Applied Intelligence (2023) 53:8286–8301
https://doi.org/10.1007/s10489-022-03858-w

STaTRL: Spatial-temporal and text representation learning for POI recommendation

Xinfeng Wang[1,2] · Fumiyo Fukumoto[3] · Jiyi Li[3] · Dongjin Yu[1] · Xiaoxiao Sun[1]

Accepted: 6 June 2022 / Published online: 27 July 2022

Abstract
With the rapid development of location-based social networks (LBSNs), point-of-interest (POI) recommendations have become a practical problem attracting more and more attention. Recent studies mostly utilize contextual features and sequential patterns of users' check-ins to recommend POIs. However, there are still many deficiencies in existing works, such as: (1) insufficiently learning relations among far-apart visits in user check-ins; (2) not effectively incorporating geographical information when modeling user-POI interactions; and (3) little exploiting the features from reviews for the POI recommendation task. To tackle the above problems, we propose spatial-temporal and text representation learning (STaTRL), which employs Transformer to learn long-term dependencies among visits in the check-ins sequence and adopts an improved approach to compute the attention between visits by applying geographical information to the self-attention layer in Transformer. Meanwhile, users' perspectives and POIs' reputations learned from textual reviews are explored to improve the performance. In addition, a multi-task objective framework is adopted to simultaneously train the hidden representations of users' historical check-ins trajectories which are shared by these two tasks. Concretely, STaTRL consists of (1) the principal task, i.e., *unvisited POI recommendation* that recommends to users the unvisited POIs, and (2) the auxiliary task, i.e., *user's POI preference learning* whose candidates include both visited and unvisited POIs. We found that the latter task helped train the embedding of visited POIs and further boosted the performance of the former task, and lacking any of both would decline the performance. Extensive experiments on three public datasets demonstrated that STaTRL vastly outperformed the state-of-the-art methods.

Keywords POI recommendation · Aspect Based Sentiment Analysis (ABSA) · Multi-task · Spatial-temporal and text representation

1 Introduction

With the exponential growth of location-based social network services, large amounts of geo-tagged data such as Foursquare and Yelp, have become available making it possible for users to share their experiences and check-in information. Check-in histories generated by users help understand users' POI preferences, and thus, provide important clues to unvisited POI recommendations. Likewise, mining an individual user's preferences obtained from POI popularity learned from other users' trajectories can help decide the user's next movement. Furthermore, the reputation of POI itself is inevitable for recommendation tasks. People would be curious to know which POIs have a good reputation, e.g., delicious food restaurants, or hotels with great services. Meanwhile, people may have several POI candidates or do not have any candidates to whom they want to go. With POI popularity obtained from all visitors, people can make their travel plans with minimal effort, even in a short time in an unfamiliar city. Exploring user preferences and POI reputations through reviews, therefore, has a significant impact on POI recommendations.

However, unlike traditional collaborative filtering-based recommendation systems, POI recommendation has major obstacles for mining users' preferences. One is that POI data often exhibit complicated short- and long-term temporal dependencies. For example, there always exist short-term dependencies between a hospital and a pharmacy, as we often go to take our prescription to the pharmacy after receiving it in the hospital. In contrast, a favorite

✉ Fumiyo Fukumoto
fukumoto@yamanashi.ac.jp

Extended author information available on the last page of the article.

barber' shop indicates long-term dependencies on check-in histories. Moreover, users visit a small number of POIs against the millions of POIs in LBSNs; as Liu et al. pointed out. the density of the POI check-in count matrix generated from the experimental data is around 0.1% [1]. These problems hamper the accurate prediction of users' preferences. To address the above problems, several authors have attempted to utilize matrix factorization (MF) [2, 3] or markov chain (MC) for recommendations [4, 5]. Nevertheless, complex non-linear user-POI interactions are difficult to model sufficiently by MF and MC-based approaches. More recently, many researchers have attempted to apply deep learning techniques, including recurrent neural networks (RNNs) [6]. They have made great progress in short-term sequences. However, every two visits in a user's check-in trajectory may be separated by hundreds of visits, or may be separated by several weeks, even several months, resulting in their failure to achieve satisfactory performances.

The other issue is that POI recommendations are affected by diverse contextual information such as spatial influence, temporal context, social influence, and textual reviews [7]. A assuming that similar users would visit similar POIs, the attempts included geographical information [8–10], temporal relations [11], both spatial-temporal relations [12–14], social relations [15, 16], categorical information [17], and textual semantics [8, 18]. However, their common shortcoming is considering only one or two types of contextual information in one entity, which makes the models lack extensibility [19].

Inspired by the previous work mentioned above, we assume that the latent representation of users' POI preferences learned from check-in trajectories and contextual information from texts can help predict user' behavior patterns, and propose a STaTRL which is illustrated in Fig. 1. By leveraging visited POIs information from users' check-in records, STaTRL learns POI popularity, i.e., spatial-temporal information, and learns a user's preferences of target POIs. Specifically, it employs a POI embedding technique to learn the historical check-ins for the complicated short- and long-term temporal dependencies between every two check-ins in historical sequence by leveraging Transformer [20]. The self-attention modules with Transformer directly model dependencies among POIs by assigning attention scores. A high score between two POIs indicates a strong dependency, while a low score implies a weak dependency. By utilizing multi-head attention modules consisting of several self-attention modules, STaTRL can adaptively assign weight scores to POIs that are at any temporal distance from the current POI. This indicates that the model can capture short and long-term dependencies between any two POIs. Moreover, to enhance the self-attention between

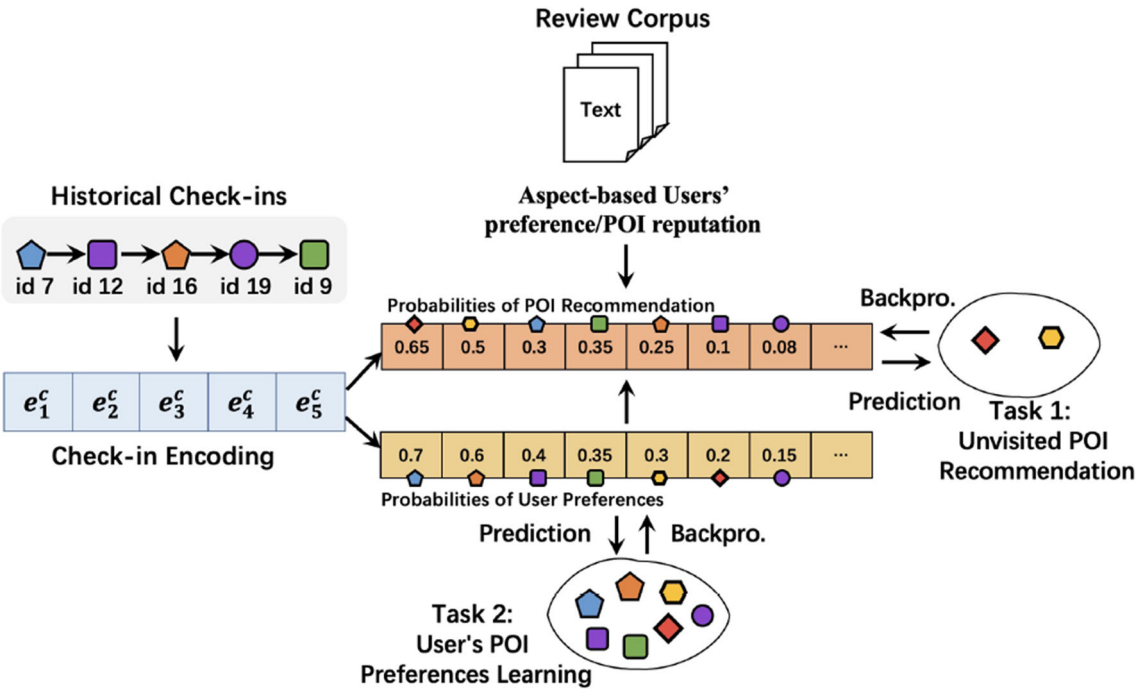


Fig. 1 Illustration of principal and auxiliary tasks

all POIs in the historical visit sequence, geographical information, namely, the physical distance between POIs, is fed into the improved self-attention layer as auxiliary features.

As shown in Fig. 1, in addition to spatial-temporal information, STaTRL also leverages text information to learn user preferences and POI reputations. More specifically, it utilizes aspect based sentiment analysis (ABSA) that identifies the sentiment polarity related to the target aspect in reviews [21–23]. The motivation is that the sentiment polarity related to a specific aspect can precisely capture the user's opinion, which is also beneficial to learn the user's preferences and POI reputation. For example, a review sentence of "The food in the restaurant was very good but I felt the service was lacking", indicates the aspect of "food" is associated with positive sentiment, while the aspect of "service" is negative. Thus, it can capture the restaurant's reputation from these aspect-based sentiments and recommend it if the user prefers delicious food. In contrast, if the user is concerned about the service, the restaurant should not be recommended.

STaTRL employs a multi-task learning framework. As illustrated in Fig. 1, STaTRL consists of two tasks: Task 1: *unvisited POI recommendation* that recommends users to the unvisited POIs, and Task 2: an auxiliary task, *user's POI preferences learning* whose candidates include both visited and unvisited ones. The assumption is that Task 2 helps to learn not only the embedding of visited POIs but also to search for more relative POIs, while Task 1 only considers the relations between visited and unvisited POIs for each specific user to discriminate users' personal preferences. Particularly, the representation of user trajectories, which includes the embedding of visited POIs and the relations between visited POIs and users' preferable POIs, are mutually restricted and promoted by the two tasks. Therefore, the model adopts the multi-task objective function and is trained to simultaneously recommend the top-k POIs for the target user and learn the user's POI preferences. Our main contributions are summarized as follows.

