Geo-BERT Pre-training Model for Query Rewriting in POI Search

Xiao Liu[1] and **Juan Hu**[1] and **Qi Shen**[2] and **Huan Chen**[1]
Didi Chuxing, Beijing, China
{samliuxiao,hujuan,chenhuan}@didiglobal.com[1]
qishen@zju.edu.cn[2]

Abstract

Query Rewriting (QR) is proposed to solve the problem of the word mismatch between queries and documents in Web search. Existing approaches usually model QR with an end-to-end sequence-to-sequence (seq2seq) model. The state-of-the-art Transformer-based models can effectively learn textual semantics from user session logs, but they often ignore users' geographic location information that is crucial for the Point-of-Interest (POI) search of map services. In this paper, we proposed a pre-training model, called Geo-BERT, to integrate semantics and geographic information in the pre-trained representations of POIs. Firstly, we simulate POI distribution in the real world as a graph, in which nodes represent POIs and multiple geographic granularities. Then we use graph representation learning methods to get geographic representations. Finally, we train a BERT-like pre-training model with text and POIs' graph embeddings to get an integrated representation of both geographic and semantic information, and apply it in the QR of POI search. The proposed model achieves excellent accuracy on a wide range of real-world datasets of map services.

1 Introduction

Point-of-Interest (POI) search plays an important role in map services, such as Google Maps, Gaode Maps, Didi, etc. Query Rewriting (QR) is critical for POI search(Rieh et al., 2006) to solve the semantic gap between queries and POIs, created by users' mistype.

Currently, lots of methods have been tried to solve the QR problem(Antonellis et al., 2008; Ali et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014; He et al., 2016; Chen et al., 2020). Recently, the Transformer-based seq2seq models (Ashish Vaswani and Polosukhin, 2017; Yu et al., 2020) significantly improve the feature representation ability and rewriting performance.

While the Transformer-based rewriting method shows its effectiveness in QR, it could be further improved in the following aspects when applyed in POI search: (1) The input of POI search is different from the general search scenario, as it may contain rich geographic information such as the user's current location. For example, when people located in $city_A$ search "the olive" (a POI in $city_B$), yet they actually want to find "ten olive"(a POI in $city_A$). However, it is extremely hard to rewrite "the olive" without the position information. (2) Sometimes the location information is useless, while user's intention city is mainly obtained through query. Effectively capturing the geographic information corresponding to the query becomes particularly crucial to QR tasks in POI search.

To solve the above challenges, we propose a pre-training model called Geo-BERT that combines geographic feature graph with textual semantics in the QR task. First, we introduce a geographic feature graph to map multiple geographic granularity information to a unified graph representation space. Specifically, we connect the neighboring POIs to each other based on the longitude and latitude, meanwhile we connect the different administrative district granularity together with the above POI. After that, we propose a pre-training model that integrates text and POIs' graph embeddings, and fuse geographic features into the text semantic space by predicting masked geographic information. Finally, we fuse the pre-training model of geographic text into a Transformer-based seq2seq model.

Our contributions can be summarized as follows.

- We construct a novel geographic feature graph to map multiple geographic granularities into a unified latent space, which helps obtain the POI embeddings with geographic information.
- We proposed a pre-training model called Geo-BERT, to combine geographic knowledge and

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2209–2214
November 7–11, 2021.

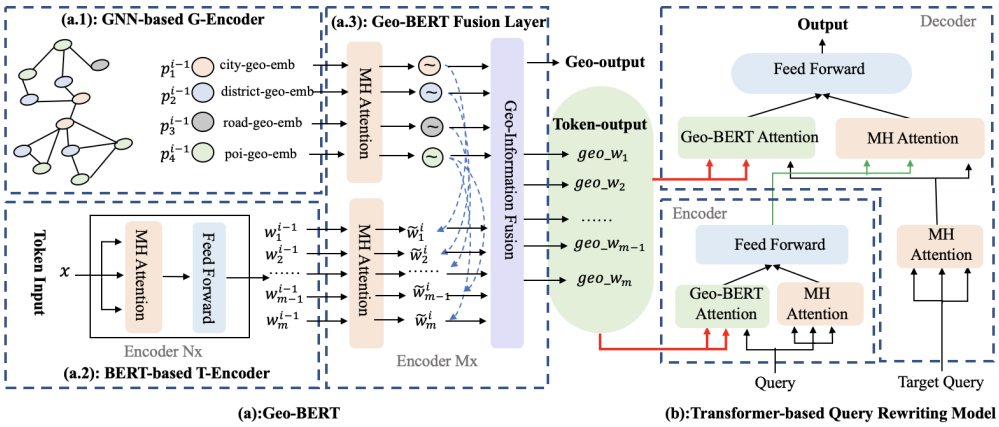


Figure 1: The overall architecture of the proposed model, including the (a):Geo-BERT layer and the (b):Transformer-based QR model. (a.1):GNN-based G-Encoder is GNN-based graph representations of multi-granularity geographic information; (a.2):BERT-based T-Encoder is the encoder layer of tokens, which is POI when pre-train and query when predict; and we fused the G-Encoder and T-Encoder by (a.3):Geo-BERT Fusion Layer. Finally, the token output of (a) is added into both the encoder and decoder of (b).

textual information, which integrates the geographic information into the text semantic space by predicting the masked geographic knowledge.

