Geo-Encoder: A Chunk-Argument Bi-Encoder Framework for Chinese Geographic Re-Ranking

Anonymous ACL submission

Abstract

Chinese geographic re-ranking task aims to find the most relevant addresses among retrieved candidates, which is crucial for location-related services such as navigation maps. Unlike the general sentences, Chinese geographic contexts are closely intertwined with geographical concepts, from general spans (e.g., province) to specific spans (e.g., road). Given this feature, we propose an innovative framework, namely *Geo-Encoder*, to more effectively integrate Chinese geographical semantics into re-ranking pipelines. Our methodology begins by employing off-the-shelf tools to associate text with geographical spans, treating them as chunking units. Then, we present a multi-task learning module to simultaneously acquire an effective attention matrix that determines chunk contributions to geographic representations. Furthermore, we put forth an asynchronous update mechanism for the proposed task, aiming to guide the model to focus on specific chunks. Experiments on two Chinese benchmark datasets, show that the *Geo-Encoder* achieves significant improvements when compared to state-of-the-art baselines. Notably, it leads to a substantial improvement in the Hit@1 score of MGEO-BERT, increasing it by 6.22% from 62.76 to 68.98 on the GeoTES dataset.

1 Introduction

Chinese geographic re-ranking (CGR) is a sub-task of semantic matching, aiming to identify the most relevant geographic context towards given queries and retrieved candidates (Zhao et al., 2019; MacAvaney et al., 2020; Yates et al., 2021). It is a crucial task that serves many downstream applications such as navigation maps (e.g., Gaode Maps), autonomous driving (e.g., Tesla), E-commerce system (e.g., Taobao), etc. (Jia et al., 2017; Avvenuti et al., 2018). Unlike general query text, Chinese geographic sentences exhibit a distinct attribute in their linear-chain structural semantics (Li et al., 2019).

Figure 1: Overview of the Chinese Geographic re-ranking task. The process begins with the user query being subjected to word chunking, segmenting it into meaningful units. Lastly, *Geo-Encoder* is employed to enhance semantic representation and re-ranking.

This peculiarity arises from the fact that Chinese addresses often comprise distinct meaningful address segments, termed as geographic chunks in linguistic terms (Abney, 1991). These chunks adhere to a specific format, organizing from the general (e.g., province) to the more specific (e.g., road). For example, as is shown in Figure 1, given a Chinese address "采荷路2号高级中学北门 *(North Gate of Caihe Road No.2 Senior High School)*", we can deconstruct it into several such chunks: "采荷路 *(Caihe Road)*", "2号 *(No.2)*", "高级中学 *(Senior High School)*", "北门 *(North Gate)*".

Conventional approaches (Reimers and Gurevych, 2019; Humeau et al., 2019; Khattab and Zaharia, 2020) addressing the CGR task often directly employ pre-trained language models (PLMs) to encode given geographic texts into embeddings, which are subsequently subjected to re-ranking through similarity calculation techniques like cosine or euclidean distance measures. Recent works (Yuan et al., 2020; Huang et al., 2022; Ding et al., 2023) in this field extend beyond mere geographic context utilization and encompass an expansive range of data sources, including point-of-interest (POI) information, multi-modal data, and user behavioral attributes (Liu et al., 2021; Hofmann et al., 2022; Huang et al., 2022) with a larger neural model. The outcome of this integration is characterized by notable enhancements, achieved through the fusion of external geographic knowledge. Furthermore, cutting-edge domain-adaptation frameworks have been introduced to facilitate effective fusion of multi-domain data, such as PALM (Zhao et al., 2019), STDGAT (Yuan et al., 2020), etc.

However, despite the effectiveness of existing attempts in leveraging geographic knowledge, these methods failed to fully harness the intrinsic potential of the geographic context itself. Therefore, in this paper, we aim to shift our focus towards the geographic context by exploiting its distinctive linear-chain attributes. To achieve this, we employ off-the-shelf tools (e.g. MGEO tagging and part-of-speech (POS) for the approximate annotation of each geographic text with pertinent geographic chunks. For example, as illustrated in Figure 1, we annotate the text "采荷路 *(Caihe Road)*" with the label *Road*, "2号 *(No.2)*" with *Num*, etc.

Firstly, building upon this foundation, we introduce an additional task that revolves learning the similarity between different components of these annotated chunks. This involves the formulation of an attention matrix, which governs the contributions of these chunks to the semantic representations. Our motivation is that general chunks tend to be less diverse across queries and candidates, and specific chunks possess a higher degree of distinctiveness. **Secondly**, we put forth a novel asynchronous update speed mechanism for the attention matrix. This mechanism is designed to empower the model to effectively focus its attention on the more specific chunks, thereby enhancing its discernment capabilities. **Lastly**, we advocate for the integration of the pure bi-encoder approach during the inference period. This strategy ensures a harmonious balance between performance and computational efficiency, safeguarding the efficacy of the model in both academic and industrial scenarios.

In summary, our key contributions can be summarized as follows: 1) We introduce a multi-task learning framework, denoted as *Geo-Encoder*, to integrate component similarity; 2) We present an asynchronous update mechanism, to distinguish specific chunks effectively; 3) Except evaluation on benchmark dataset, we collect and publish a nationwide geographic dataset in China, named GeoIND. Experimental results demonstrate the superiority of our *Geo-Encoder* over competitive methods.

2 Related Work

Semantic Matching and Re-Ranking. Semantic matching is a widely-concerned task in natural language processing, including retrieval and re-ranking process (Zhao et al., 2019; Yates et al., 2021). Within re-ranking domain, researchers employ bi-encoders to encode given queries and candidates separately by using the shared parameters, such as ESIM (Chen et al., 2017), SBERT (Reimers and Gurevych, 2019), ColBERT (Khattab and Zaharia, 2020), etc. Within the emergence of pre-trained models, such as RoBERTa (Liu et al., 2019), ERNIE (Sun et al., 2021), cross-encoders were proposed to jointly encode text and promote the information interaction (Humeau et al., 2019; Nie et al., 2020; Ye et al., 2022). Besides, to better represent sentences, external knowledge and late interactions were widely explored. For example, Xia et al. (2021) utilized a word similarity matrix and Peng et al. (2022) introduced predicate-argument spans to enhance representation.

