Towards an Integrated View of Semantic Annotation for POIs with Spatial and Textual Information

Dabin Zhang[1] , **Ronghui Xu**[1] , **Weiming Huang**[2] , **Kai Zhao**[3] and **Meng Chen**[1*]
[1]School of Software, Shandong University
[2]School of Computer Science and Engineering, Nanyang Technological University
[3] Robinson College of Business, Georgia State University
{zdb, ronghuix}@mail.sdu.edu.cn, weiming.huang@ntu.edu.sg, kzhao4@gsu.edu, mchen@sdu.edu.cn

Abstract

Categories of Point of Interest (POI) facilitate location-based services from many aspects like location search and POI recommendation. However, POI categories are often incomplete and new POIs are being consistently generated, this rises the demand for semantic annotation for POIs, i.e., labeling the POI with a semantic category. Previous methods usually model sequential check-in information of users to learn POI features for annotation. However, users' check-ins are hardly obtained in reality, especially for those newly created POIs. In this context, we present a Spatial-Textual POI Annotation (STPA) model for static POIs, which derives POI categories using only the geographic locations and names of POIs. Specifically, we design a GCN-based spatial encoder to model spatial correlations among POIs to generate POI spatial embeddings, and an attention-based text encoder to model the semantic contexts of POIs to generate POI textual embeddings. We finally fuse the two embeddings and preserve multi-view correlations for semantic annotation. We conduct comprehensive experiments to validate the effectiveness of STPA with POI data from AMap. Experimental results demonstrate that STPA substantially outperforms several competitive baselines, which proves that STPA is a promising approach for annotating static POIs in map services.

1 Introduction

Today's proliferation of geospatial data has fostered many successful stories in spatiotemporal data mining for various urban applications, e.g., venue recommendation [Sun *et al.*2021,Zhang *et al.*2022], and site selection [Liu *et al.*2021]. Among varying types of geospatial data, POIs have gained tremendous momentum, and they have been proven to be effective in many tasks of urban computing and services, such as online map search [Göbel and Kiefer2019, Li *et al.*2020b], population mapping [Ye *et al.*2019, Zhao *et al.*2019], crime prediction [Wu *et al.*2022, Zhang *et al.*2020], and housing price prediction [Xiao *et al.*2017].

While abundant POIs are being generated, the quality of POI data is still questionable. The missing property problem of POIs is prevalent, and it is particularly challenging that the key information of POIs, i.e., categories, is often missing or incorrect due to the uncertainties in the human annotation processes for POIs. Categories are among the most important properties of POIs, which carry valuable semantic information to delineate the human activities that POIs bear [Bing *et al.*2022, Xu *et al.*2023]. Therefore, it is pivotal to complete missing POI categories, to empower POIs to be better used in various downstream tasks.

Several previous studies have investigated the problem of POI semantic annotation with human mobility data. For example, Li et al. [Li *et al.*2020a] capture the similarities among different users' check-in activities to learn the POI features and use a multi-class classifier for semantic annotation. Xu et al. [Xu *et al.*2022a] model the co-occurrences of POIs and categories in check-in sequences to embed them in the same latent space and infer categories for POIs based on the similarity between POI vectors and category vectors. In addition, several works leverage multi-source data such as check-in data, user reviews, and associated images to enhance the effect of POI semantic annotation [Giannopoulos and Meimaris2019, He *et al.*2016]. Such studies have yielded remarkable advancements to tackle the problem of missing or incorrect POI categories. However, they heavily rely on sequential check-in information of users, which can be hardly obtained in reality, especially for those newly created POIs. This raises a further demand for POI semantic annotation with only static POIs, which is the most common type of POIs with smooth acquisition.

In this paper, we focus on the problem of semantic annotation for static POIs. There are primarily two types of information that are pivotal for semantic annotation in our setting. (1) **Spatial information**, i.e., geographic locations of POIs. Intuitively, it is difficult to infer POI semantics from the coordinates of POIs directly. Nevertheless, geographically close POIs tend to be also semantically related, e.g., restaurants tend to appear in clusters. How to effectively leverage such information to understand the environment and background of POIs is a challenging problem. (2) **Textual information**, i.e., POI names. POI names are generally indicative of human

*Corresponding author.

activities taking place at the locations. For example, from the POI name "Master Bao's pastry(鲍师傅糕点)", one could readily understand that the POI is a bakery. Furthermore, many POI names have sparse information that is insufficient to reflect their actual categories. For example, given the POI name "Good neighbors(好邻居)", it is barely possible to derive its actual POI category *Convenience Store* only from this weakly correlated name. Thus the short POI names pose a challenge for modeling POI textual information.

To this end, we propose a Spatial-Textual POI Annotation (STPA) model that overcomes these challenges of limited spatial information and sparse textual information of static POIs. To solve the problem of limited spatial information, STPA uses Delaunay triangulation to build a spatial graph (with POIs being graph nodes) that incorporates spatial contextual information. A graph convolutional encoder is applied to the POI spatial graph to generate POI spatial embeddings. Meanwhile, STPA uses a semantic attention component to alleviate the negative impact of textual information sparsity, which captures the influence of POI textual contexts, i.e., neighboring POIs according to the semantic distance of POI names, and yields the POI textual embeddings. Finally, the two types of POI embeddings are fused to take account of the cross-view interaction for POI semantic annotation.