- We propose a novel technique based on Transformer to tackle the complicated short- and long-term temporal dependencies between POIs in the historical visit sequence. Specifically, we apply geographical information to the self-attention layer of Transformer to learn short- and long-term dependencies between POIs.
- We leverage the sentiment polarity related to the target aspect in the reviews to capture each user's personal preferences and POI reputation from all visitors, and incorporate these as an extra constraint into check-in sequence to enhance the performance of POI recommendation.
- We adopt a multi-task learning framework consisting of the user's POI preferences learning and unvisited POI recommendation tasks. We utilize the former as the auxiliary task to help tune the embedding of visited POIs to further boost the performance of POI recommendations.
- We conducted extensive experiments on three public datasets, and STaTRL achieved significant improvement over seven state-of-the-art methods. In particular, the main results on the Yelp-2020 dataset show an improvement over the second best methods by 8.9% $\sim$ 23.4% in P@k and 13.2% $\sim$ 19.7% in R@k for k=5, 10, and 20.

The remainder of the paper is structured as follows. After discussing related work in Section 2, the preliminaries and the problem definitions are introduced in Section 3. Section 4 presents the prediction model in detail. The experimental results and discussion are provided in Section 5. Finally, Section 6 concludes the paper and outlines future work.

2 Related work

The POI recommendation is important research in recommendation systems and it has attracted extensive attention in both academic and industrial circles. Because memory-based collaborative filtering and matrix factorization-based methods can only capture the linear relationships between users and POIs, plenty of deep learning approaches have been proposed to solve the problem.

Recently, inspirited by word embedding [24] in the natural language processing (NLP) field, many researchers have attempted an embedding technique to capture users' various aspects of preference over POIs [8, 25, 26]. Furthermore, many researchers have expanded it by incorporating contextual information with POI embedding. Zhao et al. proposed a temporal POI embedding model based on the contextual check-in sequences and the various temporal characteristics on different days, e.g., users always check-in at POIs around offices on a weekday, whereas they visit shopping malls on weekends [27]. He et al. proposed a context-aware POIs embedding model [28]. Their method injects POI popularity and user preferences obtained from a probabilistic model into co-occurring POIs embedding, while our approach directly utilizes POI embeddings to predict POI popularity and users' preferences. In addition, several technologies for processing language have been employed to expand embedding-based models. Qian et al. focused on spatial-temporal contexts and embedded both users and POIs into the same space, called a transition space, and operated users and POIs by utilizing translation vectors [13]. Chang et al. leveraged not only temporal and spatial information, but also high-frequency words

that appeared in the text content collected from Instagram [8], and Liao et al. utilized the word-level encoder and decoder long short-term memory (LSTM) model to extract the sentiment expressed in each review [29]. Although Liao et al.'s approaches are conceptually similar to our work in terms of utilizing sentiment information, our work differs from their approaches in that we focus on the sentiment polarity, related to the target aspect in the review to retrieve users' opinions, precisely which is beneficial to learn users' preferences and POI reputation.

Similar to POI embedding techniques, the attention mechanism is widely applied to many tasks including recommendation tasks. Earlier attempts were made by [30, 31]. The authors utilize vanilla attention vectors to model the influence of items, while these recommendation models only focus on a single aspect of the item's importance. Zhong et al. alleviate the problem by utilizing multi-head self-attention to learn users' various aspects preferences over POIs [32]. However, their approach learns spatial-temporal contextual aspects independently by employing different learning frameworks, such as attention mechanisms and graph convolutional networks (GCNs) [33]. In contrast, we inject geographical relations among POIs into the POIs embedding technique based on Transformer to learn spatial-temporal contexts simultaneously. Similarly, Luo et al. proposed a spatio-temporal attention network (STAN), a bi-attention architecture for personalized item frequency [34]. The first layer of STAN aggregates all relevant locations from the user trajectory for updated representation, so that the second layer can match the most plausible candidates from weighted representations.

Meanwhile, self-attention-based methods exhibit their great ability to capture the importance between POIs. Ma et al. proposed an autoencoder-based model, SAE-NAD, which consists of two components: a self-attentive encoder (SAE) to adaptively compute an importance vector for each POI in a user's check-in histories and a neighbor-aware decoder (NAD) to incorporate the geographical information by computing the inner product of POI embeddings together with the radial basis function (RBF) kernel [35]. More related to our work is Zuo et al.'s model, the transformer Hawkes process (THP), which leverages the self-attention mechanism of Transformer [36]. Our approach is related to the attention mechanism and differs from the THP in that the THP utilizes an element-wise plus operation to obtain the attention output. In contrast, we compute element-wise multiplying the location similarity matrix with the sublayer output to scale up dependencies because users are more likely to visit close POIs.

In the context of text analysis, since the sentiments in the reviews help understand users' opinions and POI reputations, this makes it possible to mine underlying preferences from the prediction reviews to recommend POIs. This line of research includes rating prediction based on the analysis of reviews [37]. Yu et al. proposed a model to learn more precise latent factors of users and items by combining users' sentiments in review texts and their rating scores [38]. Cheng et al. proposed a method for modeling aspect-aware latent factors that can effectively combine reviews and ratings for rating prediction [39]. Similarly, Shen et al. focused on the inconsistency between ratings and textual reviews and proposed a reliability measure to improve the performance of recommender systems [40].

Recently, deep semantic analysis of text, i.e., ABSA was intensively applied to mine and summarize users' opinions from textual data. It aims to identify sentiment polarity toward the target aspect in the underlying reviews. Many researchers have attempted to apply bidirectional encoder representations from transformers (BERT) [24] on ABSA tasks [41–43] and gained amazing performances. We thus employ A Lite BERT (ALBERT) [44] pre-training language model, a variant of BERT, and fine-tune it to analyze sentiment polarity for the target aspect in reviews.

In the context of multi-task learning (MTL), the recent upsurge of MTL techniques have contributed to improving the performance on POI recommendation. Zhang et al. proposed an interactive MTL framework for noisy data and uncertain check-ins [45]. Halder et al. proposed a multi-task, multi-head attention transformer model to simultaneously recommend the next top-k POIs to users and predict the prospective queuing time of recommended POIs [46]. By utilizing multi-head attention, the model can integrate long-term dependencies among any POIs, while their approach does not take into account the location similarity between POIs, which has been verified to have a great impact on POI recommendation [8, 10, 47]. Xia et al. attempted to utilize generative adversarial networks (GAN) based on LSTM that simultaneously consider temporal check-ins and geographical locations [48]. Their experimental results on two medium-scale datasets on LBSNs showed that their approach performed well compared with the six baselines. Unlike the previous studies mentioned above, we assume that users' latent preferences by learning POI popularity from users' check-in records, the geographical information of POIs and POI reputations mined from reviews can support predicting the user's next movement.

3 Preliminaries

Table 1 presents the notations used in this work.

Definition 1. (Point-of-interest(POI)) *A POI* $y \in \mathcal{Y}$ *is defined as a uniquely identified location that has geographical coordinates, i.e., longitude and latitude.*

Table 1 Key notations used in this work

Notation	Definition
$\mathcal{U}, \mathcal{Y}, \mathcal{R}$	Set of users, Set of POIs, Set of reviews
$\mathcal{S}_u$	A user u visit sequence
$\mathcal{R}_u$	A user u review sequence
$\mathcal{G}$	The POI location graph
$\mathcal{A}$	Set of aspects $\mathcal{A}$ = {Food, Price, Service}
$\mathbf{E}^t$	Concatenation of check-in time embeddings $\mathbf{e}_i^t$
$\mathbf{E}^p$	Global embedding matrix of $\mathcal{Y}$
$\mathbf{E}^c$	Check-in embedding matrix of $\mathcal{S}_u$
$\mathbf{Y}$	Sequence of POIs $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_L\}$ in $\mathcal{S}_u$
$\mathbf{O}$	Location similarity matrix for all POIs $\mathcal{Y}$
$\mathbf{O}_u$	Location similarity matrix for POIs in $\mathcal{S}_u$
$\mathbf{H}$	Hidden representation of $\mathcal{S}_u$
$\mathbf{pu}$	Probability vector of user's POI preferences
$\mathbf{pr}$	Probability vector of POI recommendation
$\boldsymbol{\psi}$	Representation of user's aspect preferences
$\mathbf{W}_1^{FC}, \mathbf{b}_1, \mathbf{W}_2^{FC}, \mathbf{b}_2$	Trainable weights in feed forward layer
$\mathbf{W}_3^{FC}, \mathbf{b}_3$	Trainable weights for user's preferences prediction
$\mathbf{W}_4^{FC}, \mathbf{b}_4, \lambda, \mu$	Trainable weights for POI recommendation

Definition 2. (check-in) *A check-in* $C = (u, t, y, r)$ *shows a user* $u \in \mathcal{U}$ *has visited a POI* $y \in \mathcal{Y}$ *and leaves a review* $r \in \mathcal{R}$ *at time* t.

Definition 3. (POI location graph) *The POI location graph* $\mathcal{G} = (\mathcal{Y}, \mathbf{O})$ *shows the geographical location network between POIs, where* $\mathcal{Y}$ *denotes a set of POIs.* $\mathbf{O} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ *refers to a location similarity matrix and each element* $o_{i,j} \in \mathbf{O}$ *shows the location similarity between* y_i *and* y_j.