Chinese Geographic Text Representation. Most existing approaches focused on encoding geographic text by external knowledge in two aspects: (1) position data, such as PALM (Zhao et al., 2019), encoding positional relationship of query and candidates, STDGAT (Yuan et al., 2020), considering Spatio-temporal features, etc.; (2) geographic knowledge, such as GeoL (Huang et al., 2022), using knowledge related to user behaviors, and MGeo (Ding et al., 2023), proposing using multi-modal dataset. However, the geographic text encoding method among the above approaches is not well-explored. Besides, parsing Chinese geographic text into chunks is also a key technical issue (Li et al., 2019). Generally, address parsing is quite similar to Chinese word segmentation. Existing attempts includes conditional random fields models (Zhao et al., 2006), latent-variable variants (Sun et al., 2009), neural transition-based segmentation method (Zhang et al., 2016), and chart-based models (Stern et al., 2017; Kitaev and Klein, 2018), etc. However, while these models benefit from external geographic knowledge, exploring geographic rather than semantic representation is still crucial.

3 Our Approach

3.1 Task Definition and Overview

In Chinese Geographic Re-ranking (CGR) task setting, the available dataset $\{X\}$ is formed as query-candidate pairs. Let Q denotes queries and C as retrieved candidates. Both Q and C are composed of l-th separated tokens, where $\{X\} = \{X \in (Q, C) | X = x_1, x_2, ..., x_l\}$. The objective of CGR is to model the highest possibility of C. Thus, the bi-encoder framework, depicted in Figure 2(a), can be formalized as:

$$c = \arg\max_{C} r_\theta \left(f_\theta(Q), f_\theta(C)\right) \tag{1}$$

where f_θ denotes encoding function (we adopt PLMs here), to encode given text into vectors, c ($\in C$) is the model output and r_θ denotes similarity evaluation function, such as dot multiple and cosine similarity, to assign a similarity score for each candidate. Also, the cross-encoder framework, depicted in Figure 2(b) can be formalized as:

$$c = \arg\max_{C} r_\theta \left(f_\theta(Q, C)\right) \tag{2}$$

In our approach, we strive to enhance the encoding process through a two-step strategy. Firstly, we segment the provided geographic text into chunks and introduce a novel approach to learn both the attention matrix governing chunk contributions and component semantic representation as an additional task. Secondly, we introduce an asynchronous update mechanism for the attention matrix and model parameters. This mechanism is aimed at enabling the model to efficiently acquire the skill of focusing on specific chunks. Finally, we present our training and inference details. Our proposed framework, called *Geo-Encoder*, is illustrated in Figure 2(c).

3.2 Geographic Chunking

Chinese addresses typically consist of multiple meaningful address segments, often referred to as "geographic chunks" (Abney, 1991). These addresses follow a structured pattern, progressing hierarchically from the general (e.g., province) to specific ones (e.g., road) (Li et al., 2019). In contrast to conventional Chinese segmentation methods, geographic chunking demands tools of heightened sensitivity tailored to geographical units. These tools necessitate fine-tuning using dedicated Chinese address corpora. Consequently, we adopt the MGEO tagging tool to facilitate the acquisition of precise geographic annotations for our benchmark datasets (Wu et al., 2022a,b; Ding et al., 2023).

MGEO stands as a pre-trained model with multi-modal datasets, encompassing both geographic context and points of interest. It is designed to cater to various downstream tasks, including geographic entity alignment and address element tagging, among others. In our current framework, however, we exclusively leverage MGEO to provide chunk annotations, without employing it for the purpose of encoding contextual information. Then, dataset $\{X\}$ is extended as $\{X_u\} = \{x \in (Q, C, Q_u, C_u) | X = x_1, x_2, ..., x_n\}$, where Q_u and C_u denotes query and candidates chunking units. For example, given a Chinese address "南京市新城科技园*3*栋*5*单元*(Unit #5, Building #3, Sci-Tech Park, Nanjing City.)*", we can parse them by MGEO into: "南京市*(Nanjing City)*" – city, "新城科技园*(Sci-Tech Park)*" – devzone, "*3*栋*(Building #3)*" – houseno, "*5*单元*(Unit #5)*" – cellno.

3.3 Chunking Contribution Learning

Utilizing the chunked dataset denoted as $\{X_u\}$, we proceed to employ a pre-trained language model for the encoding process. This yields the representations of *[CLS]* embedding e^q_{cls} and token embedding $e^q_{1:l}$ from geographic text:

$$e^q_{cls}, e^q_{1:l} = Encoder(q), q \in Q \tag{3}$$

where $Encoder$ denotes PTMs. And correspondingly we can get candidates features e^c_{cls} and $e^c_{1:l}$. Given chunking annotations, we initialize a zeros query component embeddings $\{U^Q | u^q_i \in U^Q\}, i = \{1, 2, \cdots, M\}$ and we can further update query component embeddings u^q_i by:

$$u^q_i = mean(\Gamma(e^q_{1:l}, I^q_i)) \tag{4}$$

where $\Gamma(\cdot)$ is the Index function to obtain component token embeddings, M is the total amount of chunk categories, and I^q_i is the index number acquired by the tokenizer of the $Encoder$ from the chunk's location to the corresponding query. Similarly, component embeddings $\{U^C | u^c_i \in U^C\}$ can also be obtained. We can also get candidates' component embeddings u^c_i similar with Eq. 4.

To incorporate token-level embeddings, the ColBERT model (Khattab and Zaharia, 2020) introduced a multi-attention mechanism, which facilitates subsequent interactions between queries and candidates. This technique has demonstrated improved efficacy in re-ranking tasks. Nonetheless,