The contributions of this paper are as follows:

- We propose a Spatial-Textual POI Annotation (STPA) model which recognizes POI semantics from only static POIs with the information of geographic locations and names of POIs. To the best of our knowledge, STPA is the first deep model that simultaneously captures the spatial and semantic correlations of POIs from limited spatial information and sparse textual information to address the problem of semantic annotation for POIs.
- We leverage Delaunay triangulation and graph convolution to model spatial contexts of POIs to generate POI spatial embeddings. Moreover, we design an attention-based text encoder to incorporate the textual knowledge captured from semantic POI neighbors based on POI names into the POI textual embeddings.
- We justify our STPA model on two POI datasets collected from AMap and evaluate its performance via the task of semantic annotation for POIs. STPA demonstrates significant performance gains over several baseline methods based on the paired t-test.

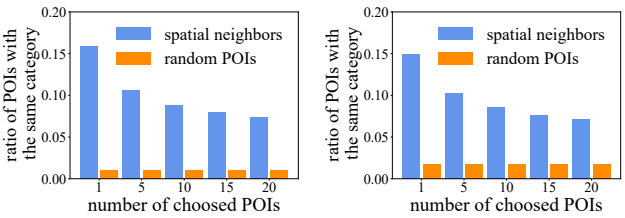
2 Preliminaries

2.1 Problem Statement

[Point of Interest (POI)] The information of a POI p_i is composed of a POI ID, a geographic location (here, latitude and longitude of a POI are known), and a POI name textualized as a bag of words.

[POI Category] A POI category c_i (e.g., *Chinese Restaurant*, *University*) represents the thematic topics of activities that are afforded at the POI p_i.

Semantic Annotation for POIs. Given the POIs' geographical coordinates and names, our goal is to predict the category labels for those unlabeled POIs. Specifically, we divide all POIs into $\mathcal{P}_{train}$ and $\mathcal{P}_{test}$, where $\mathcal{P}_{train}$ contains the POIs with category labels and $\mathcal{P}_{test}$ contains the unlabeled POIs. Our semantic annotation task is to find the category label for each POI $p \in \mathcal{P}_{test}$.



(a) Haidian data (b) Lixia data

Figure 1: Relation between POIs' categories and their spatial neighbors' categories on the Haidian and Lixia data.

2.2 Empirical Data Analysis

The POI data used in this work is collected from AMap[1], where the information of each POI consists of a POI ID, latitude and longitude, a POI name, and a POI category. We select the Haidian District of Beijing and the Lixia District of Jinan as study areas, whereas our semantic annotation method can also be generalized to other study areas. To reduce noise, we filter duplicate POIs by latitude and longitude. After this pre-processing, the Haidian dataset contains 105,577 POIs associated with 248 categories, and the Lixia dataset contains 45,280 POIs associated with 140 categories.

The First Law of Geography tells us that everything is related to everything else, but near things are more related than distant things. Therefore, we conduct the data analysis to investigate whether spatially adjacent POIs share the same category label. Specifically, for each targeting POI p_i, we find its N_k nearest neighboring POIs according to the Euclidean distances with latitude and longitude information and compute the ratio of neighboring POIs which have the same category as p_i. Similarly, we randomly choose N_k POIs from the POI set and compute the ratio of the N_k POIs which have the same category as p_i. Finally, we average the ratios of all the POIs. As shown in Figure 1, we observe that the number of spatially adjacent POIs with the same category as the targeting POI is much more than that of those random POIs on the Haidian and Lixia datasets. This is an important motivation for modeling the spatial vicinity of POIs to improve semantic annotation performance discussed in the following section.

Further, we analyze the POI names of each category. Intuitively, we believe that the words of POI names with the same category are semantically similar. We study the word distributions of categories after removing stop words and punctuation marks of POIs' names. Taking the categories *Tea house* and *Flower Shop* as examples, many POI names of *Tea house* have the keyword "tea(茶)" and those of *Flower Shop* have "flower(鲜花)". Such observations encourage us to leverage the category information of semantically similar POIs to boost the performance of semantic annotation.

[1]https://lbs.amap.com/api/webservice/guide/api/search/

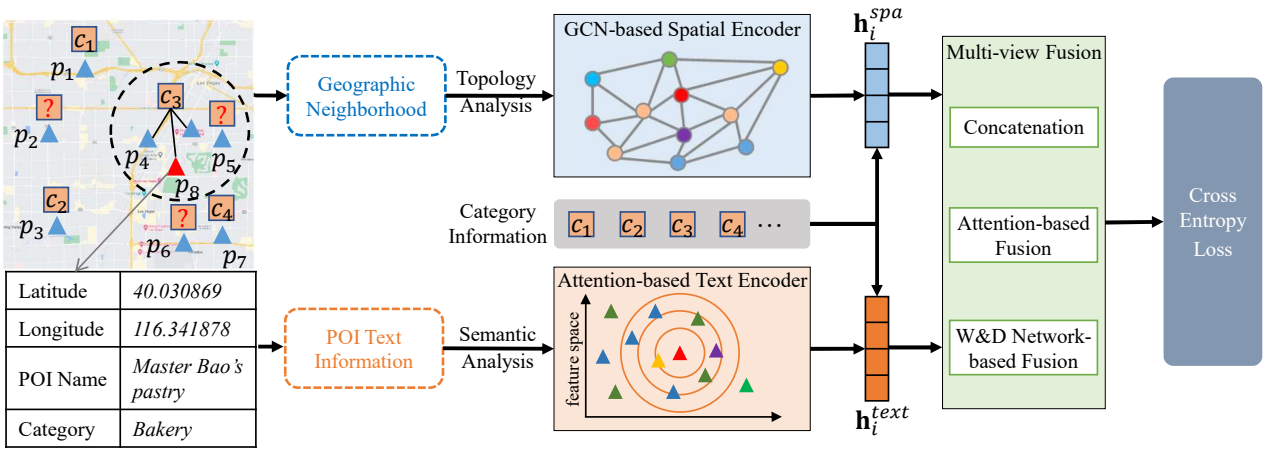


Figure 2: The framework of the proposed STPA model.