Task 1 (Unvisited POI Recommendation) Given a user $u \in \mathcal{U}$ visit sequence $\mathcal{S}_u = \{(t, y)_k\}_{k=1}^L$, where L refers to the number of POIs in the sequence, a review sequence $\mathcal{R}_u = \{(t, r)_k\}_{k=1}^L$, and a POI location graph $\mathcal{G}$, the goal is to generate a list of top-k POIs with the highest probabilities of unvisited POI candidates from $\mathbf{pr} \in \mathbb{R}^{|\mathcal{Y}|}$ for the target user u.

Task 2 (User's POI preference prediction) Given a visit sequence $\mathcal{S}_u$ and a POI location graph $\mathcal{G}$, the goal is to predict the probability of user preference for all POIs $\mathbf{pu} \in \mathbb{R}^{|\mathcal{Y}|}$.

4 Methodology

The STaTRL framework is illustrated in Fig. 2. The approach integrates geographical relations among POIs into check-in POI embeddings to learn the hidden representations, which are co-trained by both tasks. We call this module a spatial-temporal representation learning module. STaTRL utilizes ABSA in review texts to match users' opinions and POIs' reputations. We call this a text representation learning module. The last is the multi-task learning module consisting of (1) *unvisited POI recommendations*, and (2) *users' POI preferences learning*.

4.1 Spatial-temporal representation learning

Our spatial-temporal representation learning module utilizes self-attention from Transformer, a family of non-auto regressive models (NAR). For the set of POIs $\mathcal{Y}$, we obtain the global embedding matrix $\mathbf{E}^p \in \mathbb{R}^{d_m\times|\mathcal{Y}|}$, where d_m refers to the dimension of embedding and the i-th column of $\mathbf{E}^p$ indicates the $|\mathcal{Y}|$-dimensional embedding vector for POI $\mathbf{y}_i$. Subsequently, for each POI $\mathbf{y}_k$ from the historical visit sequence $\mathcal{S}_u$, we denote its embedding vector to $\mathbf{E}^p\mathbf{y}_k$, where $\mathbf{y}_k \in \mathbb{R}^{|\mathcal{Y}|}$ is a one-hot vector. Thus, the check-in POI embedding matrix $\mathbf{E}^c \in \mathbb{R}^{d_m\times L}$ of $\mathcal{S}_u$ is given by

$$\mathbf{E}^c = (\mathbf{E}^p\mathbf{Y} + \mathbf{E}^t)^\top \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_L] \in \mathbf{R}^{d_m\times L}$ denotes the POIs in $\mathcal{S}_u$. The kth column of $\mathbf{E}^p \in \mathbb{R}^{d_m\times L}$ shows the embedding of the specific POI $\mathbf{y}_k$ in $\mathcal{S}_u$. $\mathbf{E}^t = [\mathbf{e}_1^t, \mathbf{e}_2^t, ..., \mathbf{e}_L^t] \in \mathbb{R}^{d_m\times L}$ refers to the concatenation of check-in time embeddings, $\mathbf{e}_k^t$ which is obtained by the temporal encoding technique provided by THP.

To incorporate geographical relations among POIs and learn the location similarity for all POIs, we leverage graph structure and inject it into Transformer, which we call the Geo-Transformer. More specifically, following the structure THP (STHP) [36], we also construct a POI geographic location network $\mathcal{G} = (\mathcal{Y}, \mathbf{O})$, where $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{|\mathcal{Y}|}\}$

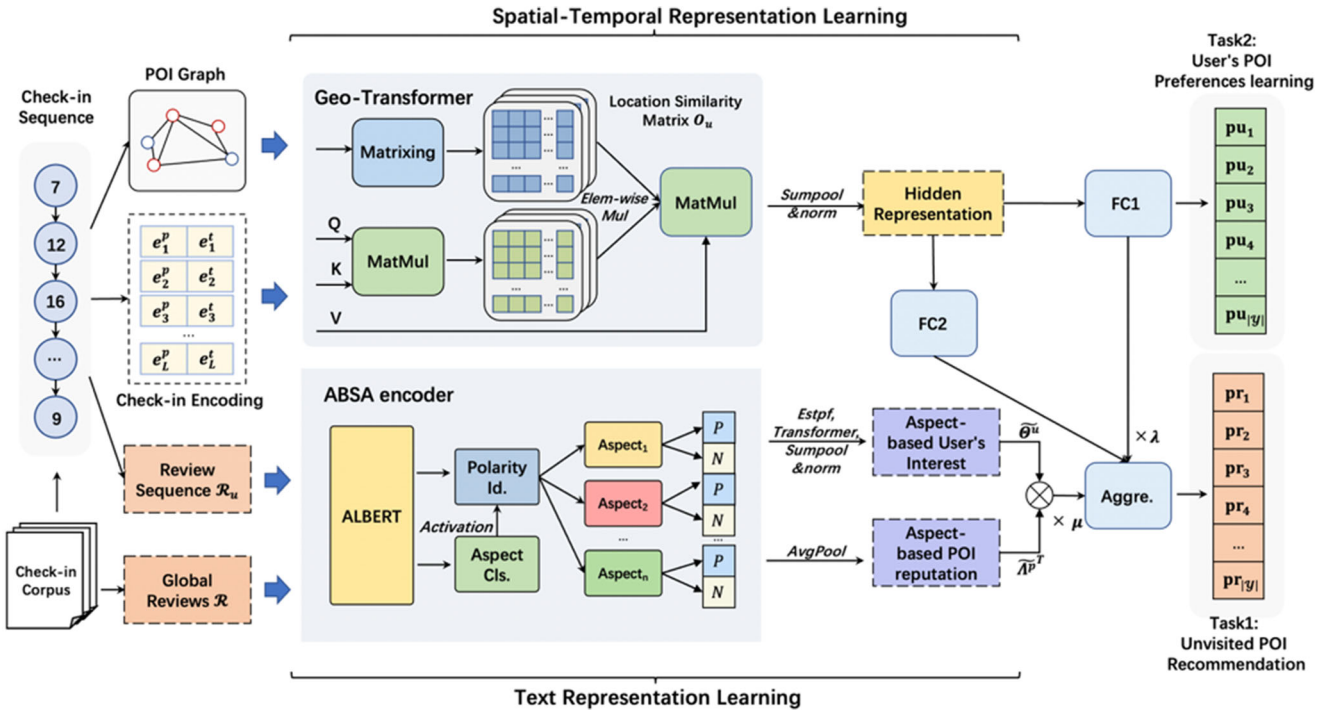


Fig. 2 Architecture of STaTRL framework

shows a set of POIs, $\mathbf{O} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ indicates the location similarity matrix, and each element $o_{i,j} \in [0, 1]$ denotes the location similarity between $\mathbf{y}_i$ and $\mathbf{y}_j$. We employ a Gaussian radial basis function (RBF) kernel [32] which is given by

$$o_{i,j} = \begin{cases} \exp(-\eta\, d_{i,j}), & \text{if } \mathbf{y}_i \neq \mathbf{y}_j \\ 1, & \text{if } \mathbf{y}_i = \mathbf{y}_j \end{cases} \tag{2}$$

where $d_{i,j}$ shows the distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ computed by the Haversine formula on latitude and longitude. $\eta > 0$ is a hyper-parameter to control the correlation strength between $\mathbf{y}_i$ and $\mathbf{y}_j$. Note that the closer the distance $d_{i,j}$, the larger the value of $o_{i,j}$. Besides, $o_{i,j}$ is set to 0 if it is less than a threshold value τ. We employ an improved self-attention mechanism to learn the check-in POI embeddings of the historical visit sequence $\mathcal{S}_u$. More specifically, we computed the attention output $\mathbf{S} \in \mathbb{R}^{d_m \times L}$:

$$\begin{aligned} \mathbf{S} &= \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top \odot \mathbf{O}_u}{\sqrt{d_k}})\mathbf{V} \\ \mathbf{Q} &= \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V \end{aligned} \tag{3}$$

where $\mathbf{O}_u \in \mathbb{R}^{L\times L}$ indicates the location similarity matrix for POIs in $\mathcal{S}_u$. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices obtained by different transformations of $\mathbf{X}$, respectively, and $\mathbf{W}^Q$, $\mathbf{W}^K \in \mathbb{R}^{d_m \times d_k}$ and $\mathbf{W}^V \in \mathbb{R}^{d_m \times d_v}$ are weight matrices for linear transformations. The initial input of $\mathbf{X}$ is $\mathbf{E}^c$.

Different from STHP, which utilizes element-wise plus operation, we element-wise multiply the location similarity matrix $\mathbf{O}_u$ with the sublayer output $\mathbf{Q}\mathbf{K}^\top$ to capture the feature that users are more likely to visit close POIs. In such a case, we deem location similarity as the weight to scale up dependencies rather than that both dependencies and location similarity between two POIs make an equal contribution to learning the features. Figure 3 illustrates this operation. Given a check-in sequence that is shown in (a), each entity $[\mathbf{Q}\mathbf{K}^\top]_{i,j}$ of the sublayer output $\mathbf{Q}\mathbf{K}^\top$ denotes the dependencies between POIs $\mathbf{y}_i$ and $\mathbf{y}_j$ which is illustrated in (b). Given the POI geographical location graph $\mathcal{G}$ shown in (c), we obtain the enhanced relations (d) by using a multiply operation.