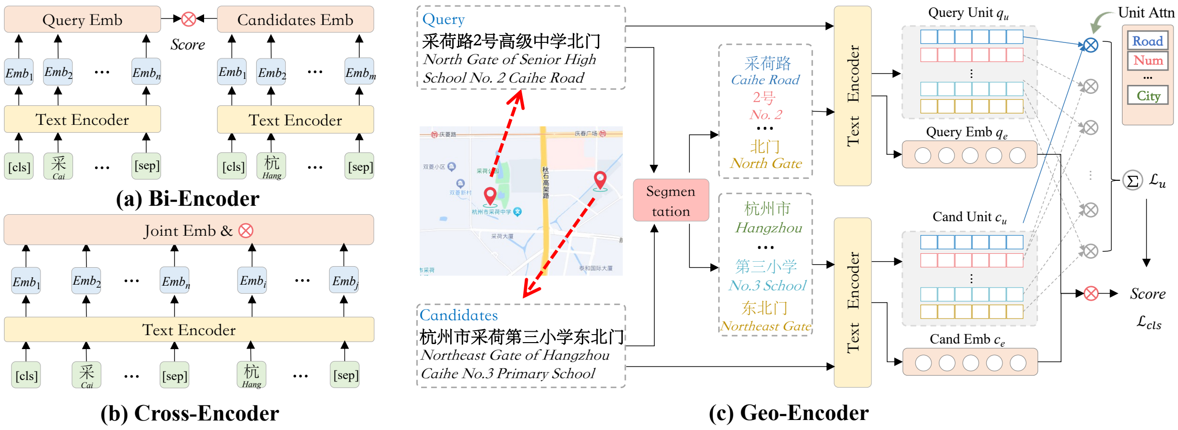


Figure 2: Architecture of re-ranking models and our proposed Geo-Encoder. The left shows the bi-encoder and cross-encoder models, and the right shows our proposed Geo-Encoder, which parsing geographic text into chunking units and jointly encoding with global features and unit attention mechanism. $\otimes$ denotes similarity calculation.

it is essential to acknowledge that the ColBERT method entails significant additional computational resources. In light of this, our work introduces an innovative multi-task learning module that incorporates only geographic chunking component embeddings and utilizes an attention matrix to fuse results. This approach is designed to address the need for efficient resource utilization while maintaining or potentially improving performance.

Specifically, we define an attention matrix that can be learned along the training process, denoted as W_U. Then, we can get the predictions from component embeddings:

$$Score_u = (U^Q * W^U) * (U^C * W^U) \quad (5)$$

We use dot multiplication to obtain the similarity scores of given queries and candidates. Thus, for components embeddings, we can obtain the component similarity loss $\mathcal{L}_u$ as:

$$\mathcal{L}_u = \Phi(Score_u, Y) \quad (6)$$

where Y represents the ground truth ranking results, and $\Phi(\cdot)$ signifies the cross-entropy loss function.

As for the primary task, we use *[CLS]* representation as sentence encoded features, and we can obtain the semantic similarity loss $\mathcal{L}_{cls}$ as:

$$\mathcal{L}_{cls} = \Phi(E^Q_{cls} * E^C_{cls}, Y) \quad (7)$$

where $e^q_{cls} \in E^Q_{cls}$ and $e^c_{cls} \in E^C_{cls}$.

3.4 Asynchronous Update Mechanism

For multi-task learning, a common concern is the disparate challenges faced by models when learning multiple tasks simultaneously, often leading to variations in convergence rates (Lu et al., 2017; He et al., 2017). In our pursuit to tackle this quandary within our designated task, we deviate from established methodologies seen in prior literature (Isonuma et al., 2017; Hashimoto et al., 2017; Nishino et al., 2019; Pfeiffer et al., 2020). Instead, we propose an innovative approach involving the integration of an asynchronous update mechanism, which allocates enhanced focus on training steps pertaining to distinct tasks. To formalize our proposition, the update of parameter $w_u (w_u \in W_U)$ is as:

$$w'_u = w_u + \lambda \cdot \nabla w_u \cdot \gamma \quad (8)$$

where γ is a hyper-parameter to adjust training speed, which can set by grid search or empirically.

3.5 Training and Inference

During the training process of CGR, we deploy our proposed framework *Geo-Encoder* of Figure 2(c). The model can be optimized by jointly minimizing the semantic similarity loss and component similarity loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_u \quad (9)$$

During the inference phase, a notable concern arises from the time-intensive nature of indexing and calculating component embeddings, particularly when extrapolated to scenarios involving an extensive pool of candidates. To circumvent this challenge, we directly adopt a bi-encoder framework for conducting inference process, as visually depicted in Figure 2(a).

4 Experiment

4.1 Datasets

Geographic Textual Similarity Benchmark (GeoTES): This large-scale dataset comprises queries meticulously crafted by human annotators and was amassed within the location of Hangzhou, China.[1] The dataset's meticulous annotation was executed by a panel of 20 participants and four domain experts. Encompassing a total of 90,000 queries, each complemented by 20/40 retrieved candidates, this dataset extends its scope beyond geographical text, using additional POI data.

Industry Geographic dataset (GeoIND): For a broader validation, we re-organize and format an additional real-world dataset named GeoIndustry, sourced from a geographic search engine. This dataset underwent rigorous cleaning and filtration procedures, effectively eliminating noise and erroneous queries. In contrast to GeoTES, this dataset exhibits an intermediary scale, yet it boasts a substantial geographical coverage. We will make it publicly available upon the publication of our work.

4.2 Baselines

To validate the effectiveness of our Geo-Encoder, we undertake a comprehensive analysis via representative bi-encoder methods. It's pertinent to mention that our assessment confines itself exclusively to geographic text data, with the exclusion of Points of Interest (POIs) or other modal data. Our selected baselines include: **Word2Vec** (Mikolov et al., 2013), a traditional method captured semantic relationships between words and encoded words as dense vector embeddings.[2] **Glove** (Pennington et al. 2014), which encapsulated both global and local semantic information and served for contextual understanding. **SBERT** (Reimers and Gurevych, 2019), a popular bi-encoder model that can effectively and efficiently serve for re-ranking task.[3] **Argument-Encoder** (Peng et al., 2022), which proposed that concatenate predicate-argument embedding as extra representations can enhance re-ranking task.[4] **MGEO** (Ding et al., 2023), which achieves state-of-the-art results in current task.[5]

Benchmark	Sets	Query	Tokens	ASL	Cands
GeoTES	Train	50,000	3,599	18.8	20
	Dev	20,000	3,322	17.2	40
	Test	20,000	3,351	17.1	40
GeoIND	Train	7,359	3,768	15.1	20
	Dev	2,453	3,376	15.1	20
	Test	2,469	2,900	15.0	20

Table 1: The statistics of two datasets. *Tokens* denotes vocabularies counts, *ASL* denotes the average sentence length, and *Cands* represents candidates numbers.