3 Spatial-Textual POI Annotation Model

3.1 Model Overview

Figure 2 presents the framework of the proposed method. Let us suppose that we have POIs with latitude and longitude information as well as text names in a study area, where some POIs are labeled with categories and others are unlabeled. For a targeting POI (e.g., p_8), on one hand, we model the spatial correlations of POIs based on the assumption that geographically close POIs tend to have the same/similar category labels. Specifically, we first construct a POI spatial graph based on their geographic locations (latitude and longitude) information, where the POIs are the nodes and the POI category information are the node features; we then aggregate the information from the spatial context via a graph convolution encoder on the spatial graph to generate the POI spatial feature vector $\mathbf{h}^{spa}$. On the other hand, we model the POI name information based on the assumption that POIs with similar textual information (names) are likely to have the same/similar category labels. Specifically, we first discover each POI's semantic neighbors through a pre-training language model and propagate the category information of semantic neighbors to each POI through an attention mechanism. This step produces a textual feature vector $\mathbf{h}^{text}$ for each POI. Finally, the two feature vectors (i.e., $\mathbf{h}^{spa}$ and $\mathbf{h}^{text}$) are fused and the entire model is optimized using the cross-entropy loss.

3.2 GCN-based Spatial Encoder

The structure of the GCN (Graph Convolutional Network)-based spatial encoder is shown in Figure 3. As the category information of POIs generally encodes POI semantics, which serve as important features for semantic annotation, we first generate category embeddings as the initial POI features. Next, we build a spatial graph for all the POIs using the latitude and longitude information based on Delaunay triangulation (DT), where the POIs become the nodes and the category embeddings become the node features. Finally, we apply a graph convolutional encoder on the POI spatial graph to generate POI embeddings, which aggregates the semantic information from POI's spatial contexts.

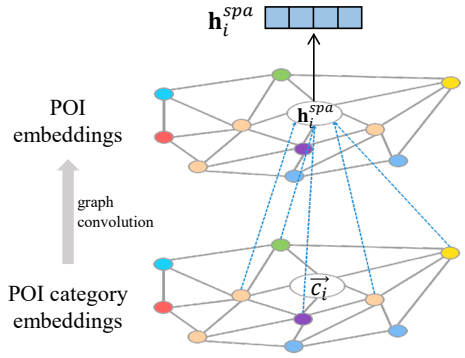


Figure 3: Architecture of the GCN-based spatial encoder.

POI Category Embeddings

POI categories largely reflect POI semantics. We choose the simple one-hot representation of categories as the initial node features in the subsequent graph learning stage, as it provides intuitive semantic label information. That is, the embedding of a category c_i is represented as

$$\vec{c_i} = (0, \cdots, 1, \cdots, 0), \tag{1}$$

where the length of $\vec{c_i}$ is equal to the number of the categories and the ith element of $\vec{c_i}$ is 1.

GCN-based POI Spatial Embeddings

After obtaining the POI category embeddings, we further learn POI embeddings by modeling the spatial contexts. According to the First Law of Geography, we know that spatial adjacency entails a strong correlation among POIs. Adjacent POIs in space are naturally more semantically similar (cf. Figure 1), e.g., restaurants are usually spatially clustered. Therefore, for a POI, we propose to leverage the category information of its spatial contexts (i.e., nearby POIs) to enhance the POI embedding.

To this end, inspired by [Huang *et al.*2023], we model POI spatial relations using a graph structure and utilize the message passing mechanism in GCN to yield POI embeddings via aggregating the category embeddings in a POI neighborhood.

Specifically, we build a spatial graph for POIs based on DT, as the effectiveness of DT graphs for capturing the interactions among spatial vector data has been validated in a series of previous studies [Huang *et al.*2022, Xu *et al.*2022b]. The DT only generates edges between closely proximal (one-hop) POIs to avoid excessive linking of POIs and eliminate noise. Along this line, we connect all the POIs in a study area using the DT to build a spatial graph, where the POIs become the nodes and the corresponding POI category embeddings serve as the node features ($X = \{\vec{c_1}, \vec{c_2}, \cdots, \vec{c_N}\}$). Note that, for a POI p_j without the category label, we average the category embeddings of its neighbor POIs as the initial node feature to ensure that all POI initial embeddings can be assigned anyhow, i.e., $\vec{c_j} = \sum_{i \in \mathcal{N}_j} \vec{c_i}/|\mathcal{N}_j|$, where $\mathcal{N}_j$ is the spatial neighbor POIs of p_j. In the POI graph, we follow [Calafiore *et al.*2021,Huang *et al.*2022] and define the weight of each edge between nodes p_i and p_j as $A_{ij} = \log[(1+L^{1.5})/(1+l_{p_ip_j}^{1.5})]$ to ensure that spatially closer POIs have larger spatial similarities, where L represents the diagonal length of the minimum bounding rectangle containing all the POIs, and $l_{p_ip_j}$ denotes the actual spatial distance between two nodes p_i and p_j. Finally, we normalize the weights into the range from 0 to 1.

Based on the constructed spatial graph, we further apply a one-layer GCN encoder to generate POI spatial embeddings,

$$\mathbf{h}^{spa} = \text{softmax}(D^{-1/2}AD^{-1/2}X\Theta), \tag{2}$$

where A is the weighted adjacency matrix of the POI graph without self-loops, D is the degree matrix of A, X is the initial node features of the graph, and Θ is a linear transformation with learnable parameters. Here we avoid self-loops in the graph, as we cannot utilize POIs' own category information in advance when aggregating neighbor information to generate POI embeddings. After graph convolution, the embedding of each POI encodes the semantic category information from its spatial contexts.

