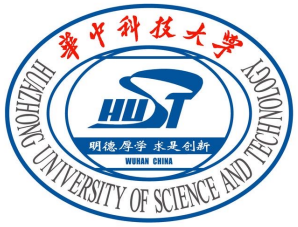


Geo-Encoder: A Chunk-Argument Bi-Encoder Framework for Chinese Geographic Re-Ranking

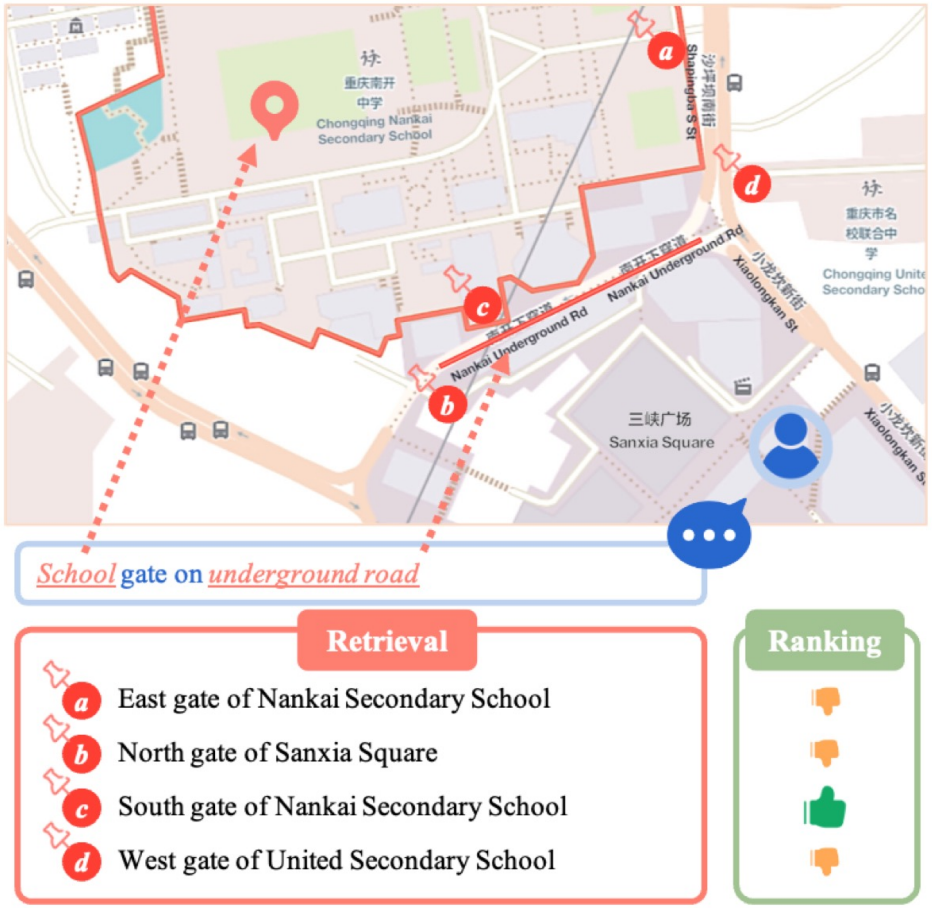
Yong Cao[1] , Ruixue Ding[2] , Boli Chen[2] , Xianzhi Li[1], Min Chen[3] , Daniel Hershcovich[4], Pengjun Xie[2], and Fei Huang[2]

[1] Huazhong University of Science and Technology
[2] Alibaba Group, Hangzhou, China
[3] School of Computer Science and Engineering, South China University of Technology
[4] Department of Computer Science, University of Copenhagen

Task

- Geographic re-ranking is a critical aspect of recent information retrieval systems