Multi-head self-attention modules can be stacked to learn high-level representations that each POI has complicated similarities with other POIs. It is beneficial for multiple-context-aware trajectory data. The multi-head attention output $\mathbf{M}$ for the POI sequence is given by

$$\mathbf{M} = [\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_z]\mathbf{W}^M \tag{4}$$

where z represents the number of heads in the multi-head attention, $\mathbf{W}^M \in \mathbb{R}^{z\cdot d_v \times d_m}$ refers to an aggregation matrix, and $\mathbf{S}_i \in \mathbb{R}^{d_m \times L}$ denotes the output of the head. The attention output $\mathbf{M}$ is fed into the position-wise feed-forward neural network and pre-hidden representation $\mathbf{F}$ of the input sequence is given by

$$\mathbf{F} = \text{ReLU}(\mathbf{M}\mathbf{W}_1^{FC} + \mathbf{b}_1)\mathbf{W}_2^{FC} + \mathbf{b}_2 \tag{5}$$

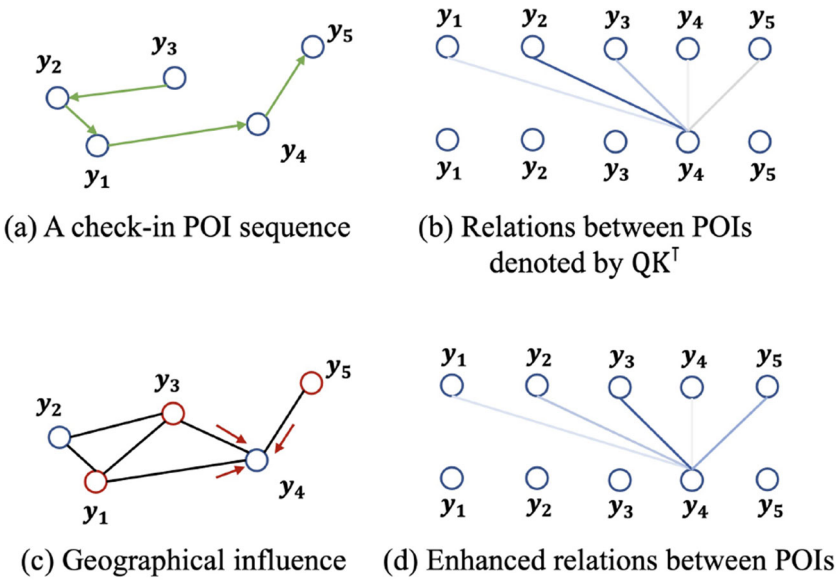


Fig. 3 Illustration of geographical influence

where $\mathbf{W}_1^{FC} \in \mathbb{R}^{d_m \times d_h}$, $\mathbf{W}_2^{FC} \in \mathbb{R}^{d_h \times d_m}$, $\mathbf{b}_1 \in \mathbb{R}^{d_h}$, $\mathbf{b}_2 \in \mathbb{R}^{d_m}$ are learnable parameters of the neural network, and d_h is the dimension of $\mathbf{W}_1^{FC}$, $\mathbf{W}_2^{FC}$, and $\mathbf{b}_1$.

After performing the Geo-Transformer layers for N times, we obtain $\mathbf{F} \in \mathcal{R}^{d_m \times L}$ which includes spatial-temporal representations for all the check-in POIs in the input sequence. Sum pooling and batch normalization are applied to $\mathbf{F}$ to obtain hidden representations $\mathbf{H} \in \mathcal{R}^{d_m}$ of the user's behavioral patterns.

4.2 Text representation learning

We also leverage sentiment polarity related to the target aspect in the underlying reviews to predict users' interest and POI reputation. As shown in Fig. 2, to obtain the sentiment features under each review $r \in \mathcal{R}$, we utilized the ALBERT [44] pre-training language model and fine-tuned the model for the ABSA task using the 2020 Yelp dataset. The model is independently fine-tuned from the other modules of our framework.

4.2.1 Fine-tuning ALBERT for ABSA

We introduced a hierarchical classification approach to identify sentiment polarity related to the target aspect as sentiment polarity and the target aspects are different categories, i.e., positive or negative and food, price, or service. Thus, we first classify a textual review into several aspects, i.e., food, price, and service, and then assign each to being either positive or negative. More specifically, as shown in Fig. 2, our ABSA encoder consists of two steps: *aspect classification* (*Aspect Cls.*), and *polarity identification* (*Polarity Id.*) for the specific aspect. The input of the encoder is a textual review r, and the output is an aspect-based sentiment embedding, a vector $\Lambda \in \mathbb{R}^{2|\mathcal{A}|}$ assigning a value of 0 or 1 for each polarity element of each aspect $a \in \mathcal{A}$.

We employ an ALBERT pre-training model [44] for each review r, and obtain the review text embedding $\mathbf{e}^r \in \mathbb{R}^{d_r}$ where d_r denotes the dimension of the review text embedding. Because of the restriction of text length and embedding size, we used the first 512-word tokens in the input review and the 768-dimension of the hidden layer.

In the first step, aspect classification, sentiment embedding $\mathbf{e}^r$ is passed to the multi-aspect classifier, which is a multi-layer perceptron (MLP). For the result, we applied a sigmoid function for the multi-label classification problem to obtain the probability value $\mathbf{p}_{a_i}^{asp} \in \mathbb{R}^2$ for the ith aspect $a_i \in \mathcal{A}$. Here, we set a threshold value α, and the ith aspect a_i, whose probability score exceeds α value is considered for identifying its aspect. After we obtained $\mathbf{p}_{a_i}^{asp}$ for each aspect a_i from $\mathbf{e}^r$, we concatenate them, which is denoted by $\mathbf{p}^{asp} \in \mathbb{R}^{|\mathcal{A}|}$. Each aspect is assigned in the first step, and is further identified polarity. We used the softmax function for this single label classification task, and obtained the polarity probability $\mathbf{p}_{a_i}^{pol}$ of the ith aspect a_i, and $\mathbf{p}_{a_i}^{pol}[0]$, $\mathbf{p}_{a_i}^{pol}[1] \in [0,1]$ are the values of negative and positive for aspect a_i. We then create the polarity vector for each aspect which is given by:

$$\widetilde{\mathbf{p}}^{pol} = \begin{cases} [0,1] & \mathbf{p}^{pol}[0] < \mathbf{p}^{pol}[1] \\ [1,0] & \mathbf{p}^{pol}[0] > \mathbf{p}^{pol}[1] \\ [0,0], & \text{otherwise} \end{cases} \tag{6}$$

Finally, we concatenate them and obtain the aspect-based sentiment embedding $\Lambda \in \mathbb{R}^{2|\mathcal{A}|}$ for review r which is given by:

$$\Lambda = [\widetilde{\mathbf{p}}_1^{pol}, \widetilde{\mathbf{p}}_2^{pol}, ..., \widetilde{\mathbf{p}}_{|\mathcal{A}|}^{pol}] \tag{7}$$

The training objective is to minimize the following loss:

$$\mathcal{L}^{asp}(\theta^{(AL)}, \phi^{(asp)}) = -(\boldsymbol{\gamma}_{a_i}^{asp})^{\top} \log([\mathbf{p}_{a_i}^{asp}])$$
$$\mathcal{L}^{pol}(\theta^{(AL)}, \phi^{(pol)}) = -(\boldsymbol{\gamma}_{a_i}^{pol})^{\top} \log([\mathbf{p}_{a_i}^{pol}]) \quad (8)$$

where a_i represents the target aspect, $\boldsymbol{\gamma}_{a_i}^{asp}$ indicates the one-hot vector of true aspect a_i. $\mathbf{p}_{a_i}^{asp}$ shows the probability value of all aspects. Similarly, $\boldsymbol{\gamma}_{a_i}^{pol}$ is the true label vector and $\mathbf{p}_{a_i}^{pol}$ indicates the probability value for the polarity. $\theta^{(AL)}$ refers to the parameters used in ALBERT, $\phi^{(asp)}$ denotes the specific parameters for estimating the probability of each aspect and $\phi^{(pol)}$ stands for the parameters for identifying the polarity of a_i.

4.2.2 Predicting aspect-based user's preference and POI reputation

So far, we have obtained the aspect-based sentiment embedding Λ from the review r by the ABSA model. We apply the procedure to the historical POI reviews $\mathcal{R}_u$ for user u and obtain $\Lambda^u \in \mathbb{R}^{L\times 2|\mathcal{A}|}$. We also apply the procedure to all of the reviews $\mathcal{R}$ related to the target POI and obtain $\Lambda^p \in \mathbb{R}^{|\mathcal{R}|\times 2|\mathcal{A}|}$ for extracting POI reputation. Similar to our spatial-temporal representation learning, we leverage the results to recommend users to the unvisited POI. On the one hand, when the polarity related to some aspect of a given target entity is positive, we can infer that the user is interested in the entity. On the other hand, when polarity is negative, the user is interested in the opposite manner. For example, in the sentence, "The food here is very expensive.", we can infer that the user prefers "not expensive food" as an entity "food" related to its aspect, "price" is negative. From the observation, we transfer the polarity of the negative state to a positive one to estimate users' personal preferences. We call this procedure, *an estimation of users' preferences regarding the target aspect* (Estpf), which is given by

$$\Theta^u = \text{Estpf}(\Lambda^u) \quad (9)$$

where $\Theta^u \in \mathbb{R}^{L\times 2|\mathcal{A}|}$ indicates the transferred aspects vector for user u. In contrast, we used the original assignment by ABSA when we captured the reputation for each POI. Note that Θ^u shows multi-aspect polarities in historical reviews for user u. As illustrated in Fig. 2, we then apply Transformer, sum pooling, and batch Normalization to Θ^u and obtain the representation of the user's interest in the aspects $\widetilde{\Theta^u} \in \mathbb{R}^{1\times 2|\mathcal{A}|}$. Similarly, the POI reputation $\widetilde{\Lambda^p} \in \mathbb{R}^{|\mathcal{Y}|\times 2|\mathcal{A}|}$ is computed by applying the average pooling operation for each POI candidate Λ^p. Finally, we obtain the representation of the aspect-based user's preference $\boldsymbol{\psi} \in \mathbb{R}^{|\mathcal{Y}|}$ which is given by:

$$\boldsymbol{\psi} = \widetilde{\Theta^u} \cdot \widetilde{\Lambda^p}^{\top} \quad (10)$$