3.3 Attention-based Text Encoder

Figure 4 shows the structure of the Attention-based text encoder. We first find the semantic neighbors for each POI based on the POI names. After that, we design a semantic attention component that weights the influence of the category information of neighboring POIs based on the distance of the textual features between the targeting POI and its neighbors. Finally, it generates the attention-based textual vector $\mathbf{h}_i^{text}$ as the output.

Discovering Semantic Neighbors Based on POI Names

POI names usually contain semantic information related to the POI categories, and POIs with similar semantic names tend to have similar category labels. For example, "Master Bao's pastry(鲍师傅糕点)" and "Master Liang's pastry(梁师傅糕点)") have similar text names and both are associated with the category *Bakery*. Therefore, we first find semantic neighbors with similar POI names for the targeting POI, which could largely mitigate the sparsity issue of POIs' textual information, and is a pivotal factor to facilitate POI semantic annotation.

Considering that the POI names are short texts and contain limited semantic information, we propose to learn features of

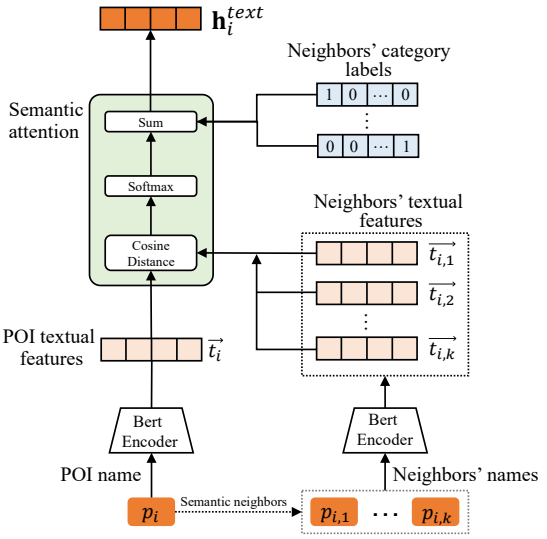


Figure 4: Architecture of the attention-based text encoder.

POI names from both word and phrase levels based on the pre-trained word and phrase vectors from Tencent AI Lab Chinese and English Term Embedding Corpora[2]. Specifically, given a POI, we average the word vectors and the phrase vectors in the POI name respectively and concatenate the two vectors as the POI's initial textual feature. With these initial features, we find the semantic neighbors for each POI based on the cosine similarities.

Semantic Attention

As semantically similar POIs tend to have similar categories, we model this prior with semantic attention by weighting the semantic neighbor POIs based on the features of POI names. As depicted in Figure 4, we use the Bert encoder to learn fine-grained textual features of POI names, and input the textual feature $\vec{t_i}$ of POI p_i and the textual features $(\vec{t_{i,1}}, \cdots, \vec{t_{i,k}})$ of the k most similar POIs to the semantic attention component. The score between the textual features of a POI and its neighbors is calculated based on the Cosine distance,

$$d(\vec{t_i}, \vec{t_{i,j}}) = 1 - \frac{\vec{t_i} \cdot \vec{t_{i,j}}}{\|\vec{t_i}\| \|\vec{t_{i,j}}\|}. \tag{3}$$

Next, we calculate the attention weight for each neighbor POI p_j based on the textual features. Specifically, we feed the distance vector $\vec{D} \in \mathbb{R}^k$ containing the Cosine distance between the targeting POI p_i and its neighbors to a fully-connected layer to generate a hidden vector $\vec{H}$,

$$\vec{H} = \mathbf{W} \cdot \vec{D} + \vec{b}, \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{k \times k}$ is a learnable parameter matrix and $\vec{b}$ is a bias factor. We then apply a softmax function on $\vec{H}$ to obtain the normalized weight for each neighbor POI p_j, i.e., $w(\vec{t_i}, \vec{t_{i,j}}) = e^{\vec{H}_j}/(\sum_{j'=1}^{k} e^{\vec{H}_{j'}})$, where $\vec{H}_j$ means the transformed semantic distance value between the targeting POI p_i and its neighbor POI p_j.

[2]https://ai.tencent.com/ailab/nlp/en/download.html

Finally, we compute the attention-based vector for the targeting POI p_i,

$$\mathbf{h}_i^{text} = \sum_{j=1}^{k} w(\vec{t_i}, \vec{t_{i,j}})\vec{c_j}, \tag{5}$$

where $\vec{c_j}$ is the category embedding of POI p_j introduced in Section 3.2. Therefore, the element of $\mathbf{h}_i^{text}$ is the weighted sum of the category information of the semantic neighbors of p_i, where the attention weights are adaptively learned. Note that, we merely encode the category information of the semantic neighbors into the attention-based vector $\mathbf{h}_i^{text}$, neglecting the textual features $\vec{t_{i,j}}$ of neighbor POIs, as these neighbors' textual features contain implicit semantic information instead of explicit category information. Experimental results demonstrate that concatenating the textual features into $\mathbf{h}_i^{text}$ hinders the annotation performance.