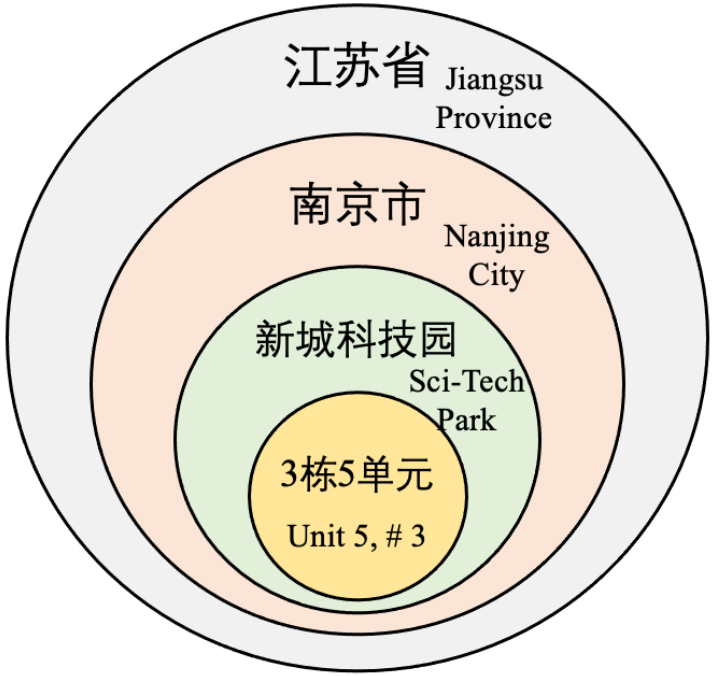
- Optimization of Chinese geographical sentence representations based solely on text remains underexplored.



Problem Statement

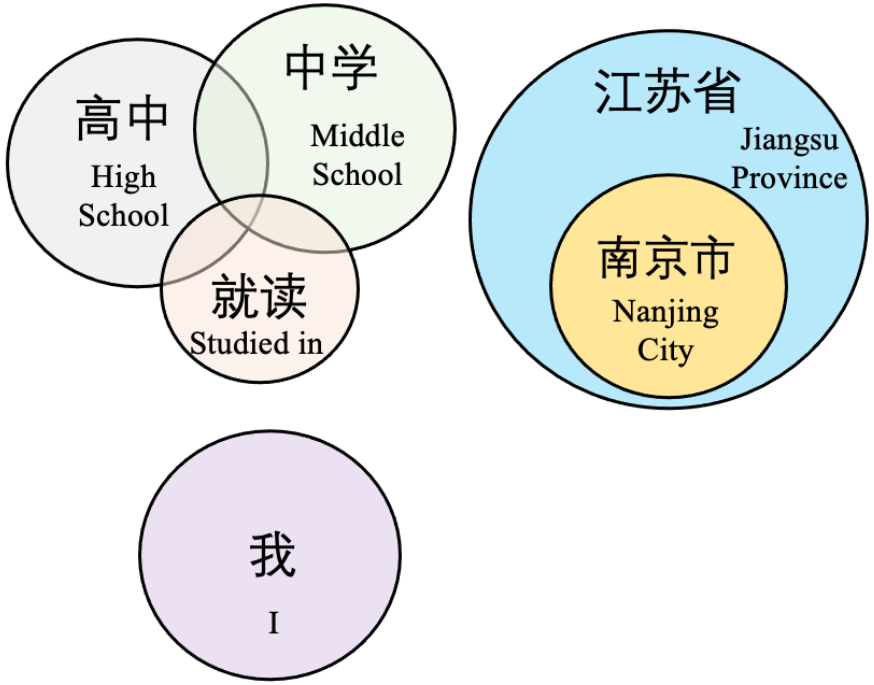
- Chinese geographic re-ranking faces unique challenges
  - Linear-chain structure
  - Chunks Contribution
  - Lack of Standardization

Linear-chain structure



江苏省南京市新城科技园3栋5单元。

Unit 5, Building 3, Sci-Tech Park, Nanjing City, Jiangsu Province.