4.3 Multi-task learning

Recall that our auxiliary task is the spatio-temporal context-based user preference prediction. Given a visit sequence $\mathcal{S}_u$ and a set of POIs $\mathcal{Y}$, we obtain hidden representations $\mathbf{H}$ of behavioral patterns. We passed it to a fully connected layer and applied a sigmoid function, as the user's POI preferences learning is a multi-label classification, where each label refers to each POI in a set $\mathcal{Y}$. The probability vector of the user's preferences $\mathbf{pu} \in \mathbb{R}^{|\mathcal{Y}|}$ is given by

$$\mathbf{pu} = \text{sigmoid}(\mathbf{H}\mathbf{W}_3^{FC} + \mathbf{b_3}) \quad (11)$$

where $\mathbf{W}_3^{FC} \in \mathbb{R}^{d_h\times|\mathcal{Y}|}$ and $\mathbf{b_3} \in \mathbb{R}^{d_h}$ are trainable parameters. The training objective is to minimize the following cross-entropy loss:

$$\mathcal{L}_{pu}(\theta) = -(\boldsymbol{\gamma}_{pu})^{\top} \log(\text{pu}) \quad (12)$$

where $\boldsymbol{\gamma}_{pu} \in \mathbb{R}^{|\mathcal{Y}|}$ refers to a true label vector. Each dimension of the vector equals 1 when the user has ever visited the POI, otherwise 0. The principal task, an unvisited POI recommendation is obtained by hidden representation $\mathbf{H}$, the aspect-based user preference $\boldsymbol{\psi}$ and the probability of the user's preferences $\mathbf{pu}$ which is given by:

$$\mathbf{pr} = \text{sigmoid}(\mathbf{H}\mathbf{W}_4^{FC} + \mathbf{b_4}) + \lambda\mathbf{pu} + \mu\boldsymbol{\psi} \quad (13)$$

where $\mathbf{pr} \in \mathbb{R}^{|\mathcal{Y}|}$ represents the probability of a POI recommendation, $\mathbf{W}_4^{FC} \in \mathbb{R}^{d_h\times|\mathcal{Y}|}$, $\mathbf{b_4} \in \mathbb{R}^{d_h}$, λ and μ are trainable parameters. The training objective is to minimize the following loss:

$$\mathcal{L}_{pr}(\theta) = -(\boldsymbol{\gamma}_{pr})^{\top} \log(\text{pr}) \quad (14)$$

where $\boldsymbol{\gamma}_{pr} \in \mathbb{R}^{|\mathcal{Y}|}$ denotes the true label vector, i.e., each dimension of the vector equals 1 when we recommend that the user visit, otherwise 0. We assume that the auxiliary user's POI preferences learning helps POI recommendation. The model adopts a multi-task objective function which is defined by:

$$\mathcal{L}^{(multi)}(\theta^{(sh)}, \phi^{(pu)}, \phi^{(pr)}) = \mathcal{L}^{(pu)}(\theta^{(sh)}, \phi^{(pu)}) + \mathcal{L}^{(pr)}(\theta^{(sh)}, \phi^{(pr)}) \quad (15)$$

where $\theta^{(sh)}$ indicates the shared parameters, and $\phi^{(pu)}$ and $\phi^{(pr)}$ stand for a parameter estimated in the user's preferences prediction and unvisited POIs recommendation, respectively. Algorithm 1 illustrates the training process of the STaTRL model. After we obtained the probability

vector of unvisited POI recommendation **pr** for user u, we recommend top-k POIs with the highest probabilities.

Algorithm 1 Training STaTRL.

Require: a POI location graph $\mathcal{G}$, a visit sequence $\mathcal{S}_u$, a review sequence $\mathcal{R}_u$ and a review set $\mathcal{R}$

Ensure: the probability vector of user's POI preferences **pu** and that of the unvisited POI recommendation **pr**

Set the number of epoch $epoch_{encoder}$, threshold value of probability $\alpha = 0.5$, and the value of the best accuracy $acc_{best} = 0$
for $epoch$ in (1, 2, ..., $epoch_{encoder}$) **do**
  for r in $\mathcal{R}$ **do**
    Obtain text embedding of r
    Calculate the probability for each aspect $p_{a_i}^{asp}$
    if $p_{a_i}^{asp} > \alpha$ **then**
      Calculate the polarity probability of the target aspect a_i
      Optimize parameters by gradient descent via (11)
    end if
  end for
  Calculate acc_{valid} in valid dataset
  if $acc_{valid} > acc_{best}$ **then**
    save parameters of encoder
    $acc_{best} \leftarrow acc_{valid}$
  end if
end for
Calculate average ratings for POIs by encoder through $\mathcal{R}$
Set the number of epoch for training the main framework $epoch_{main}$ and $acc_{best} = 0$
for $epoch$ in (1, 2, ..., $epoch_{main}$) **do**
  Calculate the vector **pu** via (11)
  Calculate the vector **pr** via (13)
  Optimize parameters by gradient descent via (15)
  Calculate acc_{valid} in valid dataset
  if $acc_{valid} > acc_{best}$ **then** save parameters of framework $acc_{best} \leftarrow acc_{valid}$
  end if
end for

5 Experiment

In this section, we first introduce the experimental setup, including the comparative methods. We then report the performance comparison on the dataset with textual reviews to examine the effectiveness of our method. Finally, we show the performance comparison on two datasets that are widely utilized as benchmark datasets in LBSNs.

5.1 Experimental setup

5.1.1 Dataset and evaluation metrics

We performed the experiments on three real-world datasets: the Gowalla [49], Yelp [1], and Yelp-2020[1] datasets. The first two datasets have been widely used as benchmarks in LBSN research. The Gowalla data were collected from February 2009 to October 2010. The Yelp data were obtained from the Yelp dataset challenge round 7. Note that most of the LBSNs datasets including these two do not provide textual information. We therefore used Yelp-2020, version Mar. 2020, to evaluate our full model. Yelp-2020 consists of a large volume of check-ins. The experimental setup is the same as other related works [1, 50]. More specifically, we randomly selected 40% of the check-ins and used them as the experimental dataset. We filtered out users who visited fewer than 10 POIs and POIs with fewer than 10 visitors. Each dataset was divided into three folds, i.e., 70% of the check-ins (oldest) and 10% (more recent) were used as the training and development sets, respectively. The remaining 20% (newest) was used as the test set. Table 2 presents the statistics on the datasets. In the Yelp-2020 dataset, the number of reviews is the same as that of check-ins, i.e.,735,092 reviews.

In addition, we used the sentiment analysis dataset taken from the SemEval-2014 Task 4 challenge [51] to train the ABSA model. We used a restaurant domain consisting of 3,041 sentences. Yelp reviews are classified into hierarchical domains. We utilized 14 domains in the top level of a hierarchy including a restaurant domain. These domains are *active*, *arts*, *auto*, *beautysvc*, *education*, *fitness*, *health*, *hotelstravel*, *nightlife*, *pets*, *physicians*, *professionals*, *restaurants*, and *shopping*. Each domain can be classified into two domains. One has three types of aspects, i.e., food, service, and price. The other has service and price aspects. For each aspect, we randomly chose sentences from Yelp reviews and manually annotated 100 sentences for positive or negative polarity. As a result, we used training data consisting of a total of 8,041 sentences. The data was divided into two sets, i.e., a training and development set with 70%, and 30%, respectively.

The performances of POI recommendation methods were evaluated by Precision at k(P@k) and Recall at k (R@k), where k is set to 5, 10, and 20. Furthermore, we evaluated our ABSA model by randomly selecting 700 reviews from the Yelp-2020 reviews. We manually annotated these reviews and used them as evaluation data. We used precision, recall, and F1-score as the evaluation metrics[2].

[1] https://www.yelp.com/dataset

[2] alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools

Table 2 Statistics of check-in dataset. "Avg. Len." and "Max. Len." denote the average and maximum length of check-in sequences, respectively

Dataset	#User	#POI	#Check-in	#Review	Avg. Len.	Max. Len.
Yelp–2020	28,038	15,745	735,092	735,092	24.1	979
Yelp challenge	30,887	18,995	860,888	–	26.6	906
Gowalla	18,737	32,510	1,278,274	–	39.6	876

5.1.2 Baselines

We compared STaTRL with state-of-the-art methods[3]. These methods are classified into three: first, self-attention-based methods that exhibit their great ability to capture the importance of POIs; second, a multiple context-based method that considers multiple contexts, including temporal, geographical, and social contexts, and finally, MF-based methods which are some of the most popular and traditional recommendation techniques.

Self-attention-based methods:

- **Transformer Hawkes process (THP)** [36]: a point process approach which leverages self-attention to capture short and long-term dependencies.
- **Transformer Hawkes process without intensive function (THP-If)** [36]: the THP without the intensity function, which enhances the relation strength between adjacent POIs for only predicting the next POI.
- **Spatio-temporal attention network (STAN)** [34]: an attention mechanism-based method, which explicitly incorporates spatiotemporal correlation within a trajectory to learn the relevance between non-adjacent locations and non-contiguous visits.
- **Self-attentive encoder and neighbor-aware decoder (SAE-NAD)** [35]: a novel autoencoder-based model that consists of a self-attentive encoder and a neighbor-aware decoder for POI recommendations.