3.4 Multi-view Fusion

As discussed previously, $\mathbf{h}^{spa}$ from the geographical view emphasizes the spatial context of POIs (the First Law of Geography), while $\mathbf{h}^{text}$ from the textual view emphasizes the semantic information about POIs. In addition, the importance of POI feature vectors from different views is still unknown for semantic annotation. Therefore, we propose a multi-view fusion layer to construct a more informative representation that preserves multi-view correlations for semantic annotation. Specifically, we consider three kinds of fusion methods:

- **Concatenation**. We simply concatenate $\mathbf{h}_i^{spa}$ and $\mathbf{h}_i^{text}$ to obtain the fusion representation for POI p_i: $\mathbf{h}_i^c = \mathbf{h}_i^{spa} \oplus \mathbf{h}_i^{text}$.
- **Attention-based Fusion**. We adopt the attention mechanism [Xi *et al.*2022] to adaptively fuse the two vectors. That is,

$$\begin{aligned}
\alpha_i^{spa} &= \vec{w^a} \cdot \tanh(\mathbf{V} \cdot \mathbf{h}_i^{spa} + \vec{b^a}),\\
\alpha_i^{text} &= \vec{w^a} \cdot \tanh(\mathbf{V} \cdot \mathbf{h}_i^{text} + \vec{b^a}),\\
\beta_i^{spa} &= \frac{\exp(\alpha_i^{spa})}{\exp(\alpha_i^{spa}) + \exp(\alpha_i^{text})},\\
\beta_i^{text} &= \frac{\exp(\alpha_i^{text})}{\exp(\alpha_i^{spa}) + \exp(\alpha_i^{text})},\\
\mathbf{h}_i^a &= \beta_i^{spa}\mathbf{h}_i^{spa} + \beta_i^{text}\mathbf{h}_i^{text},
\end{aligned} \tag{6}$$

  where $\vec{w^a}$, $\vec{b^a}$, and $\mathbf{V}$ are learnable parameters.
- **Wide** & **Deep Network-based Fusion**. The Wide & Deep network contains a wide layer and a deep layer [Park *et al.*2019]. In the deep layer, $\mathbf{h}_i^{spa}$ and $\mathbf{h}_i^{text}$ are concatenated and fed to a MLP to yield a deep vector $\mathbf{g}_i$,

$$\mathbf{g}_i = \mathbf{W}^g \cdot (\mathbf{h}_i^{spa} \oplus \mathbf{h}_i^{text}) + \vec{b^g}, \tag{7}$$

  where $\mathbf{W}^g$ and $\vec{b^g}$ denote the weight and the bias of the layer, respectively. In the wide layer, the outer product of $\mathbf{h}_i^{spa}$ and $\mathbf{h}_i^{text}$ is computed and then flattened to a wide vector $\mathbf{u}_i$,

$$\mathbf{u}_i = \mathbf{W}^u \cdot F(\mathbf{h}_i^{spa} \otimes \mathbf{h}_i^{text}) + \vec{b^u}, \tag{8}$$

  where $\mathbf{W}^u$ and $\vec{b^u}$ denote the weight and the bias, respectively. F denotes a flattening function that converts a matrix into a single vector. Later, given a POI p_i, the wide vector $\mathbf{u}_i$ is directly concatenated to the deep vector $\mathbf{g}_i$ to yield the fused vector: $\mathbf{h}_i^{wd} = \mathbf{g}_i \oplus \mathbf{u}_i$.

3.5 Training Objective

After obtaining the fused vector, we directly concatenate it with the pre-trained text vector of the POI name to yield the final POI representation $\mathbf{h}_i^{final}$ for POI p_i, as the POI name of p_i contains the semantic information related to its category label. Next, we feed $\mathbf{h}_i^{final}$ into a three-layer MLP to generate the final output $\hat{\mathbf{y}}_i$, which is the probability that POI p_i associated with each category label and mathematically calculated as follows,

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W_3} \cdot (\mathbf{W_2} \cdot (\mathbf{W_1} \cdot \mathbf{h}_i^{final} + \vec{b_1}) + \vec{b_2}) + \vec{b_3}), \tag{9}$$

where $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{W_3}$, $\vec{b_1}$, $\vec{b_2}$, and $\vec{b_3}$ are parameters.

Finally, we adopt the classification training objective and minimize the cross-entropy loss function,

$$\mathcal{L} = -\frac{1}{|\mathcal{P}|}\sum_{i=1}^{|\mathcal{P}|}\sum_{j=1}^{|\mathcal{C}|} y_{ij} log(\hat{\mathbf{y}}_{ij}), \tag{10}$$

where $|\mathcal{P}|$ and $|\mathcal{C}|$ denote the number of unique POIs and categories in the dataset respectively, y_{ij} denotes whether POI p_i is labeled with the category c_j, and $\hat{\mathbf{y}}_{ij}$ is the predicted probability that p_i is labeled with the category c_j.

4 Experiment

4.1 Experimental Settings

Datasets. We adopt the Haidian and Lixia POI data crawled from AMap as the datasets (cf. Section 2.2). For both datasets, we randomly split them into two collections in the proportion of 8:2 as the training set and test set. We run the model 5 times and report the mean of the results.

Evaluation metrics. Semantic annotation for POIs is a multi-class classification problem. We adopt the two well-known metrics including *Accuracy* and *Macro-F1* to evaluate the performance. In addition, based on $\hat{\mathbf{y}}_i$, we could generate the ranking list of the predicted category labels. Thus we also leverage the metric *MRR* which considers the position of real labels in the ranking lists. Specifically, it is defined as

$$\text{MRR} = \frac{1}{|\mathcal{P}_{test}|}\sum_{i=1}^{|\mathcal{P}_{test}|}\frac{1}{rank_i}, \tag{11}$$

where $|\mathcal{P}_{test}|$ is the number of POIs in the test set $\mathcal{P}_{test}$ and $rank_i$ is the rank of the real category in the predicted list.

4.2 Baselines

We compare the proposed method against the following feature learning baselines of three types.

I. **Textual view methods**

- **Word-based Textual Feature (WTF)**: Based on POI names, we concatenate word and phrase embeddings pre-trained on massive Chinese corpus to yield word-based features for POIs following [Liu *et al.*2020].