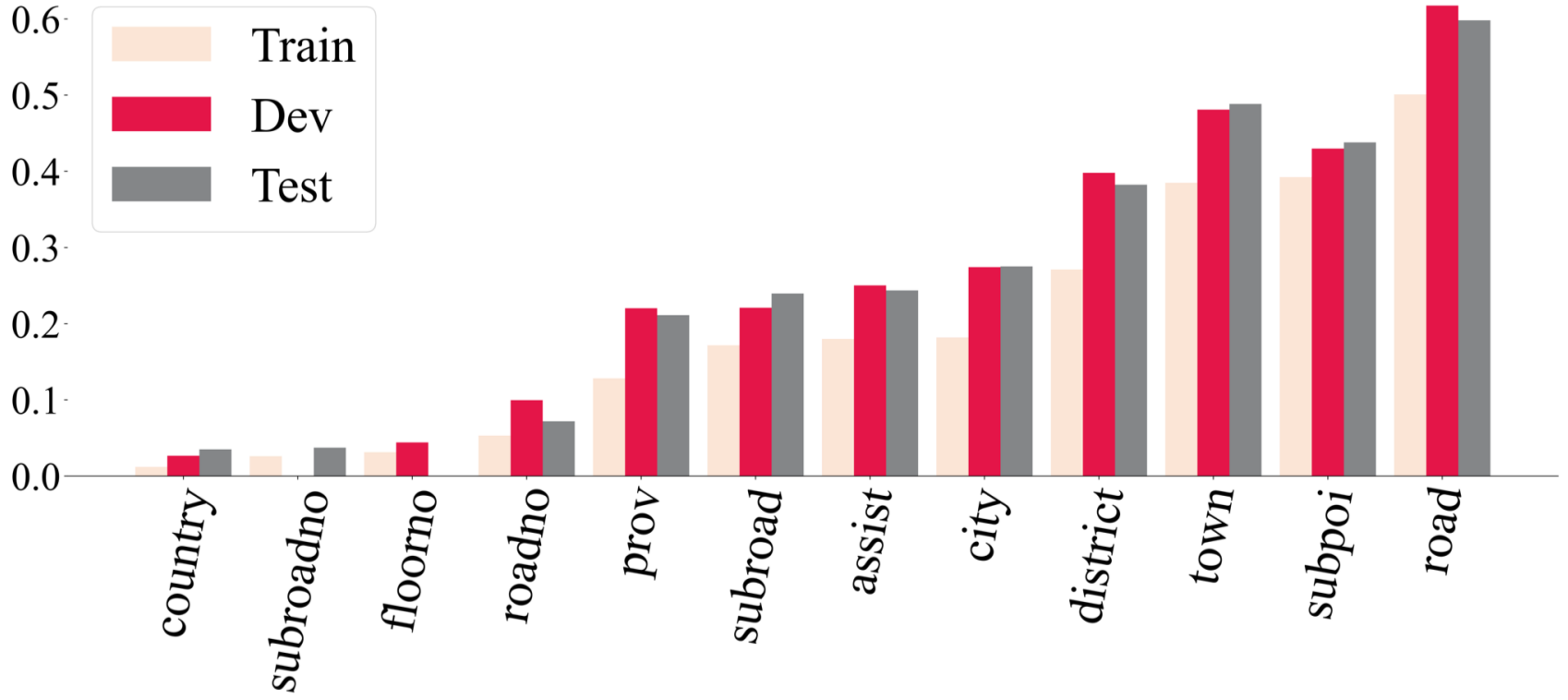


我高中就读于江苏省南京市的中学。

My high school is one of middle schools of Nanjing, Jiangsu Province.

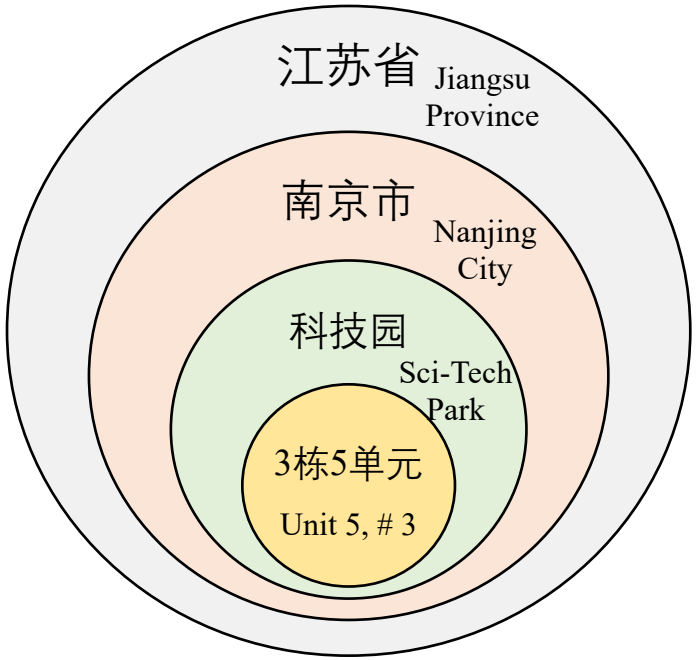
Chunks Contribution

Specific chunks (e.g., road) exhibit greater diversity compared to general ones (e.g., country).