Multiple context-based method:

- **Spatio-temporal activity-centers prediction (STACP)** [52]: a joint geographical and temporal modeling approach based on matrix factorization.

MF-based methods:

- **Local geographical logistic matrix factorization (LGLMF)** [53]: a logistic MF method based on geographical information to leverage a user's personal, geographic profile, and a location's geographic popularity.
- **Contextualized point-of-interest recommendation (CPIR)** [50]: a method that utilizes MF on user and POI similarity to recommend top-k POIs to users.

[3]We downloaded the authors' source code provided by their URL.

5.1.3 Implementation and parameter settings

We utilized the ALBERT-base-v2 model[4] as a pre-training model of ABSA. The search ranges of the hyperparameters in our model were as follows: The dimension of check-in encoding d_m was searched in $\{128 \times i\}_{i=4}^{16}$. The number of dimensions, d_h, d_k and d_v were searched in $\{128 \times i\}_{i=4}^{8}$. The hyper-parameter η in the RBF kernel, the threshold value τ of location similarity, and the threshold value α of the aspect probability were searched in [0, 1]. The learning rate of STaTRL and that of the ABSA encoder were searched in [1e-5, 1e-4], and {1e-6, 5e-6, 1e-5, 5e-5, 1e-4}, respectively. The number of the Geo-Transformer layers N was searched in {1, 2, 3, 4, 5}, and the number of heads in multi-attention z was searched in {4, 8, 12, 16, 20}. After tuning, we utilized the hyperparameters with the best performance on the validation set as follows: d_m is 1,024 for the Yelp-2020 and Yelp challenge, and 2,048 for the Gowalla dataset. d_h, d_k and d_v were set to 1,024, 1,024, and 896, respectively. The η, τ and α were set to 0.37, 0.14 and 0.5, respectively. The learning rates of the STaTRL and ABSA encoders were 1e-4 and 1e-6, respectively. The number of layer N of the Geo-Transformer was 2, and the number of heads z was 12. These hyperparameters were tuned using Optuna [5]. The parameters for all the baselines were tuned to attain the best performance or set as proposed by the authors. The experiments were conducted using Pytorch on Nvidia GeForce RTX 3090 (24GB memory).

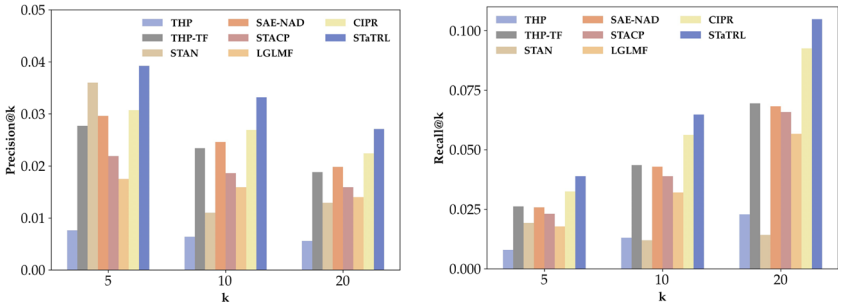
5.2 Results

5.2.1 Performance comparison on Yelp-2020

Recall that Yelp-2020 consists of a large volume of check-ins with textual reviews. Thus, we used the data to evaluate our full model. Figure 4. shows the performance comparisons of STaTRL model and baselines. First, STaTRL achieved the best performance on all evaluation metrics, specifically, the P@5 value obtained by our STaTRL performed better than the second-best method, STAN, by 8.9%. Similarly, STaTRL achieved an improvement over the second-best method, CPIR, by 23.4%, 21.0%, 19.7%,

[4]huggingface.co/transformers/pretrained.models.html

[5]https://github.com/pfnet/optuna



(a) Precision@k on Yelp-2020 dataset (b) Recall@k on Yelp-2020 dataset

Fig. 4 Performance comparison with the baselines on Yelp-2020 dataset

15.3%, and 13.2% in P@10, P@20, R@5, R@10, and R@20, respectively. This shows that the latent representation of users' preferences learned from diverse contexts helps improve the POI recommendation task.

Second, THP produced the worst results with the except of STAN with R@10 and R@20, although the method leverages the self-attention mechanism. This indicates that the conditional intensity function which interpolates two observed time stamps gives a negative effect capturing two POIs located far in the time interval.

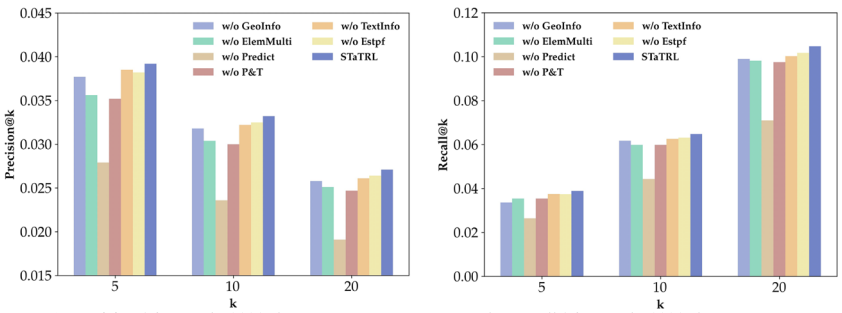
Similarly, STAN did not work well except for the result by P@5, while the approach was also based on the attention mechanism. This is reasonable because the architecture of STAN is designed for the next POI recommendation, i.e.,to recommend the next POI for users at a specific time given users' historical check-in data, leading to poor performance. The result by P@5 shows significant performance compared with other baselines and also shows the evidence. SAE-NAD works well in attention-based methods, but it is worse than our STaTRL. This indicates that a multi-dimensional attention mechanism can predict more POIs with rich structured data, but to predict more accurate POI visit sequences, we need to leverage more variety of contextual information obtained from the data.

Third, the results obtained by the multi-context-based approach, i.e., STACP, indicate that geographical, social, and spatial-temporal contexts helped improve the performance, while it did not achieve the best performance. One possible reason is that the method utilizes kernel estimation or matrix factorization techniques, while our STaTRL model based on the Transformer uses a POI encoder that can effectively capture the semantic meaning of contexts. The observation that the performance obtained by THP-If utilizing the POI embedding technique is better than their methods also supports this. Similarly, MF-based methods achieve better results, while StaTRL captures latent users' behavior patterns more effectively.

5.2.2 Ablation study

We performed an ablation study to empirically examine the impact of each module used in the STaTRL. The results are shown in Fig. 5. The caption "w/o P&T" indicates that the model recommended unvisited POIs by utilizing only the POI encoder. "w/o Estpf" shows the results that we directly utilize the result of ABSA for the user's historical sequence. "w/o ElemMulti" denotes that not element-wise multiplication but element-wise plus was applied to incorporate the geographical information.

We can see from Fig. 5 that the user's POI preference learning task was most effective, as "w/o Predict" exhibits the worst performance in all evaluation metrics. The



(a) Precision@k on Yelp-2020 dataset (b) Recall@k on Yelp-2020 dataset

Fig. 5 Ablation experiment: "w/o X" denotes the removed components, i.e., geographical information (GeoInf), user's preferences prediction (Predict), both user's preferences prediction & text information (P&T) by ABSA, Text Information by ABSA (TestInfo), and estimation of user's preference regarding the target aspect (Estpf), respectively. Similarly, "w/o ElemMulti" indicates that not element-wise multiply but element-wise plus operation is applied

Table 3 Performance of ABSA model

Aspect	Positive			Negative		
	Recall	Precision	F1	Recall	Precision	F1
Food	0.836	0.893	0.864	0.770	0.807	0.788
Price	0.856	0.868	0.862	0.697	0.723	0.710
Service	0.747	0.849	0.794	0.767	0.814	0.790

second worst performance on "w/o P&T" indicates that the POI embeddings obtained by the check-in sequence worked well in the user's POI preferences learning task and were more effective compared with geographical information, i.e.,location similarity between every two POIs, and text information by ABSA. The results obtained by "w/o GeoInf" were slightly worse than those of "w/o TextInfo" and "w/o Estpf". This shows that the injection of location similarity into Transformer is effective compared with ABSA. We note that the results obtained by "w/o ElemMulti" are worse than our full model, STaTRL, and even worse than the results obtained by a model without geographical information, i.e., "w/o GeoInf". This proves our conjecture that geographical distance provides important clues for users to decide their next movement, and that our model utilizing an element-wise multiply operation is able to capture geographical features well.

Table 3 shows the ABSA result for each aspect. F1 refers to the Micro-F1 score. Overall, the results of "food" and "price" aspects were better than those of "service", and the results that were correctly assigned as positive were better than those with a negative sentiment polarity for all aspects. From the results of the error analysis, we observed that the sentiment polarity toward the specific aspects in the underlying review often appears in the first and the last parts of that review. However, we used only the first 512-word tokens in the input review sentences because of the input restriction of ALBERT. As a result, the ALBERT could not correctly identify the polarity for the aspect. There were 38.3% of the errors that were classified into this type. Beltagy et al. addressed the limitation that Transformer-based models, including ALBERT, are unable to process long sequences, due to their self-attention operation, and proposed a method called Longformer [54]. Its mechanism is a drop-in replacement for self-attention and combined locally windowed attention with task-motivated global attention, which makes it easy to process documents of thousands of tokens or longer. Similarly, Kitaev et al. attempted to improve the efficiency of transformers by utilizing dot-product attention and reversible residual layers [55]. There is considerable room for further improvement of ABSA by incorporating these models.

