POIReviewQA: A Semantically Enriched POI Retrieval and Question Answering Dataset

Gengchen Mai, Krzysztof Janowicz
STKO Lab, UCSB
{gengchen_mai,janowicz}@geog.ucsb.edu

Cheng He, Sumang Liu, Ni Lao
SayMosaic Inc.
{cheng.he,sumang.liu,ni.lao}@mosaix.ai

ABSTRACT

Many services that perform information retrieval for Points of Interest (POI) utilize a Lucene-based setup with spatial filtering. While this type of system is easy to implement it does not make use of semantics but relies on direct word matches between a query and reviews leading to a loss in both precision and recall. To study the challenging task of semantically enriching POIs from unstructured data in order to support open-domain search and question answering (QA), we introduce a new dataset POIReviewQA[1]. It consists of 20k questions (e.g."is this restaurant dog friendly?") for 1022 Yelp business types. For each question we sampled 10 reviews, and annotated each sentence in the reviews whether it answers the question and what the corresponding answer is. To test a system's ability to understand the text we adopt an information retrieval evaluation by ranking all the review sentences for a question based on the likelihood that they answer this question. We build a Lucene-based baseline model, which achieves 77.0% AUC and 48.8% MAP. A sentence embedding-based model achieves 79.2% AUC and 41.8% MAP, indicating that the dataset presents a challenging problem for future research by the GIR community. The result technology can help exploit the thematic content of web documents and social media for characterisation of locations.

CCS CONCEPTS

• **Information systems → Question answering**; *Relevance assessment*;

KEYWORDS

POI, Search, Question Answering, Semantic Enrichment

ACM Reference Format:
Gengchen Mai, Krzysztof Janowicz and Cheng He, Sumang Liu, Ni Lao. 2018. POIReviewQA: A Semantically Enriched POI Retrieval and Question Answering Dataset. In *12th Workshop on Geographic Information Retrieval (GIR'18), November 6, 2018, Seattle, WA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3281354.3281359

[1]http://stko.geog.ucsb.edu/poireviewqa/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GIR'18, November 6, 2018, Seattle, WA, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6034-0/18/11.
https://doi.org/10.1145/3281354.3281359

1 INTRODUCTION

Location-based services (LBS) and the underlying Point of Interest (POI) datasets play a increasingly important role in our daily interaction with mobile devices. Platforms such as Yelp, Foursquare, Google Map allow users to search nearby POIs based on their names, place types, or tags, which requires manual data annotation. In fact, besides these structured data, POIs are typically associated with abundant unstructured data such as descriptions and users' reviews which contain useful information for search and question answering purpose. For example questions like "Is this restaurant dog friendly?" or "Is this night club 18+?" can be answered by relevant text in reviews such as "Great dog friendly restaurant" or "18+ night club". This information can also help accomplishing search needs such as "find night clubs near me which are 18+".

There are only a few existing GIR benchmark datasets (e.g., GeoCLEF [4]) and they often lack in rich annotations as would be required for the examples above. Recently many datasets have been produced for reading comprehension such as SQuAD [5]. However, they do not have a spatial/platial component. Here we present a POI search and question answering dataset called POIReviewQA with detail annotations of context and answers. Baseline models are implemented to demonstrate the difficulty of this task.

Our work provides an evaluation benchmark for geographic information retrieval and question answering systems. It follows the idea of semantic signatures for social sensing [2] by which we can study POI types using patterns extracted from human behavior, e.g., what people write about places of a particular type. Intuitively, questions about age limits only arise in the narrow context of a few such types, e.g., nightclubs, movie theaters, and so on. Furthermore, unstructured data such as reviews are often geo-indicative without the need for explicit geographic coordinates. For instance, people may be searching for a *central but quiet hotel* [3]. It is those questions that we will address in the following.

2 THE POIREVIEWQA TASK

We created POIReviewQA based on the Yelp Challenge 11 (YC11) dataset[2] and the QA section of POI pages.

Query Set Generation. We create the question answer dataset from the "Ask the Community" section[3] of POI pages. The Yelp platform is dominated by popular business types such as restaurants. In order to produce a balanced query set for all business types we performed stratified sampling: 1) count the frequencies of POI name suffixes (single words) in YC11; 2) for every suffix with at least frequency 10 we create a quoted search query restricting to the Yelp

[2]https://www.yelp.com/dataset/challenge
[3]https://www.yelpblog.com/2017/02/qa

Table 1: The Statistic of POIReviewQA

# of Annotated question	4,100
% of questions WITHOUT related reviews	11.4%
Avg. # of related *reviews* per question	4.61
Avg. # of 1 rater agreeing on relevant *sentence* per question	2.19
Avg. # of 2 raters agreeing on relevant *sentence* per question	1.08
Avg. # of 3 raters agreeing on relevant *sentence* per question	0.83

business QA domain[4], and collect community QA page URLs from Google search engine; 3) collect questions and answers from the community QA pages. In total, 1,701 quoted search queries results are collected from Google with up to 100 search results for each query. Since the last term often indicates the place type of a POI, the collected Yelp business question pages have a wide coverage of different place types. In total 20K questions were collected from Yelp business question pages for 1022 Yelp business types. Each question is associated with one or multiple POIs with several POI types, e.g., *Echoplex* (`Music Venues`, `Bars`, `Dance Clubs`) or *Paper Tiger Bar* (`Cocktail Bars`, `Lounges`).