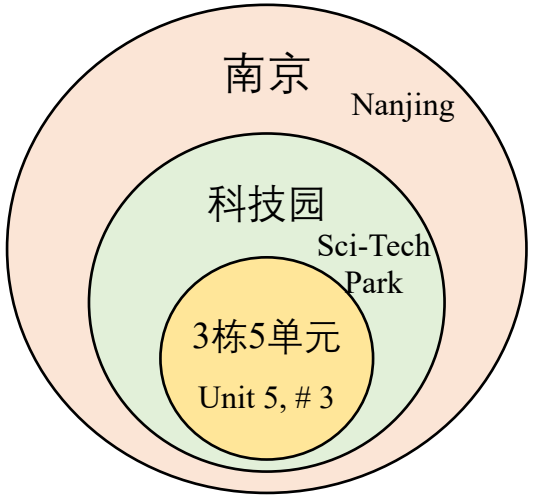
The information entropy of Chinese Address dataset

Lack of Standardization



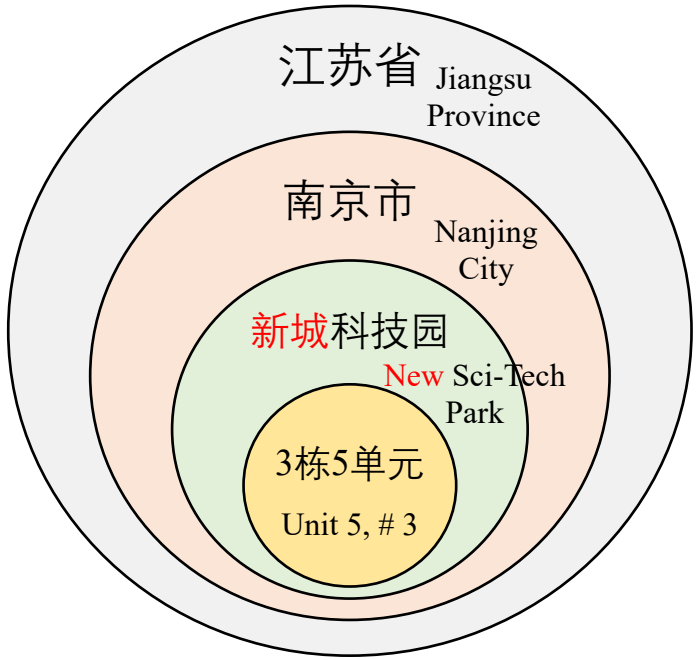
江苏省南京市科技园3栋5单元。

Unit 5, Building 3, Sci-Tech Park, Nanjing City, Jiangsu Province.



~~江苏省~~南京~~市~~科技园3栋5单元。

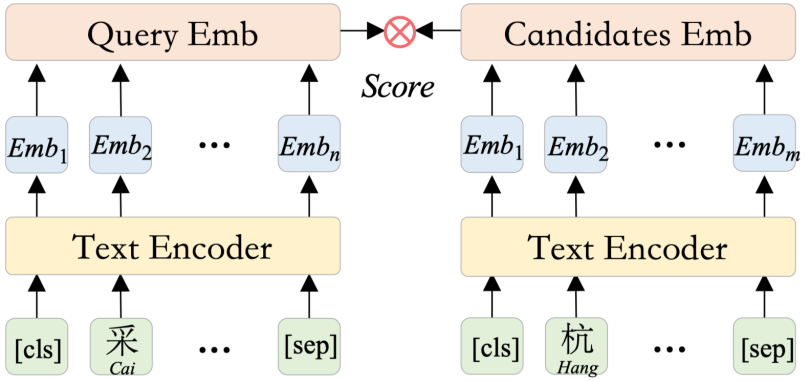
Unit 5, Building 3, Sci-Tech Park, Nanjing ~~City, Jiangsu Province.~~