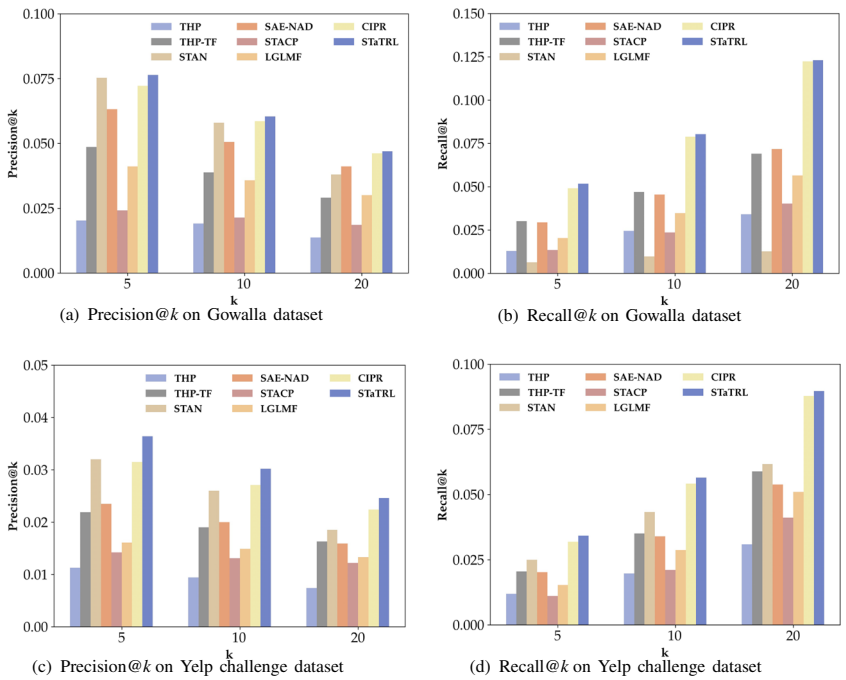


Fig. 6 Performance comparison with the baselines on Gowalla and Yelp challenge dataset

(a) Precision@k on Gowalla dataset

(b) Recall@k on Gowalla dataset

(c) Precision@k on Yelp challenge dataset

(d) Recall@k on Yelp challenge dataset

5.2.3 Performance comparison on Yelp challenge and Gowalla datasets

We note that throughout the ablation study, the user's POI preferences learning module plays a crucial role in STaTRL, leading to a significant improvement of STaTRL over the state-of-the-art POI recommendation methods in the Yelp-2020 dataset. We thus conducted experiments using the Yelp challenge and Gowalla benchmark datasets that are widely utilized as benchmark datasets in LBSNs to examine the effectiveness of our STaTRL model without text representation, i.e., the model without text information by the ABSA model. The evaluation results for the two datasets are summarized in Fig. 6.

As shown in Fig. 6, STaTRL without text representation outperformed all the baselines by both metrics and the two datasets. More precisely, in the Gowalla dataset, our model obtained an improvement over all the baselines by 1.5% ∼276.4%, 3.1% ∼ 216.2% and 1.7% ∼ 243.1%, in P@5, P@10, and P@20, respectively, and by 5.3% ∼ 720.6%, 1.8% ∼ 727.8% and 0.5% ∼ 868.5% in R@5, R@10, and R@20, respectively. Similarly, on the Yelp challenge dataset, our model obtained an improvement over all the baselines by 13.8% ∼ 222.1%, 11.4% ∼ 221.3% and 9.8% ∼ 232.4% in P@5, P@10, and P@20, respectively, and by 7.2% ∼ 208.1%, 4.2% ∼ 186.8%, and 2.2% ∼ 190.3% in R@5, R@10, and R@20, respectively. These observations clearly support the effectiveness of our STaTRL model.

5.2.4 Computational complexity

The training procedure of our approach consists of two stages: STaTRL framework and ABSA encoder. The computational complexity of the former is determined by the Geo-Transformer, i.e., $O(n^2)$, where n refers to the length of each sequence, and its computational complexities are significantly larger than those of other modules such as the fully connected layers and normalization layers. The computational complexity of the latter depends on that of the ALBERT-base-v2 model[6] and it is $O(n^2)$, where n denotes the number of words in each review.

Table 2 shows the statistics of the check-in sequences from the real-world datasets used in our experiments. In each dataset, the average time for training the main framework in a total of 30 epochs using Pytorch on Nvidia GeForce RTX 3090 are 72.3, 76.8, and 115.8 min, each. Similarly, it takes a total of 6.8 min to fine-tune the encoder in 10 epochs, which is acceptable in our current experiments.

[6] huggingface.co/transformers/pretrained.models.html

6 Conclusion

In this article, we proposed an unvisited POI recommendation method by learning spatial-temporal and text representation. Our STaTRL model based on Transformer learns POI popularity by leveraging visited POIs from users' check-in records, and predicts user preferences. It adopts multi-task learning and is trained to simultaneously recommend top-k POIs and predict user preferences. The experimental results showed that STaTRL outperforms the state-of-the-art POI recommendation methods on three real-world datasets: Gowalla, Yelp challenge, and Yelp-2020. Moreover, throughout the ablation study, we demonstrated that the users' POI preferences learning helps improve the unvisited POI recommendation task. We will open our source code and datasets to facilitate the comparison of future studies on POI recommendations.

There are several interesting directions for future research. We should be able to obtain further advantages in the efficacy, especially, our ASBA model with ALBERT. Recall that we utilized the first 512 word tokens in the input review and 768-dimension of the hidden layer, as ALBERT restricts text length and embedding size because of the time complexity, which may lead to a lack of contextual information about sentiment polarity related to the aspects. Various attempts have been made to reduce the overall self-attention complexity [54–57]. We would improve our ASBA model by incorporating its pre-training models to process reviews consisting of long token sequences.

We would also incorporate more context information, such as scores of aspect, queuing time at POIs [46], and social relations, into the model to further improve POI recommendation performance.

Acknowledgements This work is supported by the Grant-in-aid for JSPS, Grant Number 21K12026, and Suzuki foundation.

CRediT authorship contribution statement **Xinfeng Wang**: Conception and design, Data collection, Software, Analysis and interpretation of results, Writing- original & editing. **Fumiyo Fukumoto**: Conception and design, Analysis and interpretation of results, Writing- review draft, Supervision. **Jiyi Li**: Analysis and interpretation of results, Writing- review draft. **Dongjin Yu**: Writing- review draft, Supervision. **Xiaoxiao Sun**: Writing- review draft

Declarations

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest ; and expert testimony or patent-licensing arrangements, or non-financial interest (such as personal or professional

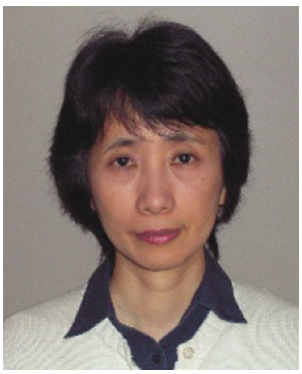
relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