- We conduct extensive experiments to fuse Geo-BERT into the Transformer-based seq2seq model. The results show that it can achieve an excellent improvement on real-world datasets.

2 Background

Usually, incorporating external knowledge could enhance the performance of NLP tasks(Liu et al., 2020; Zhou et al., 2020; Han et al., 2018). Graph-based representation is able to express structured external knowledge effectively, (Hamilton et al., 2017) and leverages node feature information to infer unseen data by aggregating subsampled local neighborhoods. (Grover and Leskovec, 2016) incorporate breadth-first search and depth-frst search in neighborhood sampling to learn node embeddings. (Chiang et al., 2019) use subgraph sampling to reduce time and memory cost when using graph convolutuoin neural networks to learn larger graphs.

Recently, pre-training models such as BERT (Jacob Devlin and Toutanova, 2019) have shown their power in both understanding and generative tasks (Zhu et al., 2020). (Zhang et al., 2019) raise a BERT-like model to incorporate informative entities in knowledge graphs. Considering that POIs' geographic neighborhood relationship can be also expressed as graphs, we follow (Zhang et al., 2019) to incorporate geographic information in Transformer-based query rewrite models.

3 Methodology

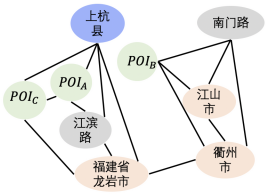
In this section, we present the overall framework(See Figure 1) of the proposed model.

3.1 Graph for Geographic Information

Queries in POI search may contain the administrative region information, e.g. city, district and road, so we consider constructing a fine-grained geographic graph.

	POI	City	District	Road	Distance from POI_A (km)
POI_A	福建省龙岩市上杭县江滨路21号	福建省龙岩市	上杭县	江滨路	0
POI_B	衢州市江山市南门路1号	衢州市	江山市	南门路	590.30
POI_C	福建省龙岩市上杭县成龙润发商行	福建省龙岩市	上杭县	——	0.21

(a): POI examples



(b): Geographical graph

Figure 2: The illustration of geographic graph. The distance between POI_A and POI_B is below 1 km and thus they are connected. POI_C is over 590 km far from the above two POIs, so there is no edges between them.

Considering the inclusion relationship among four geographic granularities, we build an undirected graph through the available geographic information with the following rules,

- Consider each POI as a node and connect adjacent nodes whose distance is less than 1 km;

- Consider each administrative region (city, district and road) as a node and connect it to the POI nodes in this region;
- Connect the administrative region nodes with their inclusive regions, and all the city nodes are connected;
- All the edges are unweighted.

Figure 2 illustrates the geographic graph. The graph is not only based on the neighborhood relationship between POIs, but also fuses the inclusion relationship between administrative regions. It is unweighted because two following reasons: (1) we have no idea about the path between two POIs for the lack of complete map information; (2) we hope to simplify the graph to make the learned representations more robust.

We use graph embedding algorithms, e.g. node2vec (Grover and Leskovec, 2016), to get the node representations that contain geographic information.

3.2 Geo-BERT Architecture

The whole pre-training model Geo-BERT consists of two stacked modules: (1) the underlying textual encoder (**T-Encoder**) responsible for capturing basic lexical and syntactic information from the input tokens; (2) the upper geographic encoder (**G-Encoder**) responsible for integrating extra token-oriented geographic information into textual information from the underlying layer.

Let a token sequence be $w_1, ..., w_n$, where n is the length of the token sequence. Meanwhile, we denote the POI sequence aligning to the given tokens as $p_1, ..., p_n$. Furthermore, we denote the whole vocabulary as $\mathbb{V}$, and the POI list in the geographic graph as $\mathbb{P}$. If a token $w \in \mathbb{V}$ has a corresponding POI geographic sequence $p \in \mathbb{P}$, their alignment is defined as $f(w) = p$. Besides, we denote the number of **T-Encoder** layers as N, and the number of **G-Encoder** layers as M. In this paper, we hope that each word in a query could reconstruct geographic information through pre-training. Thus, we align a geographic phrase to every corresponding token as shown in Figure 3.