4.3 Experimental Setting

Evaluation Metrics. Following previous re-ranking tasks (Qu et al., 2021; Ding et al., 2023), we use Hit@K(K=1,3), NDCG@1 (Järvelin and Kekäläinen, 2002) and MRR@3 to evaluate the performance across all models. Specifically, Hit@K quantifies the proportion of retrieved candidates that include at least one correct item within the top K ranks. NDCG@1 is a graded relevance measure that takes into account the positions of relevant items in the ranked list. MRR@3 calculates the average of the reciprocal ranks of the top-3 correct answers in the ranked list.

Hyper-parameters. For finetuing, we set the learning rate is set as 1e-5 for RoBERTa and 5e-5 for BERT and ERNIE. We finetune models for 50 epochs with early stopping after 3 epochs of no improvement in Hit@1 on the validation set. We conduct our experiment on a single A100 GPU and optimize all the models with Adam optimizer, where the batch size is set to 32. And followed by Ding et al. (2023), we decrease the embedding dimension from 768 to 256.

4.4 Main Results

We have conducted a rigorous comparison between our method with the aforementioned baselines and the results are presented in Table 2.

Firstly, it is evident that our proposed approach achieves a remarkable state-of-the-art performance across all evaluated metrics, surpassing the performance exhibited by all alternative methods. Particularly, our method improves the Hit@1 score of BERT by 6.62% from 62.76 to 68.98 on GeoTES and by 2.59% from 64.12 to 66.71 on GeoIND.

[1]The dataset can be downloaded here: `https://modelscope.cn/datasets/damo/GeoGLUE/summary`.

[2]Reproduced by text2vec package(Xu, 2023): `https://github.com/shibing624/text2vec`.

[3]https://github.com/UKPLab/sentence-transformers.

[4]We reproduce this method by replacing the predicate-argument with specific geographic-argument.

[5]We compare three backbone models with MGEO in text-only modal data, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ERNIE 3.0 (Sun et al., 2021).

Model	GeoTES				GeoIND			
	Hit@1	Hit@3	NDCG@1	MRR@3	Hit@1	Hit@3	NDCG@1	MRR@3
Word2vec (Mikolov et al., 2013)	19.26	30.60	28.79	24.15	47.79	71.69	66.15	58.27
Glove (Pennington et al. 2014)	48.02	67.33	63.32	59.35	52.38	74.87	71.95	69.35
SBERT (Reimers and Gurevych, 2019)	24.22	51.22	46.65	35.80	42.20	71.24	64.56	54.92
Argument-Encoder (Peng et al., 2022)	56.54	80.01	73.47	67.08	59.58	85.54	78.61	71.19
MGEO-BERT (Ding et al., 2023)	62.76	80.89	75.95	70.87	64.12	88.66	81.35	75.04
Geo-Encoder	**68.98**	**85.82**	**81.11**	**76.56**	**66.71**	**89.35**	**82.78**	**76.99**
MGEO-ERNIE (Ding et al., 2023)	67.50	84.54	79.60	75.15	63.95	87.89	81.06	74.60
Geo-Encoder	**68.66**	**85.64**	**80.75**	**76.30**	**65.33**	**89.06**	**82.10**	**75.98**
MGEO-RoBERTa (Ding et al., 2023)	68.74	85.16	80.63	76.15	63.63	88.70	81.62	74.81
Geo-Encoder	**70.39**	**86.69**	**81.97**	**77.72**	**67.27**	**90.28**	**83.61**	**77.56**

Table 2: Main results on GeoTES and GeoIND, where bold values indicate the best performance within each column. Our proposed method consistently outperforms all three baselines across all metrics on both datasets.

Secondly, RoBERTa performs emerges as the superior candidate, surpassing both BERT and ERNIE. This advantage can be attributed to RoBERTa's augmented network depth and its exposure to a comprehensive training corpus, endowing it with a heightened capacity for contextual comprehension and modeling than other models.

Thirdly, a notable trend is that the GeoTES dataset is marginally more amenable to learning compared to the GeoIND dataset, a phenomenon primarily attributed to its significantly larger scale, which is 6.76 times greater. This distinction is corroborated by the highest attained Hit@1 score of 70.39 on the GeoTES dataset, as opposed to the score of 67.27 observed on the GeoIND dataset.

Furthermore, we can also conclude that conventional encoding methodologies such as word2vec, GloVe, and SBERT exhibit subpar performance in CGR tasks. And cosine similarity tends to exhibit suboptimal performance compared to dot multiplication for CGR task, which is evident from the fact that SBERT yields lower performance scores across both datasets. Similarly, the argument-enhancement techniques and the MGEO bi-encoder manifest a consistently underwhelming performance across both datasets.

5 Analysis and Discussion

5.1 Fix Contribution vs. Learning Weight

We constructed an experimental framework wherein the dynamic interplay of chunk contributions is examined. This is realized by configuring the attention matrices within the *Geo-Encoder* architecture as constant values, effectively precluding gradient updates. Initialization is undertaken by the values of 0.1, 0.5, and 1.0 respectively, thereby probing the impact of different attention allocation strategies on the learning process.

Method	Hit@1	Hit@3	NDCG@1	MRR@3
GeoTES				
baseline	62.76	80.89	75.95	70.87
w Fixed_1.0	68.08	85.35	80.48	75.84
w Fixed_0.5	66.02	83.91	78.97	74.03
w Fixed_0.1	68.19	84.95	80.31	75.70
w POS ***(Ours)***	68.25	85.55	80.65	76.02
w Geo ***(Ours)***	**68.98**	**85.82**	**81.11**	**76.56**
GeoIND				
baseline	64.12	88.66	81.35	75.04
w Fixed_1.0	65.61	89.59	82.47	76.39
w Fixed_0.5	65.69	89.06	82.28	76.23
w Fixed_0.1	64.20	87.85	81.14	74.77
w POS ***(Ours)***	65.21	89.59	82.24	76.06
w Geo ***(Ours)***	**66.71**	**89.35**	**82.78**	**76.99**

Table 3: Ablation study including exclude automatic attention update mechanism and geographic chunking.

As is shown in Table 3, we can find that the imposition of fixed attention matrices contributes to a reduction in the performance of the *Geo-Encoder* across both datasets. Besides, the diverse initialization schemes for these attention matrices yield distinct effects among datasets. Within the GeoTES dataset, an initialization ratio of 0.1 yields optimal results, indicating a higher reliance on the sentence-level *[CLS]* representation. Conversely, the GeoIND dataset attains peak performance when the ratio is set to 1.0, implying a contrasting attention distribution trend. Lastly, we find that even exclude the automatic update of attention matrices, the resultant performance still surpasses that of the baseline models. This outcome underscores the benefits derived from the incorporation of chunking information, substantiating its constructive impact on enhancing the overall model performance.