Relevance and Answer Annotation. For each question, 10 review candidates are selected by stratified sampling from the search result of a lucene-based setup, i.e., applying *Elastic Search* to POI reviews based on the question with constraint to the associated POI types. We developed a crowd-facing Web server and deployed it on Amazon Mechanical Turk to let raters annotate each sentence of these 10 reviews with respect to whether it answer the current question and what the corresponding answer is. The annotation results are collected for each question. To date, we have collected about 4100 questions. Basic statistic for these are shown in Tab. 1. In order to study the relationship between raters (given 3 raters per review sentence) and the accuracy of the raters, we divide the sentences into 4 sets based on the number of raters that agreed on each sentence, denoted as R_0, R_1, R_2, R_3. Then we randomly sample 20 sentences from each of the last three sets (R_1, R_2, R_3). By manually inspecting the relevance of these sentences to the corresponding questions. The resulting accuracy of each sample set is 45% for R_1, i.e., 9/20 sentences, 90% for R_2, 100% for R_3. We treat the sentences in R_2, R_3 as relevant, and the rest are labeled as irrelevant sentences. These labels are used to evaluate different models.

Evaluation Metrics. Area under curve (AUC) and mean average percision (MAP) are used as evaluation metrics.

3 EXPERIMENT WITH BASELINE MODELS

In order to provide a similar search functionality to Yelp's new review-based POI search[5], we developed a *TF-IDF based model* to search through all sentences from 10 reviews based on a question. An evaluation using the POIReviewQA dataset gives 77% AUC and 48.8% MAP. We also applied the *sentence embedding model* proposed by Sanjeev Arora et al. [1]. It improves the average word embeddings using SVD and gives what the authors call "tough-to-beat" results for a text similarity tasks. We use the pretrained Google News 300 dimension Word2Vec embeddings to generate the sentence level embedding for both questions and review sentences.

[4]Search "site:https://www.yelp.com/questions/ 'Restaurant" via Google
[5]https://engineeringblog.yelp.com/2017/06/moving-yelps-core-business-search-to-elasticsearch.html

Table 2: Examples of POIReviewQA. Each example consists of a question Q, one or more POI types T, a context sentence C from the POI reviews, and an answer A. The ranking of sentence (C) based on human judgements (H), Lucene (L), and sentence embedding (E) is also shown.

Reason	Example	Ranking (H/L/E)
Paraphrase	**Q:** About **how long** should I expect my visit to be? **T:** Venues & Event Spaces; Kids Activities **C:** We **were there for** about 2 hours, including the show. **A:** took 2 hrs	1/107/88 out of 158
Hyponym	**Q:** Any good **vegan choices**? **T:** Restaurants→Cajun/Creole **Sent:** After scanning the menu for a bit however, I was able to find the **tofu** wings. **A:** Tofu wings could be a choice	2/49/18 out of 83
Synonymy	**Q:** Any recommendations on how to **score a table**? ... **T:** Restaurants→French **C:** I **made a reservation** a day in advance thinking it will be busy. **A:** A day in advance	1/63/14 out of 98
Deduction	**Q:** Are there classes for **seniors**? **T:** Art Galleries; Art Schools **C:** Great studio for **all**, including kids! **A:** There are classes for seniors	1/60/15 out of 72
Common Sense	**Q:** Do they **buy comic books**? **T:** Shopping→Comic Books **C:** The concerns: The store currently has **no consignment or new issues**. **A:** No	1/45/53 out of 62

Then their cosine similarities are used to rank the sentences given a question. Evaluation by POIReviewQA gives 79.2% AUC and 41.0% MAP. Comparing to the TF-IDF model, the sentence embedding-based model gives a higher AUC (which is sensitive to overall rankings) but lower MAP (which is sensitive to top rankings). The results from both baseline models indicate that the POIReviewQA dataset presents a challenging task. Table 2 shows examples for which the baseline model fails. Correctly predicting relevant sentence requires an understanding of language and common sense. We hope that the dataset will enable further GIR research about question answering as it relates to place types.

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations*.
[2] Krzysztof Janowicz, Grant McKenzie, Yingjie Hu, Rui Zhu, and Song Gao. 2018. Using Semantic Signatures for Social Sensing in Urban Environments. In *Mobility Patterns, Big Data and Transport Analytics*.
[3] Krzysztof Janowicz, Marc Wilkes, and Michael Lutz. 2008. Similarity-based information retrieval and its role within spatial data infrastructures. In *International Conference on Geographic Information Science*. Springer, 151–167.
[4] Thomas Mandl, Paula Carvalho, Giorgio Maria Di Nunzio, Fredric Gey, Ray R Larson, Diana Santos, and Christa Womser-Hacker. 2008. GeoCLEF 2008: the CLEF 2008 cross-language geographic information retrieval track overview. In *Working Notes for CLEF 2008 Workshop*. Springer, 808–821.
[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).