geolocalisation of user generated Short Text based on LBSN, where the weights of different entities are determined according to their occurrences in UGSTs. Experiments on real datasets validate the effectiveness of the proposed method.

ACKNOWLEDGMENTS This work was partly supported by the Shaanxi Provincial Natural Science Foundation, China (2014JM2-6104).

REFERENCES

[1] Farzindar Atefeh and Wael Khreich. 2015. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence* 31, 1 (2015), 132–164.

[2] Wen-Haw Chong and Ee-Peng Lim. 2017. Tweet Geolocation: Leveraging Location, User and Peer Signals. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 1279–1288.

[3] Jorge David Gonzalez Paule, Yashar Moshfeghi, Joemon M. Jose, and Piyushimita Thakuriah. 2017. On Fine-Grained Geolocalisation of Tweets. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, New York, NY, USA, 313–316.

[4] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*. ACM, New York, NY, USA, 61–68.

[5] Kisung Lee, Raghu K. Ganti, Mudhakar Srivatsa, and Ling Liu. 2014. When Twitter Meets Foursquare: Tweet Location Prediction Using Foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST, Brussels, Belgium, 198–207.

[6] Yongjun Li, Zhen Zhang, You Peng, Hongzhi Yin, and Quanqing Xu. 2018. Matching user accounts based on user generated content across social networks. *Future Generation Computer Systems* 83 (2018), 104 – 115.

[7] Christoper Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60.

[8] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2016. EAIMS: Emergency Analysis Identification and Management System. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 1101–1104.

[9] Anastasios Noulas, Colin Moffatt, Desislava Hristova, and Bruno Gonçalves. 2018. Foursquare to the Rescue: Predicting Ambulance Calls Across Geographies. In *Proceedings of the 2018 International Conference on Digital Health*. ACM, New York, NY, USA, 100–109.

[10] Pavlos Paraskevopoulos and Themis Palpanas. 2015. Fine-Grained Geolocalisation of Non-Geotagged Tweets. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, New York, NY, USA, 105–112.