Data	View	Method	Accuracy	Macro-F1	MRR
Haidian	Textual	WTF	64.97	59.64	76.18
		ATF	65.59	59.92	76.63
	Spatial	GSF	13.17	4.43	31.10
		GPS2Vec	4.58	0.26	11.79
	Integrated	EHC	65.77*	60.67*	76.88*
		WTF+GPS2Vec	64.65	58.17	75.97
		STPA	**67.73**	**62.69**	**78.26**
		Improvements	2.98	3.33	1.80
Lixia	Textual	WTF	62.60	56.93	74.89
		ATF	63.41	57.63	75.42
	Spatial	GSF	13.54	4.57	24.53
		GPS2Vec	6.33	0.28	15.11
	Integrated	EHC	63.41*	58.03*	75.56*
		WTF+GPS2Vec	62.48	57.03	74.87
		STPA	**65.52**	**60.04**	**77.01**
		Improvements	3.33	3.46	1.92

Table 1: Performance comparison of different methods (in percentage), where the performance improvements of STPA are compared with the best of these baseline methods, marked by the asterisk.

- **Attention-based Text Feature (ATF)**: Given a POI, we use the attention mechanism introduced in Section 3.3 to obtain the weighted textual feature vector of semantic neighbors, and concatenate it with the word-based textual vector as the final feature.

II. **Spatial view methods**

- **Grid-based Spatial Feature (GSF)**: Following [Liu *et al.*2020], we divide the study area into grids and use the TF-IDF transformation to obtain multi-scale geographic features for POIs based on the category distribution in the grids.
- **GPS2Vec**: Following [Yin *et al.*2021], we train a neural network to extract geo-aware features for POIs, which could learn the semantic embeddings from the initial GPS encoding of POIs.

III. **Multiple view methods**

- **EHC**: This is an Ensemble POI Hierarchical Classification framework, which mainly consists of a textual and geographic feature extraction component and a hierarchical classifier [Liu *et al.*2020]. As we do not include the category hierarchy in our model, we adapt EHC with a SVM classifier instead of the hierarchical classifier.
- **WTF+GPS2Vec**: We concatenate the word-based textual features and the geo-aware features learned with GPS2Vec as the fused features of POIs.

These baselines generate multiple types of features of POIs. Based on these POI features, we train a multi-class classifier (i.e., SVM [Chang and Lin2011]) to predict categories for those unlabeled POIs.

4.3 Comparison with Baselines

We report the comparative results in Table 1. It can be observed that:

1) The performance of spatial view methods is unsatisfactory on both datasets, as the limited spatial information of POIs cannot encode enough semantics for POI annotation. GPS2Vec directly uses a MLP network to encode the POI coordinates and performs the worst; GSF leverages the category information of other POIs in the same grid and outperforms GPS2Vec, indicating that spatial context contains useful, not enough though, information for inferring the category of the targeting POI.

2) Textual view methods perform better than spatial view methods, as POI names usually contain richer and more direct semantic information related to POI categories. In addition, ATF has better performance than WTF, as it leverages the information of semantic neighbors, which also proves the effectiveness of the designed attention mechanism.

3) Integrated view methods obtain decent results, and our proposed STPA performs the best. Compared with the best of the baseline methods (EHC), STPA achieves an average improvement of 2.7% on the Haidian data and 2.9% on the Lixia data in terms of the three metrics. Furthermore, the superiority paired t-test results show that the improvement of STPA over these baselines is of practical significance with p value < 0.01.

4.4 Ablation Study

Study of Performances in Different Views

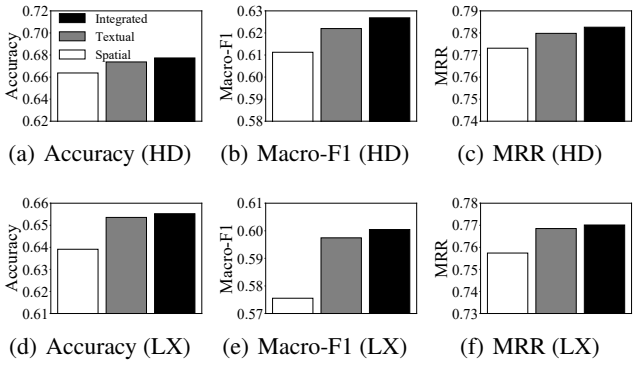
Figure 5 shows the results of two single views and an integrated view on both datasets. We observe that the performances of the textual view are better than those of the spatial view, as POI names contain richer semantic information than geographic locations for our task. Further, it is interesting to observe that the performances of the integrated view are better than those of all the single view methods. Such observations indicate that the integrated view could make up for the deficiency of every single view and capture relatively complete semantics about POIs.

Study of Performances of Different Fusion Methods

Figure 6 shows the results of three kinds of multi-view fusion methods introduced in Section 3.4 on both datasets. We observe that the simple concatenation method outperforms the other two methods. A potential explanation is that the POI features learned with limited spatial information and sparse textual information are relatively simple, and fusing them using a complex function with many parameters may lead to the problem of overfitting.

4.5 Parameter Sensitivity

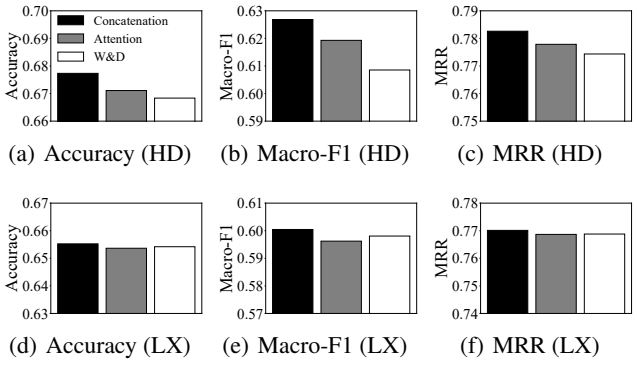
We investigate the sensitivity of results over the number of semantic neighbors (k) in the attention-based text encoder. We increase k from 5 to 40 with a step of 5 and report the results in Figure 7. Evidently, the performances improve when we increase k from 5 to 15, and then remain relatively stable when we increase it further.