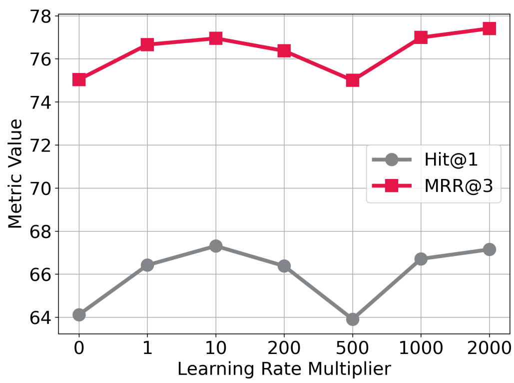


Figure 3: Comparing performance with varying learning rate multiplier ratios on the GeoIND dataset. The learning rate multiplier signifies the ratio of attention matrix learning rate to model parameter learning rate.

5.2 Geo Chunking vs. General Chunking

Subsequently, our investigation delves deeper into the influence of geographic chunks (Geo) by conducting a substitution experiment wherein these chunks are replaced with Part-of-Speech (POS) tagging results. To achieve this, we employ the jieba POS tagging tool to restructure the two datasets[6]. It is essential to note that the core distinction between POS and Geo lies in the target of segmentation: while GEO is geared towards geographic ontology, POS is more focused on semantic components.

The results, as depicted in Table 3, yield an interesting observation that employing POS tagging can benefit both datasets, signified by the obvious superior performance of POS when compared to the baseline. This favorable outcome can be attributed to the additional representation and multi-task learning introduced by our approach. Nevertheless, it is noteworthy that despite the advantageous performance of POS, it lags behind Geo in terms of performance. This discrepancy further underscores the pivotal role played by geographic chunks in the context of the CGR task. Irrespective of the approach used for segmentation, our framework consistently exhibits better performance, thereby reinforcing *Geo-Encoder*'s adaptability and efficacy. Therefore, our proposed framework transcends the confines of the Chinese task, and holds relevance and applicability to other languages or tasks characterized by sentence structures that align with linear-chain attributes.

[6]To ensure a fair comparison, we manually select relevant POS labels (e.g., quantity, noun, position, etc.), while excluding irrelevant ones (e.g., tone, punctuation, preposition, etc.). Further details can be found in the Appendix.

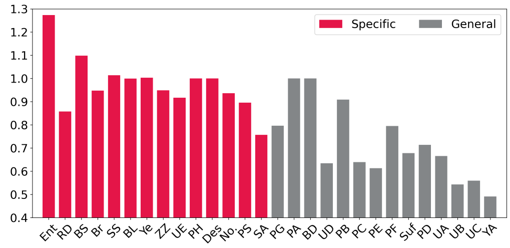
Method	GeoTES		GeoIND	
	Training (hour)	Inference (ms/case)	Training (hour)	Inference (ms/case)
Word2vec	–	5.9	–	3.5
Augment-Encoder	6.24	32.7	1.52	15.8
MEGO-BERT	4.50	33.8	0.92	18.9
Geo-Encoder (Ours)	5.94	35.6	1.25	19.5

Table 4: The statistics of training and inference time across different bi-encoder baseline models and our proposed *Geo-Encoder* on GeoTES and GeoIND datasets.

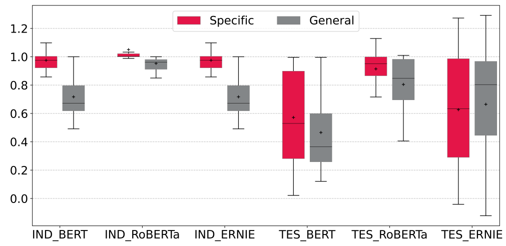
5.3 Parameter Sensitivity and Efficiency

Considering the pivotal impact of the dynamic attention matrix on model performance, we have conducted an additional experiment involving different update speed for model parameters and the attention matrix, which we called asynchronous learning rate updates. The outcomes, as is shown in Figure 3, underline the sub-optimal nature of synchronously updating metrics with model parameters (i.e. ratio=1). Contrarily, we have identified that employing a more extended update step for the attention matrix yields improved results; for instance, setting learning rate ratio at 10 and 2000 for the GeoIND dataset. This trend suggests that the attention matrix carries a weightier importance than general model parameters. Our finding is consistent with similar endeavors focused on adaptively weighted learning (He et al., 2017). Specifically, within our CGR task, a swifter acquisition of focus by the model on specific geographic chunks reveals to enhanced performance.

Furthermore, in line with our commitment to addressing real-world challenges, it becomes imperative to substantiate the efficacy of the proposed *Geo-Encoder*. To this end, we present an empirical analysis of training and inference times, as detailed in Table 4. Evidently, when comparing the results with MGEO-BERT, our training process exhibits a marginal increase in duration due to the incorporation of chunking attention matrix learning and supplementary representation fusion. However, it's noteworthy that our inference times remain remarkably similar, underscoring the effectiveness of our algorithm without causing substantial disparities in computational efficiency. The inference time of all models are acceptable for various industry application scenarios. Moreover, our training time is actually shorter than that of the Augment-Encoder approach (Peng et al., 2022), demonstrating the effectiveness of multi-task learning rather than geographic component feature concatenation.



(a) BERT chunk attention weights on GeoIND dataset



(b) Statistical distribution of attention matrix

Figure 4: Attention matrix weights visualization. We mark specific chunks as red and general chunks as grey. Weights of specific chunks are higher than general ones.

5.4 Chunking Weight Distribution

The attention matrix stands as a pivotal element warranting meticulous examination. Thus, this section delves into an in-depth analysis to discern whether the model demonstrates the capacity to effectively focus on specific chunks as opposed to the more general ones. Using the MGEO tagging tool, we manually labeled the subsequent categories as specific chunks: bus stations, house numbers, etc., encompassing a total of 14 distinct kinds. Conversely, the remaining chunks are classified as general (comprising 15 kinds), such as country, province, city, town, prefix, conjunction, etc.