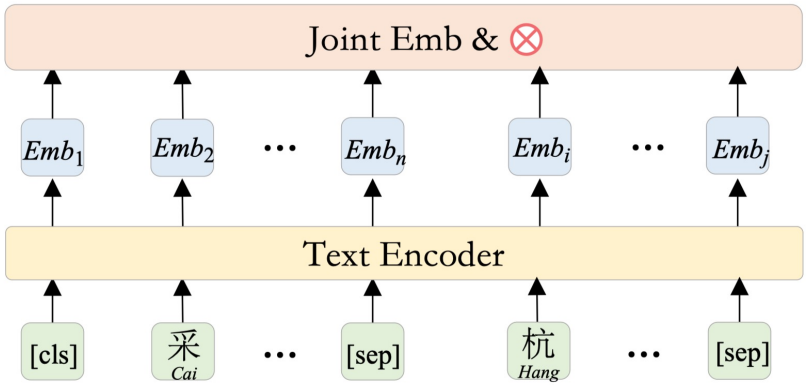
江苏省南京市新城科技园3栋5单元。

Unit 5, Building 3, New Sci-Tech Park, Nanjing City, Jiangsu Province.

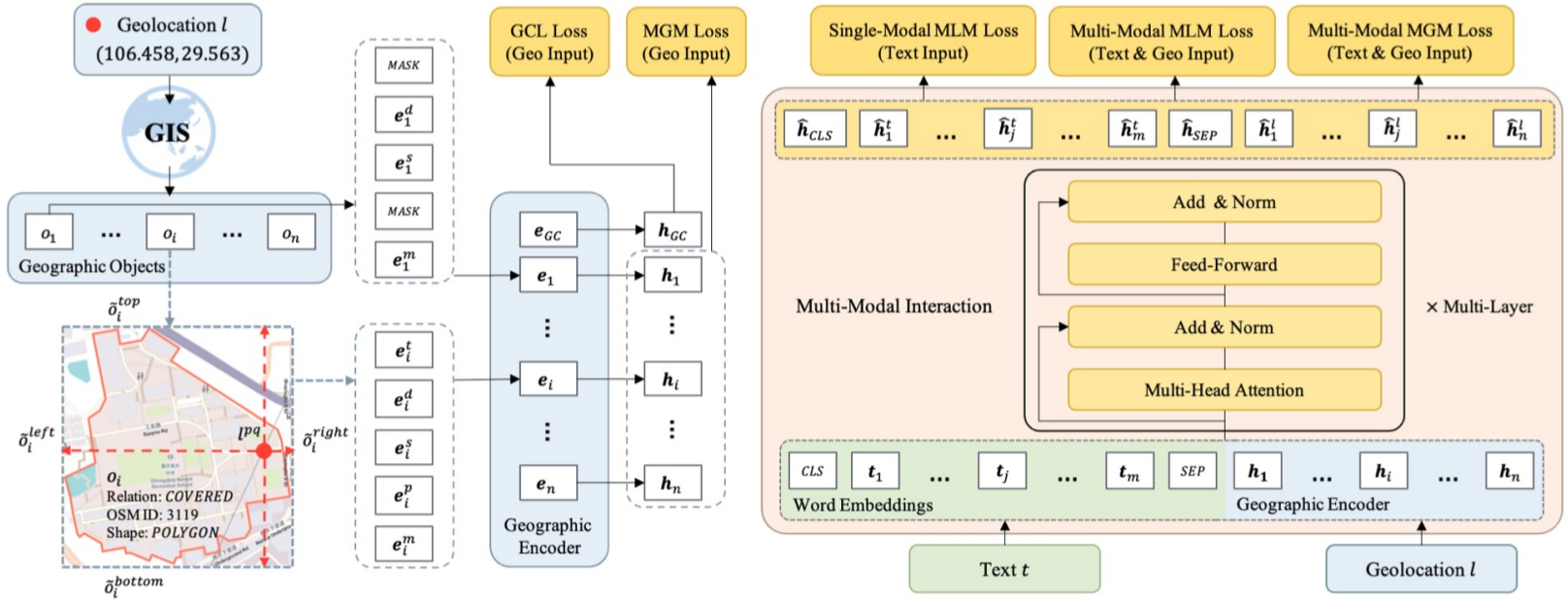
Background



(a) Bi-Encoder



(b) Cross-Encoder



- Existing framework and approaches: Simple Sentence Representation Based on BERT [CLS]