(a) Accuracy (HD) (b) Macro-F1 (HD) (c) MRR (HD)

(d) Accuracy (LX) (e) Macro-F1 (LX) (f) MRR (LX)

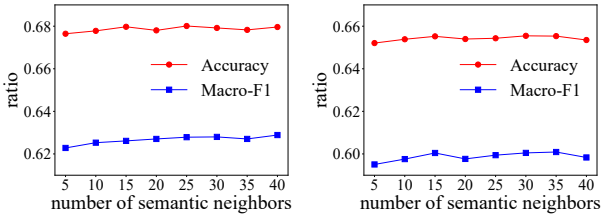
Figure 5: Performance comparison in different views (HD: Haidian data, LX: Lixia data).



(a) Accuracy (HD) (b) Macro-F1 (HD) (c) MRR (HD)

(d) Accuracy (LX) (e) Macro-F1 (LX) (f) MRR (LX)

Figure 6: Performance comparison of different fusion methods (HD: Haidian data, LX: Lixia data).



(a) Haidian Data (b) Lixia Data

Figure 7: Effect of number of semantic neighbors.

5 Related Work

For the problem of POI semantic annotation, traditional methods usually model users' check-ins to construct POI-related features (e.g., distribution of check-in time and the number of check-ins) and train multi-label classifiers [Chang *et al.*2014, Li *et al.*2020a, Ye *et al.*2011]. Further, some studies learn representations of POIs and categories from check-in sequences based on embedding methods and predict the category labels of POIs based on the similarity between POI vectors and category vectors [Wang *et al.*2017, Rahmani *et al.*2019, Xu *et al.*2022a]. In addition, except for the check-in information, some works leverage external semantic information (e.g., users' ratings and reviews) of POIs to recognize POI semantics for better annotation [He *et al.*2016, Giannopoulos and Meimaris2019].

Besides considering the check-in sequential information, some studies model the spatial information of POIs for learning POI feature representations. For example, Yan et al. [Yan *et al.*2017] regard the spatial neighbors of a POI as its spatial context and adopt the word2vec method to model the POI co-occurrence information to learn semantic embeddings for POIs. Liu et al. [Liu *et al.*2020] leverage the category distribution in the neighborhoods and construct the multi-scale geographic features for POIs. Yin et al. [Yin *et al.*2021] present a model named GPS2Vec to extract geo-aware features for venues worldwide. Specifically, they divide the region into fine-grained cells and perform the initial GPS encoding; then they train a network to learn the semantic embeddings for the GPS encoding with geotagged documents (e.g., images and tweets) being the training labels. Huang et al. [Huang *et al.*2022] utilize random walks in a spatial network to capture the spatial contexts of POIs, and a manifold learning algorithm to capture POIs' categorical semantics.

There are also some methods that extract venue-related features from user-generated content (e.g., images) instead of check-ins and make semantic annotation for venues accordingly. For example, Meng et al. [Meng *et al.*2017] model the text-image pairs and predict the categories of venues based on a feature-level fusion method. Zhang et al. [Zhang *et al.*2016] construct features from visual, acoustic, and textual modalities, and label each unseen micro-video with venue categories based on a multi-task multi-modal learning model.

6 Conclusion

In this study, we investigate the problem of semantic annotation for static POIs with only POI names and geographic locations. We present a Spatial-Textual POI Annotation (STPA) model which fully captures the information contained in the locations and names of POIs. Specifically, we design a GCN-based spatial encoder to model the spatial correlations among POIs and generate the POI spatial feature embeddings by using the category information of spatial neighbors; we design an attention-based text encoder that models the POI names and weights the influence of the semantic neighbors' category information to yield the POI textual feature embeddings; we finally construct a more informative fused POI representation that preserves multi-view correlations for semantic annotation. We perform comprehensive experiments using POI data from AMap to demonstrate the effectiveness of STPA. In addition, we observe that textual information is more indicative than spatial information for deriving missing POI categories, while the latter is a useful addition to mitigate the sometimes sparse and weakly correlated textual information.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61906107, the Young Scholars Program of Shandong University. W.H. was supported by the Knut and Alice Wallenberg Foundation.

References

[Bing *et al.*, 2022] Junxiang Bing, Meng Chen, Min Yang, Weiming Huang, Yongshun Gong, and Liqiang Nie. Pre-trained semantic embeddings for poi categories based on multiple contexts. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12, 2022.

[Calafiore *et al.*, 2021] Alessia Calafiore, Gregory Palmer, Sam Comber, Daniel Arribas-Bel, and Alex Singleton. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems*, 85:101539, 2021.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[Chang *et al.*, 2014] Chih-Wei Chang, Yao-Chung Fan, Kuo-Chen Wu, and Arbee LP Chen. On the semantic annotation of daily places: A machine-learning approach. In *Proceedings of the 4th International Workshop on Location and the Web*, pages 3–8, 2014.

[Giannopoulos and Meimaris, 2019] Giorgos Giannopoulos and Marios Meimaris. Learning domain driven and semantically enriched embeddings for poi classification. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, pages 214—-217, 2019.

[Göbel and Kiefer, 2019] Fabian Göbel and Peter Kiefer. Poitrack: Improving map-based planning with implicit poi tracking. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research amp; Applications*, 2019.