1. Liu Y, Pham T-AN, Cong G, Yuan Q (2017) An experimental evaluation of point-of-interest recommendation in location-based social networks. In: Proc. of the VLDB Endowment, vol 10, pp 1010–1021
2. Chen C, Liu Z, Zhao P, Zhou J, Li X (2018) Privacy preserving point-of-interest recommendation using decentralized matrix factorization. In: Proc. of the AAAI Conference on artificial intelligence, vol 32
3. Rahmani HA, Aliannejadi M, Baratchi M, Crestani F (2020) Joint geographical and temporal modeling base on matrix factorization for point-of-interest recommendation. Advances in Information Retrieval, p 205–219
4. Zhang Y, Shi Z, Zuo W, Yue L, Liang S, Li X (2020) Joint personalized markov chains with social network embedding for cold-start recommendation. Neurocomputing 386:208–220
5. Liu S, Wang L (2018) A self-adaptive point-of-interest recommendation algorithm based on a multi-order markov model. Futur Gener Comput Syst 89:506–514
6. Li G, Chen Q, Zheng B, Yin H, Nguyen QVH, Zhou X (2020) Group-based recurrent neural networks for poi recommendation, ACM/IMS Trans Data Sci 1(1)
7. Zhao P, Zhu H, Liu Y, Xu J, Li Z, Zhuang F, Sheng VS, Zhou X (2019) Where to go next: A spatio-temporal gated network for next poi recommendation. AAAI, p 5877–5884
8. Chang B, Park Y, Park D, Kim S, Kang J (2018) Content-aware hierarchical point-of-interest embedding model for successive poi recommendation. IJCAI, p 3301–3307
9. Wang H, Shen H, Ouyang W, Cheng X (2018) Exploiting poi-specific geographical influence for point-of-interest recommendation. IJCAI, p 3877–3883
10. Liu W, Wang Z-J, Yao B, Yin J (2019) Geo-alm: Poi recommendation by fusing geographical information and adversarial learning mechanism. IJCAI 7:1807–1813
11. Yao Z (2018) Exploiting human mobility patterns for point-of-interest recommendation. In: Proc. of the Eleventh ACM International conference on web search and data mining, pp 757–758
12. Cui Y, Sun H, Zhao Y, Yin H, Zheng K (2021) Sequential-knowledge-aware next poi recommendation: A meta-learning approach. ACM Trans Inf Syst (TOIS) 40(2):1-22
13. Qian T, Liu B, Viet Q, Nguyen H, Yin H (2019) Spatiotemporal representation learning for translation-based poi recommendation. ACM Trans Inf Syst 37(2):1–24
14. Lim N, Hooi B, Ng S-K, Wang X, Goh YL, Weng R (2020) Stp-udgat: Spatial-temporal-preference user dimensional graph attention network for next poi recommendation. In: Proc. of the 29TH ACM International Conference on Information and Knowledge Management (CIKM'20), pp 845–855
15. Gao Q, Trajcevski G, Zhou F, Zhang K, Zhong T, Zhang F (2018) Trajectory-based social circle inference. In: Proc. of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp 369–378
16. Zhou F, Yue X, Trajcevski G, Zhong T, Zhang K (2019) Context-aware variational trajectory encoding and human mobility inference. In: Proc. of World Wide Web Conference, pp 3469–3475
17. Xing S, Liu F, Wang Q, Zhao X, Li T (2019) Content-aware point-of-interest recommendation based on convolutional neural network. Appl Intell 49(3):858–871
18. Werneck H, Santos R, Silva N, Pereira AdrianoCM, Mourão F, Rocha L (2021) Effective and diverse poi recommendations through complementary diversification models. Expert Syst Appl 175:114775
19. Han P, Li Z, Liu Y, Zhao P, Li J, Wang H, Shang S (2020) Contextualized point-of-interest recommendation. In: Proc. of the 29th International conference on artificial intelligence, pp 2484–2490
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems 30
21. Mowlaei ME, Abadeh MS, Keshavarz H (2020) Aspect-based sentiment analysis using adaptive aspect-based lexicons. Expert Syst Appl 148:113234
22. Peng H, Xu L, Bing L, Huang F, Lu W, Si L (2020) Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In: Proc. of the AAAI Conference on Artificial Intelligence, vol 34, pp 8600–8607
23. Alqaryouti O, Siyam N, Monem AA, Shaalan K (2020) Aspect-based sentiment analysis using smart government review data. Applied Computing and Informatics
24. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805
25. Feng S, Tran LV, Cong G, Chen L, Li J, Li F (2020) Hme: A hyperbolic metric embedding approach for next-poi recommendation. In: Proc. of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1429–1438
26. Ho NL, Lim KH (2021) User preferential tour recommendation based on poi-embedding methods. In: Proc. of 26th International conference on intelligent user interfaces, pp 46–48
27. Zhao S, Zhao T, King I, Lyu MR (2017) Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In: Proc. of the 26th international conference on world wide web companion, pp 153–162
28. He J, Qi J, Ramamohanarao K (2019) A joint context-aware embedding for trip recommendations. In: Proc. of 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, pp 292–303
29. Liao J, Liu T, Yin H, Chen T, Wang J, Wang Y (2021) An integrated model based on deep multimodal and rank learning for point-of-interest recommendation. World Wide Web 24(2):631–655
30. Chen J, Zhang H, He X, Nie L, Liu W, Chua T-S (2017) Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In: Proc. of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, pp 335–344
31. Manotumruksa J, Macdonald C, Ounis I (2018) A contextual attention recurrent architecture for context-aware venue recommendation. In: Proc. of The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp 555–564
32. Zhong T, Zhang S, Zhou F, Zhang K, Trajcevski G, Wu J (2020) Hybrid graph convolutional networks with multi-head attention for location recommendation. World Wide Web 23:3125–3151
33. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th International conference on learning representations, pp 1–14
34. Luo Y, Liu Q, Liu Z (2021) Stan: Spatio-temporal attention network for next location recommendation. In: Proc. of the Web Conference 2021, pp 2177–2185
35. Ma C, Zhang Y, Wang Q, Liu X (2018) Point-of-interest recommendation: Exploiting self-attentive autoencoders with neighbor-aware influence. In: Proc. of the 27th ACM International

Conference on Information and Knowledge Management, pp 697–706
36. Zuo S, Jiang H, Li Z, Zhao T, Zha H (2020) Transformer hawkes process. In: Proc. of international conference on machine learning, pp 11692–11702
37. Lai C-H, Hsu C-Y (2021) Rating prediction based on combination of review mining and user preference analysis. Inf Syst 99:101742
38. Yu D, Mu Y, Jin Y (2017) Rating prediction using review texts with underlying sentiments. Inf Process Lett 117:10–18
39. Cheng Z, Ding Y, Zhu L, Kankanhalli M (2018) Aspect-aware latent factor model: Rating prediction with ratings and reviews. In: Proc. of the 2018 World Wide Web conference, pp 639–648
40. Shen R-P, Zhang H-R, Yu H, Min F (2019) Sentiment based matrix factorization with reliability for recommendation. Expert Syst Appl 135:249–258
41. Karimi A, Rossi L, Prati A (2021) Adversarial training for aspect-based sentiment analysis with bert. 2020 25th International Conference on Pattern Recognition (ICPR), 8797–8803, IEEE
42. Song Y, Wang J, Liang Z, Liu Z, Jiang T (2020) Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference, arXiv:2002.04815
43. Xu H, Shu L, Yu PS, Liu B (2020) Understanding pre-trained bert for aspect-based sentiment analysis, arXiv:2011.00169
44. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: A lite bert for self-supervised learning of language representations, arXiv:1909.11942
45. Zhang L, Sun Z, Zhang J, Lei Y, Li C, Wu Z, Kloeden H, Klanner F (2020) An interactive multi-task learning framework for next poi recommendation with uncertain check-ins. CAL 301(985):13954
46. Halder S, Lim KH, Chan J, Zhang X (2021) Transformer-based multi-task learning for queuing time aware next poi recommendation. In: Proc. of Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 510–523
47. Lian D, Wu Y, Ge Y, Xie X, Chen E (2020) Geography-aware sequential location recommendation. In: Proc. of the 26th ACM SIGKDD International conference on knowledge discovery & data mining, pp 2009–2019
48. Xia B, Bai Y, Yin J, Li Q, Xu L (2020) Mtpr: A multi-task learning based poi recommendation considering temporal check-ins and geographical locations. Appl Sci 10(19):6664
49. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proc. of KDD, ACM, pp 1082–1090
50. Han P, Li Z, Liu Y, Zhao P, Li J, Wang H, Shang S (2020) Contextualized point-of-interest recommendation. In: Proc. of twenty-ninth international joint conference on artificial intelligence and seventeenth pacific rim international conference on artificial intelligence IJCAI-PRICAI-20
51. Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) Semeval-2014 task4: Aspect-based sentiment analysis. In: Proc. of the 8th international workshop on semantic evaluation, pp 27–35
52. Rahmani HA, Aliannejadi M, Baratchi M, Crestani F (2020) Joint geographical and temporal modeling based on matrix factorization for point-of-interest recommendation. In: Proc. of European conference on information retrieval, pp 205–219
53. Baratchi M, Afsharchi M, Crestani F (2020) Lglmf: Local geographical based logistic matrix factorization model for poi recommendation. In: Proc. of information retrieval technology, vol 12004, p 66
54. Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer, arXiv:2004.05150v2
55. Kitaev N, Kaiser L, Levskaya A (2020) Reformer: The efficient transformer, arXiv:2001.04451
56. Lu J, Yao J, Zhang J, Zhu X, Xu H, Gao W, Xu C, Xiang T, Zhang L (2021) Soft: Softmax-free transformer with linear complexity. Advances in Neural Information Processing Systems 34
57. Xiong Y, Zeng Z, Chakraborty R, Tan M, Fung G, Li Y, Singh V (2021) Nyströmformer: A nystöm-based algorithm for approximating self-attention. AAAI 35(16):14138

Xinfeng Wang received a B.S. degree in Software Engineering from Hangzhou Dianzi University Information Engineering School and a dual M.S. degree from Hangzhou Dianzi University in China and University of Yamanashi in Japan. His current research interests are recommendation systems and natural language processing. His e-mail is kaysen@hdu.edu.cn.



Fumiyo Fukumoto received MSc in the centre for Computational Linguistics of UMIST, the UK in 1993, and a Ph.D. degree in the Department of Information Science, University of Tokyo, Japan in 1997. She is currently a Professor at the Interdisciplinary Graduate School of Medicine and Engineering, Univ. of Yamanashi. Her current research interest is Natural Language Processing. She is a member of the Association for Computational Linguistics, the Association for Natural Language Processing, and the Information Processing Society of Japan.

Jiyi Li received his B.S. and M.S. from Nankai University, China, in 2005 and 2008, respectively. He received his Ph.D. degree from Department of Social Informatics, Graduate School of Informatics, Kyoto University in 2013. He is currently an assistant professor at Department of Computer Science and Engineering, University of Yamanashi, Japan. His current interests include natural language processing and data mining.

Dongjin Yu is currently a professor at Hangzhou Dianzi University, China. His research efforts include intelligent software engineering, service computing and data engineering. He is the director of Big Data Institute, and the director of Computer Software Institute of Hangzhou Dianzi University. He is a senior member of IEEE, and a senior member of China Computer Federation (CCF). He also serves as the executive members of the technical committees of Software Engineering and Service Computing CCF.

Xiaoxiao Sun is currently a lecturer at Hangzhou Dianzi University. She received the Ph.D. degree from Zhejiang University, Hangzhou, China in 2017. Her current research interests include spatio-temporal data mining, business process management, etc. Her e-mail is sunxiaoxiao@hdu.edu.cn.

Affiliations

Xinfeng Wang[1,2] · Fumiyo Fukumoto[3] · Jiyi Li[3] · Dongjin Yu[1] · Xiaoxiao Sun[1]

✉ Dongjin Yu
yudj@hdu.edu.cn

Xinfeng Wang
kaysen@hdu.edu.cn

Jiyi Li
jyli@yamanashi.ac.jp

Xiaoxiao Sun
sunxiaoxiao@hdu.edu.cn

1 School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China

2 Integrated Graduate School of Medicine, Engineering, Agricultural Sciences, Faculty of Engineering, University of Yamanashi, Kofu, 400-8511, Japan

3 Faculty of Engineering, Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Kofu, 400-8511, Japan