Linking Fine-Grained Locations in User Comments
(Extended abstract)

Jialong Han *, Aixin Sun †, Gao Cong †, Wayne Xin Zhao ‡, Zongcheng Ji †, Minh C. Phan †

* *Tencent AI Lab, Shenzhen, China*

† *School of Computer Science and Engineering, Nanyang Technological University, Singapore.*

‡ *School of Information, Renmin University of China, Beijing, China.*

{jialonghan,batmanfly,jizongcheng}@gmail.com, {axsun,gaocong}@ntu.edu.sg, phan0050@e.ntu.edu.sg

***Abstract*—Many domain-specific websites host a profile page for each entity (*e.g.,* locations on Foursquare, movies on IMDb, and products on Amazon), and users can post comments on it. When commenting on an entity, users often mention other entities for reference or comparison. Compared with web pages and tweets, disambiguating the mentioned entities in user comments has not received much attention. This paper investigates linking fine-grained locations in Foursquare comments. We demonstrate that the *focal location*, *i.e.,* the location that a comment is posted on, provides rich contexts for linking. To exploit such information, we represent the Foursquare data in a graph, which includes locations, comments, and their relations. A probabilistic model named FocalLink is proposed to estimate the probability that a user mentions a location when commenting on a focal location, by following different kinds of relations. Experimental results show that FocalLink is consistently superior to different baselines.**

I. Introduction

With the prevalence of GPS-enabled devices such as smartphones and tablets, people are sharing on location-based social networks (LBSNs) their experiences about *fine-grained locations*, *e.g.,* restaurants, malls, and parks. On typical LBSNs like Foursquare, Yelp, and Google Maps, a dedicated profile page is hosted for each location. A user can open a location's page to view information, or post ratings and comments on it.

In Fig. 1, we exemplify with data from Foursquare. In this figure, locations are represented by ellipses. Location profile pages are indicated by rectangular boxes, where comments on the respective locations are posted. For instance, on the page of IMM Building (a shopping mall in Singapore), a user left a comment saying "Go for *daiso*, 3rd fl, the 2dollar shop". For clarity, given a comment, we refer to the location being commented on (*e.g.,* IMM Building) as the focal location.

When commenting on the focal location, users may possibly mention some other locations like *daiso* (a chain miscellaneous store) in the above example. If those mentions could be identified and linked to the right pages as in Fig. 1, the following applications could be achieved or enhanced:

- **Comment gathering**. Comments mentioning business places like the above *daiso* could be sent to their owners for reference, even if they were posted elsewhere.
- **Sentiment analysis**. Sentiment analysis with all comments on a location page could be applied with care, because sentiments expressed in comments mentioning other locations may not be intended for the focal location.

The TKDE journal version [1] of this extended abstract was done when Jialong Han was a Research Fellow with School of Computer Science and Engineering, Nanyang Technological University, Singapore.

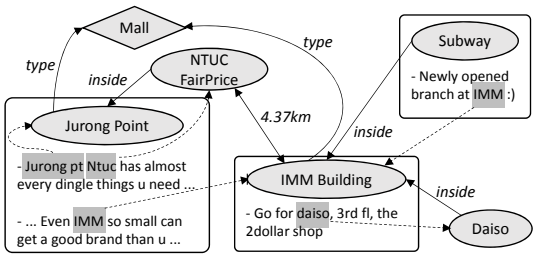

Fig. 1. A motivating example of Foursquare data graph.

- **Location recommendation**. A comment on one location but mentioning another may indicate certain connections between both locations, which could be a strong signal to exploit in (next) location recommendation.

Besides the above benefits, we note that our study is generalizable to other domains like movie (*e.g.,* IMDb) and e-commence (*e.g.,* Amazon), since websites in those domains also possess similar structures as in Fig. 1, where interconnected pages for entities are hosted to receive comments.

Resolving ambiguous entity mentions to the correct entries in a database, *a.k.a. entity linking* [2], has been relatively well studied for formal documents like web pages. However, efforts on the same task for user comments remain limited. Moreover, like tweets, the short[1] nature of comments also renders the task here challenging. For example, given that *daiso* has more than ten branches in Singapore, it is difficult to judge solely from the above short comment which one it is referring to.

Since text information is scarce in comments, any extra information should be exploited to assist with the task. Luckily, for a given comment, the focal location is always available and unambiguous, which provides important contextual information. For example, if we know that the comment about *daiso* is posted on IMM Building, we know that the "*daiso*" here tends to refer to the branch located inside IMM, because of their spatial containment *relation* indicated by "3rd fl".

In this paper, we investigate how focal locations could serve as additional clues to link locations in comments. Specifically, we view all locations and their relations, *e.g.,* space containments, type overlaps, and geographical distance, as a graph like Fig. 1. We propose FocalLink, a probabilistic model that estimates how likely a given mention refers to a location. In this model, we seamlessly incorporate the focal location context along with other useful information to assist linking.

[1]The length of comments is 15 words on average, according to our Foursquare dataset of 0.44 million comments on Singapore locations.

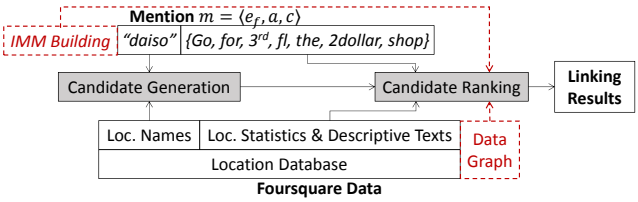

Fig. 2. An overview of our solution.