Masked Mechanism: the pre-training contains two tasks, one of which is the masked language model (MLM(Jacob Devlin and Toutanova, 2019)) to learn semantic features and the other is masked geographic information model (MGM) to learn geographic features. The MGM, which is designed for learning geographic information, masks geographic granularities with a probability of 0.5.

T-Encoder firstly sums the token embedding, segment embedding, positional embedding for each token to compute its input embedding, and then computes deep features $\mathbf{w}_1, ..., \mathbf{w}_n$ as $\mathbf{w}_1, ..., \mathbf{w}_n =$ **T-Encoder**$(w_1, ..., w_n)$.

Then, the i-th aggregator integrates token and geographic sequence through a fusion layer, and computes the output embedding for each token and geographic entity. The information fusion process is as follows,

$$\begin{aligned} \mathbf{h}_j &= \sigma(\tilde{\mathbf{W}}_t^{(i)} \tilde{\mathbf{w}}_t^{(i)} + \tilde{\mathbf{W}}_p^{(i)} \tilde{\mathbf{p}}_k^{(i)} + \tilde{b}^{(i)}) \\ \mathbf{w}_j^{(i)} &= \sigma(\mathbf{W}_t^{(i)} h_j + b_t^{(i)}) \\ \mathbf{p}_k^{(i)} &= \sigma(\mathbf{W}_p^{(i)} h_j + b_p^{(i)}) \end{aligned} \tag{1}$$

where h_j is the inner hidden state integrating the information of both tokens and geographic entities. $\sigma(\cdot)$ is a non-linear activation function, which is set as GELU (Hendrycks and Gimpel, 2016) in the experiments.

For simplicity, the i-th aggregator operation is denoted as follows,

$$\begin{aligned} \mathbf{w}_1^{(i)}, ..., \mathbf{w}_n^{(i)}, \mathbf{p}_1^{(i)}, ..., \mathbf{p}_n^{(i)} = Aggregator(\\ \mathbf{w}_1^{(i-1)}, ..., \mathbf{w}_n^{(i-1)}, \mathbf{p}_1^{(i-1)}, ..., \mathbf{p}_n^{(i-1)}). \end{aligned} \tag{2}$$

The output embeddings of both tokens and POI geographic entities computed by the top aggregator will be used as the final output embeddings of the geographic encoder **G-Encoder**.

Figure 3: The example of pre-training dataset. The geographic labels "C", "D", "R" and "P" respectively denote the graph embeddings of "City", "District", "Road" and "POI coordinate". "M" denotes the masked label used for the masked language model and the masked geographic information model.

3.3 Fusion in Sequence-to-sequence Model

An illustration of the overall QR framework is shown in Figure 1. Any input $x \in \mathbb{X}$ is progressively processed by the Geo-BERT, encoder and decoder. The entire procedure of our algorithm is as follows, Step-1: Given any token input $x = w_1, ..., w_n$, **Geo-BERT** first encodes it into representation $H_B =$ **Geo-BERT**(x). Step-2:

Then H_B is fused into Transformer-based Seq2Seq Model as the same method in (Zhu et al., 2020).

4 Experiments

4.1 Dataset

The QR data in the paper is from the internal real-world dataset[1]. Each sample is a pair of source-query and target-query, and the source-query is the real search text and the target-query is the one with click behavior in session. The dataset is divided into a training with 7.5M examples and a test set with 8.1K examples. Especially, we construct a geographic-related test set named Geo-test whose examples are subjectively chosen according to whether their rewritting relies on geographic information. The pre-training dataset contains over 10.7M POI samples. Each sample includes name, address, longitude and latitude of POI.

4.2 Results and Analysis

4.2.1 QR Performance

Baseline: The baseline models are a vanilla Transformer-based NMT model (Ashish Vaswani and Polosukhin, 2017) and its version fused with BERT (Zhu et al., 2020). When using BERT, we respectively take two kinds of methods. One is to directedly finetune it with the NMT model on QR dataset and the other, called POI-BERT, is to pre-train BERT on our own POI corpus.

Experimental settings: Most experimental settings of Geo-BERT follow (Zhang et al., 2019). Especially, the geographic graph embedding size is set to 128. We pre-train Geo-BERT on POI dataset for 3 epochs. Most experimental settings of the NMT model follow (Zhu et al., 2020). The maximum training iteration is set to 300K. We keep total number of tokens in each batch below 12K.