[He *et al.*, 2016] Tieke He, Hongzhi Yin, Zhenyu Chen, Xiaofang Zhou, Shazia Sadiq, and Bin Luo. A spatial-temporal topic model for the semantic annotation of pois in lbsns. *ACM Transactions on Intelligent Systems and Technology*, 8(1):1–24, 2016.

[Huang *et al.*, 2022] Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. Estimating urban functional distributions with semantics preserved poi embedding. *International Journal of Geographical Information Science*, 36(10):1905–1930, 2022.

[Huang *et al.*, 2023] Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:134–145, 2023.

[Li *et al.*, 2020a] Yanhui Li, Xiangguo Zhao, Zhen Zhang, Ye Yuan, and Guoren Wang. Annotating semantic tags of locations in location-based social networks. *GeoInformatica*, 24(1):133–152, 2020.

[Li *et al.*, 2020b] Ying Li, Jizhou Huang, Miao Fan, Jinyi Lei, Haifeng Wang, and Enhong Chen. Personalized query auto-completion for large-scale poi search at baidu maps. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(5), 2020.

[Liu *et al.*, 2020] Shaopeng Liu, Jifan Yu, Juanzi Li, and Lei Hou. Geographical information enhanced poi hierarchical classification. In *International Conference on Web Information Systems and Applications*, pages 108–119. Springer, 2020.

[Liu *et al.*, 2021] Yu Liu, Jingtao Ding, and Yong Li. Knowledge-driven site selection via urban knowledge graph. *arXiv preprint arXiv:2111.00787*, 2021.

[Meng *et al.*, 2017] Kaidi Meng, Haojie Li, Zhihui Wang, Xin Fan, Fuming Sun, and Zhongxuan Luo. A deep multimodal fusion approach for semantic place prediction in social media. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 31–37, 2017.

[Park *et al.*, 2019] Donghyeon Park, Keonwoo Kim, Yonggyu Park, Jungwoon Shin, and Jaewoo Kang. Kitchenette: Predicting and ranking food ingredient pairings using siamese neural network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

[Rahmani *et al.*, 2019] Hossein A Rahmani, Mohammad Aliannejadi, Rasoul Mirzaei Zadeh, Mitra Baratchi, Mohsen Afsharchi, and Fabio Crestani. Category-aware location embedding for point-of-interest recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 173–176, 2019.

[Sun *et al.*, 2021] Huimin Sun, Jiajie Xu, Kai Zheng, Pengpeng Zhao, Pingfu Chao, and Xiaofang Zhou. Mfnp: A meta-optimized model for few-shot next poi recommendation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 3017–3023, 2021.

[Wang *et al.*, 2017] Yue Wang, Meng Chen, Xiaohui Yu, and Yang Liu. Lce: A location category embedding model for predicting the category labels of pois. In *Proceedings of the 2017 International Conference on Neural Information Processing*, pages 710–720, 2017.

[Wu *et al.*, 2022] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. Multi-graph fusion networks for urban region embedding. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 2312–2318, 2022.

[Xi *et al.*, 2022] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference*, 2022.

[Xiao *et al.*, 2017] Yixiong Xiao, Xiang Chen, Qiang Li, Xi Yu, Jin Chen, and Jing Guo. Exploring determinants of housing prices in beijing: An enhanced hedonic regression with open access poi data. *ISPRS International Journal of Geo-Information*, 6, 2017.

[Xu *et al.*, 2022a] Haoran Xu, Ronghui Xu, Meng Chen, Yang Liu, and Xiaohui Yu. Cave-sc: Inferring cate-

gories for venues using check-ins. *Information Sciences*, 611:159–172, 2022.

[Xu *et al.*, 2022b] Yongyang Xu, Bo Zhou, Shuai Jin, Xuejing Xie, Zhanlong Chen, Sheng Hu, and Nan He. A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. *Computers, Environment and Urban Systems*, 95:101807, 2022.

[Xu *et al.*, 2023] Ronghui Xu, Meng Chen, Yongshun Gong, Yang Liu, Xiaohui Yu, and Liqiang Nie. Tme: Tree-guided multi-task embedding learning towards semantic venue annotation. *ACM Transactions on Information Systems*, 2023.

[Yan *et al.*, 2017] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2017.

[Ye *et al.*, 2011] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 520–528, 2011.

[Ye *et al.*, 2019] Tingting Ye, Naizhuo Zhao, Xuchao Yang, Zutao Ouyang, Xiaoping Liu, Qian Chen, Kejia Hu, Wenze Yue, Jiaguo Qi, Zhansheng Li, and Peng Jia. Improved population mapping for china using remotely sensed and points-of-interest data within a random forests model. *Science of The Total Environment*, 658:936–946, 2019.

[Yin *et al.*, 2021] Yifang Yin, Ying Zhang, Zhenguang Liu, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. Gps2vec: Pre-trained semantic embeddings for worldwide gps coordinates. *IEEE Transactions on Multimedia*, 24:890–903, 2021.

[Zhang *et al.*, 2016] Jianglong Zhang, Liqiang Nie, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat Seng Chua. Shorter-is-better: Venue category estimation from micro-video. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 1415–1424, 2016.

[Zhang *et al.*, 2020] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 4431–4437, 2020.

[Zhang *et al.*, 2022] Lu Zhang, Zhu Sun, Ziqing Wu, Jie Zhang, Yew Soon Ong, and Xinghua Qu. Next point-of-interest recommendation with inferring multi-step future preferences. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 3751–3757. International Joint Conferences on Artificial Intelligence Organization, 2022.

[Zhao *et al.*, 2019] Yuncong Zhao, Qiangzi Li, Yuan Zhang, and Xin Du. Improving the accuracy of fine-grained population mapping using population-sensitive pois. *Remote Sensing*, 11(21), 2019.