II. Problem Definition and Solution Framework

Problem Definition. We denote by $e \in \mathcal{E}$ a location in a database. By leveraging CRF++[2] on a given comment with focal location e_f, multiple *mentions* could be extracted, each denoted by $m = \langle e_f, a, c \rangle$ with an anchor a and a surrounding context c. For m, we want the entry $e^{(m)} \in \mathcal{E}$ it refers to.

Solution Framework. We decompose the task into two subtasks, namely *candidate generation* and *candidate ranking*, as illustrated in Fig. 2. In the first stage, locations potentially matching m by the name are retrieved. They are fed to the second stage, which ranks them and links m to the top one. In the following, we concentrate on the second stage.

III. FocalLink: Exploiting Focal Locations

We adopt a probabilistic framework, and model the generative process of a mention $m = \langle e_f, a, c \rangle$ in **three steps**:

Step 1: Draw a relation to follow. We consider a set R of six single- or multi-step relations (*e.g., Inside*, *Co-Type*, and *Near*) on the graph $\mathcal{G}$. Here, user picks $r \in R$ from an unknown multinomial distribution $\boldsymbol{\pi}$.

Step 2: Draw e to mention. Starting from e_f and walking along r on graph $\mathcal{G}$, user picks a location e to mention. Path-Constrained Random Walk [3] is used to estimate $P_{\mathcal{G}}(e|e_f, r)$, *i.e.,* how likely a user will end up with (or mention) e.

Step 3: Write the surrounding context. User wraps the anchor a with some words w about e, e_f, and r as the surrounding context c. We regard w as being independently generated from a mixture of $P(w|d_e)$, $P(w|\boldsymbol{\theta}_r)$, and $P(w|D)$, *i.e.,* words related to e, r, and background noise. Here $P(w|d_e)$ and $P(w|D)$ are directly estimated from the data, while $P(w|\boldsymbol{\theta}_r)$ is to be learnt by FocalLink.

Given a collection of unlabeled mentions $m = (e_f, a, c)$ with candidates, we adopt the **E**xpectation-**M**aximization (EM) algorithm to estimate parameters $\Phi = (\{\pi_r\}, \{\boldsymbol{\theta}_r\})$.

IV. Experiments

Dataset. We collected 321,943 Singapore locations and 442,803 comments on them from Foursquare. A sample of 4,000 comments were annotated, with 828 mentions found.

Baselines. We compare with the three baselines in Table I. As their names indicate, they have different access to information like location popularity, surrounding context, and geographical coordinates of focal locations.

[2]https://taku910.github.io/crfpp/

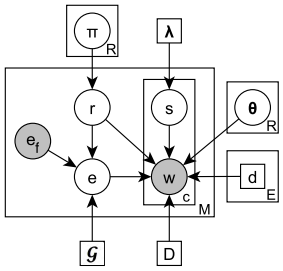

Fig. 3. Bayesian graphical representation of FocalLink.

TABLE I
Linking results of all methods.

Approaches	Prec	Rec	F_1
Popularity	.563	.508	.534
PopContext	.619	.558	.587
PopContextDist	.679	.611	.643
FocalLink	**.690**	**.621**	**.653**

TABLE II
Top-10 relation words learnt by FocalLink, sorted by $P(w|\boldsymbol{\theta}_r)$.

Self-Ref	at, to, in, on, the, it, of, s, is, from
Inside	st, at, cross, outlet, locate, rd, open, on, road, middle
Contains	at, stall, noodle, on, from, must, fry, mee, serve, has
Co-Inside	plaza, thomson, at, go, hungry, atm, check, out, beside, wordpress
Co-Type	better, than, at, on, much, compare, cheaper, as, to, price
Near	to, from, at, bus, walk, locate, road, mrt, take, go

Results. In Table I, the performance of all baselines and FocalLink increases in the order they are presented. The PopContext baseline outperforms Popularity by 5 points in terms of all metrics. This is because the compatibility between related words of a candidate and the context words may imply that the comment is mentioning the candidate. When the distance between a candidate and the focal location is taken into consideration, PopContextDist can improve all three metrics by at least 5 points. This suggests that, for most of the time, users tend to use the focal location as a geographical context and mention nearby locations. Finally, FocalLink manages to achieve one point of improvement over PopContextDist, the most competitive baseline.

Table II presents top-10 relation-indicative words learnt by FocalLink w.r.t. $P(w|\boldsymbol{\theta}_r)$ for each relation. For example, for relation *Inside*, words like "at", "cross", "locate", "on" and "middle" modify the mentioned location where e_f is inside, *e.g.,* ⟨ `Crowne Plaza Changi Airport` (an airport hotel), "Conveniently **located** in terminal 1 ..."⟩. Meanwhile, "st", "rd", and "road" are used to give further directions. For relation *Co-Type*, readers may notice that almost all words here are related to comparisons. As indicated by the order, the most common comments for comparisons are like ⟨e_f, "**Better than** e."⟩. Besides, users often write comparative degrees of adjectives following the word "much", though most of the time they care for "much cheaper price".

References

[1] J. Han, A. Sun, G. Cong, W. X. Zhao, Z. Ji, and M. C. Phan, "Linking fine-grained locations in user comments," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 59–72, 2018.
[2] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data." in *EMNLP-CoNLL*, vol. 7, 2007, pp. 708–716.
[3] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine Learning*, vol. 81, no. 1, pp. 53–67, 2010.