For enhanced clarity, we manually categorize all chunk types into general and specific classifications, and present the BERT attention matrices in Figure 4(a) on GeoIND dataset. Notably, the trend discernible in this figure reveals that specific chunks (red) garner higher weights than general ones (grey). Further, we investigate the tendency across all models and datasets, as depicted in Figure 4(b). The congruence of these outcomes is evident, except for the case of ERNIE on the GeoTES datasets. This discrepancy aligns with the consistent low correlation scores observed between ERNIE and other models, as presented in Table 5.

Moreover, to probe the consistency across diverse learning processes, we compute spearman correlation coefficients (Spearman, 1961) across different datasets. Illustrated in Table 5, all of these correlation coefficients exhibit positive correlations and most of the results are statistically significant, underscoring uniform learning outcomes in component weights. It is worth noting that, except for the ERNIE model on the GeoTES dataset, the majority of models and datasets exhibit robust correlations, which is obviously evidenced by the high correlation scores. This result aligns with the observation that the ERNIE backbone model attains marginal enhancement, as shown in Table 2. Lastly, models trained on the same datasets yield notably high correlation scores among themselves. For instance, the scores between indBERT and indRoBERTa, and similarly between tesBERT and tesRoBERTa, surpass the 0.78 threshold.

Model	IndBERT	IndRoBERTa	IndERNIE
IndBERT	–	0.796*	0.785*
IndRoBERTa	0.796*	–	0.932*
IndERNIE	0.785*	0.932*	–

Model	TesBERT	TesBERTa	TesERNIE
TesBERT	–	0.819*	0.604*
TesRoBERTa	0.819*	–	0.374
TesERNIE	0.604*	0.374	–

Model	IndBERT	IndRoBERTa	IndERNIE
TesBERT	0.614*	0.409*	0.501*
TesRoBERTa	0.713*	0.634*	0.672*
TesERNIE	0.253	0.035	0.175

Table 5: Spearman correlation scores on GeoTES (Tes) and GeoIND (Ind) datasets. Statistically significant results are marked with *, where *p-value* < 0.05.

6 Conclusion

In this paper, we proposed a novel framework called *Geo-Encoder* for Chinese geographic re-ranking task by deploying multi-task learning module and synchronous update mechanism. The key idea is to encode geographic text using an additional component learning representations from address chunks. This approach allows the *Geo-Encoder* to effectively leverage linear-chain characteristic of geographic text, which guides the model to capture subtle distinctions among different candidates. Extensive experiments demonstrate that our proposed method leads to significant improvements over several competitive baselines. Future work could be incorporating our approach in multi-modal and multi-lingual tasks.

7 Limitation

While our work has achieved good performance and shown promising results in enhancing Chinese geographic re-ranking task through incorporation of geographic representations, there are still limitations in our work. Specifically, the Geo-Encoder we have developed exhibits a specificity towards textual data possessing linear-chain or structural characteristics, thereby constraining the method's applicability primarily to within-domain scenarios. However, we believe that this study is still useful in highlighting the challenges of geographic encoding. Moreover, our approach demonstrates notable effectiveness and efficiency when employed in industrial applications, owing to its minimal augmentation of parameters.

In the future, we plan to explore the feasibility of collecting multi-modal datasets, which can be potential to provide further insights into incorporating geographic understanding with our proposed framework into CGR task.

References

Steven P Abney. 1991. Parsing by chunks. In *Principle-based parsing*, pages 257–278. Springer.

Marco Avvenuti, Stefano Cresci, Leonardo Nizzoli, and Maurizio Tesconi. 2018. Gsp (geo-semantic-parsing): geoparsing and geotagging with machine learning on top of linked data. In *European Semantic Web Conference*, pages 17–32. Springer.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. 2023. Mgeo: A multi-modal geographic pre-training method. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

Keke He, Zhanxiong Wang, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. 2017. Adaptively weighted multi-task deep network for person attribute classification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1636–1644.

Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B Pierrehumbert, and Hinrich Schütze. 2022. Geographic adaptation of pretrained language models. *arXiv preprint arXiv:2203.08565*.

Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3029–3039.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2101–2110, Copenhagen, Denmark. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Ya-Hui Jia, Wei-Neng Chen, Tianlong Gu, Huaxiang Zhang, Huaqiang Yuan, Ying Lin, Wei-Jie Yu, and Jun Zhang. 2017. A dynamic logistic dispatching system with set-based particle swarm optimization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9):1607–1621.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Hao Li, Wei Lu, Pengjun Xie, and Linlin Li. 2019. Neural Chinese address parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3421–3431, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Liu, Juan Hu, Qi Shen, and Huan Chen. 2021. Geo-BERT pre-training model for query rewriting in POI search. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2209–2214, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5334–5343.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Efficient document re-ranking for transformers by precomputing term representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 49–58, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ping Nie, Yuyu Zhang, Xiubo Geng, Arun Ramamurthy, Le Song, and Daxin Jiang. 2020. Dc-bert: Decoupling question and document for efficient contextual encoding. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1829–1832.

Toru Nishino, Shotaro Misawa, Ryuji Kano, Tomoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2019. Keeping consistency of sentence generation and document classification with multi-task learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3195–3205, Hong Kong, China. Association for Computational Linguistics.

Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-argument based bi-encoder for paraphrase identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Charles Spearman. 1961. The proof and measurement of association between two things.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.

Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2009. A discriminative latent variable chinese segmenter with hybrid word/character information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 56–64.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Boli Chen, Pengjun Xie, Fei Huang, and Min Zhang. 2022a. Forging multiple training objectives for pre-trained language models via meta-learning. *CoRR*, abs/2210.10293.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Pengjun Xie, Fei Huang, and Min Zhang. 2022b. Adversarial self-attention for language understanding.

Tingyu Xia, Yue Wang, Yuan Tian, and Yi Chang. 2021. Using prior knowledge to guide bert's attention in semantic textual matching tasks. In *Proceedings of the Web Conference 2021*, pages 2466–2475.

Ming Xu. 2023. Text2vec: Text to vector toolkit. https://github.com/shibing624/text2vec.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1154–1156.

Wenwen Ye, Yiding Liu, Lixin Zou, Hengyi Cai, Suqi Cheng, Shuaiqiang Wang, and Dawei Yin. 2022. Fast semantic matching via flexible contextualized interaction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1275–1283.