	Regular test		Geo-test	
	Top1	**Top3**	**Top1**	**Top3**
NMT(Transformer)	55.46	69.91	54.98	69.43
+BERT	57.61	70.68	58.33	69.82
+POI-BERT	58.10	71.58	57.82	69.45
+Geo-BERT-SG	62.82	74.32	65.14	75.40
+Geo-BERT-MG	**65.51**	**77.78**	**66.78**	**79.24**

Table 1: The top1/top3 accuracy comparison on test set. "Geo-BERT-SG"denotes Geo-BERT with the single geographic granularity, that is POI longitude and latitude; "Geo-BERT-MG" denotes Geo-BERT with multiple geographic granularities.

[1]The data are collected through Didichuxing in China.

Results: Table 1 shows that Geo-BERT has overall improvement on both regular dataset and Geo-test dataset. Compaired to baselines, a simple NMT model fused with Geo-BERT achieves at least 4.59% and 6.93% top1 accuracy gains as well as 2.68% and 5.62% top3 accuracy gains on two datasets. Note that Geo-BERT helps QR models more on Geo-test set, we believe that it could learn useful geographic information while retaining semantic information. An interesting fact in Table 1 is that pre-training Geo-test data with BERT ("NMT + POI-BERT") leads to 0.45% top1 decrease and 0.36% top3 decrease compared to "NMT + BERT" on Geo-test set. That means, in geographic-correlated QR tasks, Geo-BERT is definately neccessary because a vanilla BERT cannot actually learn geographic representations.

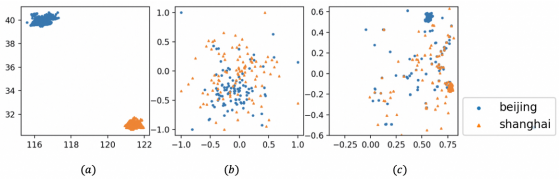


Figure 4: The TSNE visualization of POIs' geographic distribution in two cities and their pre-trained representations. (a) POIs' latitude and longitudes; (b) POIs' BERT pre-trained represention; (c) POIs' Geo-BERT pre-trained representations.

Figure 4 shows the learned geographic information of Geo-BERT, we respectively choose 300 POIs in Beijing and Shanghai to display their latitude and longitudes as well as the pre-trained representations of their address. Different from BERT, in Geo-BERT, we find that the representations of POIs in the same city tend to gather while those in different cities tend to seperate. Obviously, the Geo-BERT model benefits extracting the geographic feature.

4.2.2 Ablation Study

	Regular test		Geo-test	
Geo-BERT-NMT	**Top1**	**Top3**	**Top1**	**Top3**
with all granularities	65.51	77.78	66.78	79.24
without road	63.15	75.80	64.77	76.66
without district	64.95	77.25	65.30	77.35
without city	65.23	77.58	66.58	79.13

Table 2: The top1/top3 accuracy of Geo-BERT-NMT with various geographic granularities on test set.

According to POI address, we can extract the corresponding city, district, town or road.

Their proportion in POI dataset is respectively $46.37\%, 46.85\%, 15.13\%, 42.81\%$. Except the sparse town information, we improve Geo-BERT through three frequent geographic granularities, including city, district and road.

Table 2 shows the influence of each geographic granularity on two test set. As can be seen, the "city" granualrity has weakest impact on both regualar test set and Geo-test set. On the other hand, the "road" granularity is most effective.

5 Conclusion

In this paper, we proposed a pre-training model called Geo-BERT, and applied it to the QR task in POI search. Specially, we adopt a multiple geographic granularity graph and combine texual semantics with geographic infomation of POIs. The proposed pre-trained model adopts sepcial masked strategy to learn meaningful geographic features. Experimental results show that our model outperforms many strong baselines on a wide range of real-world datasets of map services.

References

Alnur Ali, Jianfeng Gao, Xiaodong He, Bodo Von Billerbeck, and Sanaz Ahari. 2014. Enhanced query rewriting through statistical machine translation. US Patent 8,732,151.

Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. 2008. Simrank++ query rewriting through link analysis of the clickgraph (poster). In *Proceedings of the 17th international conference on World Wide Web*, pages 1177–1178.

Niki Parmar Jakob Uszkoreit Llion Jones Aidan N Gomez Łukasz Kaiser Ashish Vaswani, Noam Shazeer and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, page 5998–6008.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Zheng Chen, Xing Fan, and Yuan Ling. 2020. Pre-training for query rewriting in a spoken language understanding system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7969–7973. IEEE.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. 2019. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 257–266.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216*.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yunlong He, Jiliang Tang, Hua Ouyang, Changsung Kang, Dawei Yin, and Yi Chang. 2016. Learning to rewrite queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1443–1452.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2020. *Representation Learning for Natural Language Processing*. Springer Nature.

Soo Young Rieh et al. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3):751–768.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1933–1936.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.