Zixuan Yuan, Hao Liu, Yanchi Liu, Denghui Zhang, Fei Yi, Nengjun Zhu, and Hui Xiong. 2020. Spatio-temporal dual graph attention network for query-poi matching. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 629–638.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 421–431, Berlin, Germany. Association for Computational Linguistics.

Hai Zhao, Changning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective tag set selection in chinese word segmentation via conditional random field modeling. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 87–94.

Ji Zhao, Dan Peng, Chuhan Wu, Huan Chen, Meiyu Yu, Wanji Zheng, Li Ma, Hua Chai, Jieping Ye, and Xiaohu Qie. 2019. Incorporating semantic similarity with geographic correlation for query-poi relevance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1270–1277.

A Appendix

A.1 Dataset Details

As previously mentioned, we utilize the MGEO geographic tagging tool[7] to thoroughly annotate the provided geographical text. To elaborate further, we present a demonstrative instance in Table 11. This example highlights the effectiveness and comprehensive nature of the MGEO in annotating geographical information within the text.

[7] https://modelscope.cn/models/damo/mgeo_geographic_elements_tagging_chinese_base.

A.2 POS Implement

We utilize the Jieba tagging tools[8] , which enable the segmentation of all geographical text into meaningful segments. Following this initial breakdown, a rigorous selection process is undertaken, wherein 28 specific parts-of-speech categories are identified as pertinent and aligned with our Geo tagging system. These categories are chosen based on their close relevance to geographical references, thereby ensuring the precision of the tagging process. A comprehensive list of these valid part-of-speech tags is provided in Table 9.

In this context, it's important to emphasize that even though manual selection involves a degree of subjectivity, we have maintained consistent tag categories with geographical references to ensure a fair comparison. Additionally, although certain POS tags may not directly pertain to geographic terminology, we have arranged them based on their relative correlations across all POS tags. We have also provided a list of POS tags that are deemed invalid in Table 10, consisting of 24 specific parts-of-speech categories.

Moreover, we compute the fuzzy similarity[9] between the results of POS tagging and Geo chunking, as shown statistically in Table 6.

Set	Avg. Geo	Avg. POS	Similarity
GeoTES			
Train	5.11	10.71	80.56 ± 7.39
Dev	4.69	9.47	80.46 ± 7.35
Test	4.66	9.41	80.60 ± 7.41
GeoIND			
Train	4.38	8.59	78.50 ± 6.46
Dev	4.38	8.60	79.71 ± 6.65
Test	4.37	8.57	79.77 ± 6.68

Table 6: Valid POS categories and their respective definitions, comprising a total of 28 categories.

As depicted in Table 6, it becomes evident that the average count of Geo chunking units is less than that of POS. Concurrently, a noteworthy inference can be drawn that the chunking outcomes exhibit resemblance. This is supported by the substantial similarity scores (exceeding 78.00) between the results on both datasets.

[8] POS tagging is based on jieba: https://github.com/fxsjy/jieba.

[9] https://pypi.org/project/fuzzywuzzy/

Parameter	GeoTES	GeoIND
Learning rate(BERT)	$5e^{-5}$	$5e^{-5}$
Learning rate(RoBERTa)	$1e^{-5}$	$1e^{-5}$
Learning rate(ERNIE)	$5e^{-5}$	$5e^{-5}$
Batch size	32	32
Test Batch size	16	16
Early Stop	3	3
Embed_dim	256	256
Optimizer	AdamW	AdamW
Attn_init	1.0	1.0
Weight_decay	0.02	0.02

Table 7: The hyper-parameters of the best results on GeoTES and GeoIND dataset.

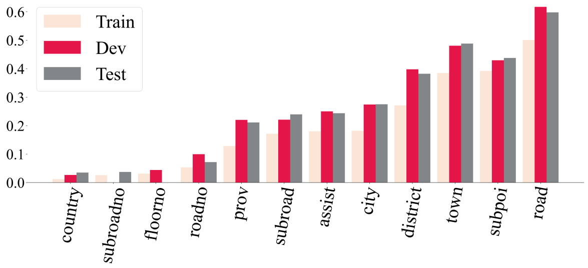


Figure 5: The information entropy of Li et al. (2019), indicate that specific chunks (e.g., *road*) exhibit greater diversity compared to general ones (e.g., *country*).

Chunks	Definition
	General
PA	Country
PB	Province
PC	City
PD	District
PE	Township
PF	Street
PG	Village
PH	Administrative Term / Business District
PS	Other Administrative Term
UA	Door Address: Road xx, No.xx / Lane xx
UB	Door Address: Building xx / Area xx
UC	Door Address: Building No. xx
UD	Door Address: Additional Description
	Specific
BS	Bus Station
BL	Bus and Subway Route
RD	Road, Highway, Furuin Street, Tunnel, Bridge, Overpass
Entity	General Name for Point of Interest (POI)
Brand	Well-known Brand
CategorySuffix	Category Suffix Word
Ent	Point of Interest (POI)
Br	Brand
No.	Number
UE	Door Address: East Entrance, South Gate
SA	Direction Modifier
PH	Administrative Term / Business District
Ye	Semantic Connector
Des	Descriptor
ZZ	Unknown

Table 8: Translation of Chunking Terms.

A.3 Geo Chunks

We have compiled a comprehensive table (Table 8), that outlines various chunking categories along with their corresponding definitions of Geo chunks. Drawing from our accumulated expertise, we have classified all chunk categories into two distinct groupings: "general" and "specific."

This categorization is guided by a systematic process that sorts these categories based on their relative degrees of correlation. To elaborate on this process, we strategically designate the first 50% of the selection as general chunks, while the subsequent 50% are categorized as specific chunks. By employing this division strategy, we achieve a balanced representation of both general and specific chunk types.

A.4 Entropy of Geo Chunks

Most current attempts directly deploy PTMs to encode geographic texts into embeddings (Yuan et al., 2020; Huang et al., 2022; Ding et al., 2023), ignoring the linear-chain structure characteristic of geographic text. To quantify this distinction, we calculate the entropy score of geographic chunking datasets from (Li et al., 2019) as shown in Figure 5. Obviously, the specific chunks (e.g. road, town, etc.) hold a higher entropy score among all sets, revealing more diversity than the general chunks (e.g. country, province, etc.). Therefore, it can be further inferred that specific chunk components contribute unequally to the semantic representation of sentences, indicating that specific chunks play a more substantial role than general ones.

A.5 Hyper-parameter Setting

In an effort to support the reproducibility of the *Geo-Encoder* and its demonstrated reasoning performance, we are providing a compilation of the optimal hyperparameters that yielded the best outcomes on two benchmark datasets, as illustrated in Table 7.

In the process of establishing the baseline, it's important to note that all scores presented in Table 2 have undergone training and validation on a consistent hardware platform. Additionally, we are committed to making our baseline code publicly available for reference, which will coincide with the release of our paper.

A.6 More Discussion

Chunking Contribution Learning Due to the components of each geographic text being quite different, introducing feature concatenation strategy in CGR task is not reasonable. Therefore, we pro-

Valid POS tag	Definition
nz	Other Proper Noun
a	Adjective
m	Numeral
q	Measure Word
t	Time Word
mg	Measure Word for Quantity
ns	Place Name
ng	Noun as Morpheme
ag	Adjective as Morpheme
f	Locative
z	Status Word
nt	Organization Name
eng	English Word
an	Noun
mq	Measure Word for Quantity
ad	Adverb as Adjective
b	Differentiation Word
j	Abbreviation
n	Noun
c	Conjunction
uv	Auxiliary Word
k	Following Part
h	Preceding Part
v	Verb
uz	Status Word
ug	Tense Word
df	Differentiation Word
yg	Modal Particle

Table 9: Valid POS categories and their respective definitions, comprising a total of 28 categories.

Invalid POS tag	Definition
e	Interjection
i	Idiom
d	Adverb
l	Idiomatic Expression
p	Preposition
u	Particle
y	Modal Particle
g	Morpheme
x	Non-Morpheme Character
vg	Verbal Morpheme
vn	Nominal Verb
zg	State Morpheme
r	Pronoun
dg	Adverbial Morpheme
tg	Tense Morpheme
o	Onomatopoeia
uj	Particle
ud	Particle
nr	Personal Name
rg	Modal Particle
ul	Tense Particle
s	Locative Noun
nrt	Personal Name
nrfg	Personal Name

Table 10: Invalid POS categories and their respective definitions, consisting of a total of 24 categories.

posed to use an universal component embeddings for queries U^Q and candidates U^C, and initialize them as zero matrices. It follows that empty components would yield no contributions to the final representations. Similarly, components that do not align appropriately between the queries and candidates would also have no impact.

Asynchronous Update Mechanism Our insights is that the fast distinction of specific geographic chunks should conceivably be more amenable and expedited for the model's learning process. Consequently, the matrix W_U could feasibly adapt to more substantial increments in learning steps compared to those attributed to language model parameters.

Training and Inference The rationale for introducing components stems from a deliberate consideration of the trade-off between training and inference aspects. The underlying objective is to facilitate the model in exhibiting a heightened sensitivity towards specific chunks as opposed to general ones. This endeavor has yielded demonstrably effective outcomes in our experimental evaluations. Conversely, during the inference phase, we eliminate the necessity for component predictions, thereby leading to a marked improvement in computational efficiency. This assertion will be substantiated in the subsequent section.

Field	Content
Query	浙江省杭州市人民检察北东院侧广播电视台东门南 South of the East Gate of People's Procuratorate North East Radio and Television Station, Hangzhou City, Zhejiang Province.
Query_Geo_Chunks	浙江省-prov / 杭州市-city / 人民检察-poi/ 东院-subpoi / 侧-assist / 广播电视台-subpoi / 东门-subpoi / 南-assist Zhejiang Province / Hangzhou City / People's Procuratorate / East Door / of / Radio and Television Station / East Gate / South Procuratorate of Hangzhou City, Zhejiang Province.
Query_POS_Chunks	浙江省-ns / 杭州市-ns / 人民-n / 检察-vn / 北东-ns / 院侧-n / 广播-vn / 电视台-n / 东门-ns / 南-ns Zhejiang Province / Hangzhou City / People / Procuratorate / North East / of / Radio Television Station / East Gate / South Procuratorate of Hangzhou City, Zhejiang Province.
Candidates	浙江省人民北路路旁播州区人民检察院 People's Procuratorate of Bozhou District, beside Renmin North Road, Zhejiang Province. 浙江省人民检察院 Zhejiang Provincial People's Procuratorate. 浙江省浙江北路136号山东广播电视台 Shandong Radio and Television Station, No. 136 Zhejiang North Road, Zhejiang Province. 台州路1号杭州市拱墅区人民检察院 People's Procuratorate of Gongshu District, Hangzhou City, No. 1 Taizhou Road.
Candidates_Geo_Chunks	浙江省-prov / 人民北路-road / 路旁-assist / 播州区人民检察院-poi Zhejiang Province / Renmin North Road / beside / People's Procuratorate of Bozhou District. 浙江省-prov / 人民检察院-poi Zhejiang Province / Provincial People's Procuratorate. 浙江省-prov / 浙江北路-road / 136号-roadno / 山东广播电视台-poi Zhejiang Province / Zhejiang North Road / No. 136 / Shandong Radio and Television Station 台州路-road / 1号-roadno / 杭州市-city / 拱墅区-district / 人民检察院-poi Taizhou Road / No. 1 / Hangzhou City / Gongshu District / People's Procuratorate
Candidates_POS_Chunks	浙江省-ns / 人民-n / 北路-ns / 路旁-s / 播州-ns / 区-n / 人民检察院-nt Zhejiang Province / Renmin / North Road / beside / Bozhou / District / People's Procuratorate. 浙江省-ns / 人民检察院-nt Zhejiang Province / Provincial People's Procuratorate. 浙江省-ns / 浙江-ns / 北路-ns / 136-m / 号-m / 山东-ns / 广播-vn / 电视台-n Zhejiang Province / Zhejiang / North Road / 136 / No. / Shandong / Radio / Television Station 台州-ns / 路-n / 1-m / 号-m / 杭州市-ns / 拱墅区-ns / 人民检察院-nt Taizhou / Road / 1 / No. / Hangzhou City / Gongshu District / People's Procuratorate

Table 11: A representative illustration sourced from the GeoTES dataset is provided. We are showcasing a subset of potential options in this context. The English was meticulously translated, as this information isn't inherently present in our initial dataset.