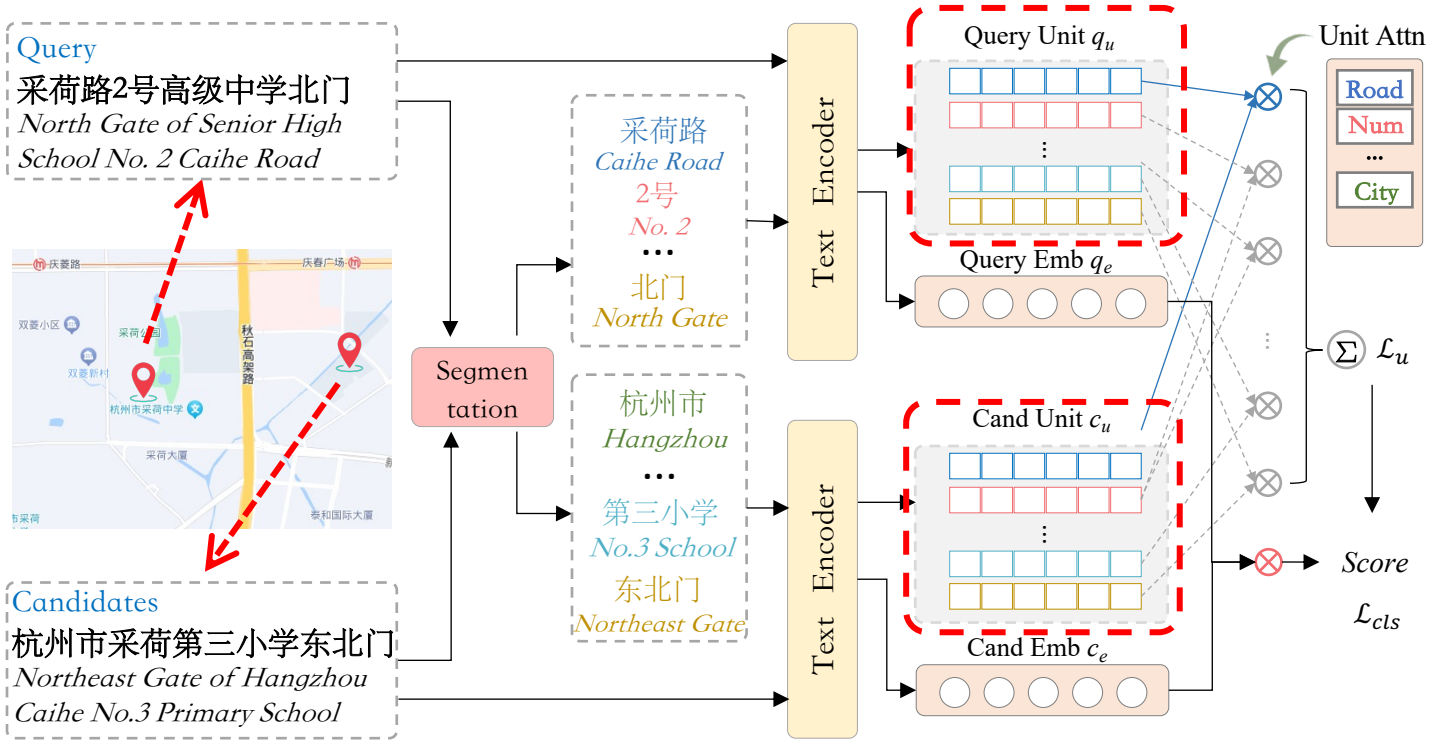
Methodology

- Chunking Contribution Learning

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

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

$$Score_u = (U^Q * W^U) * (U^C * W^U)$$



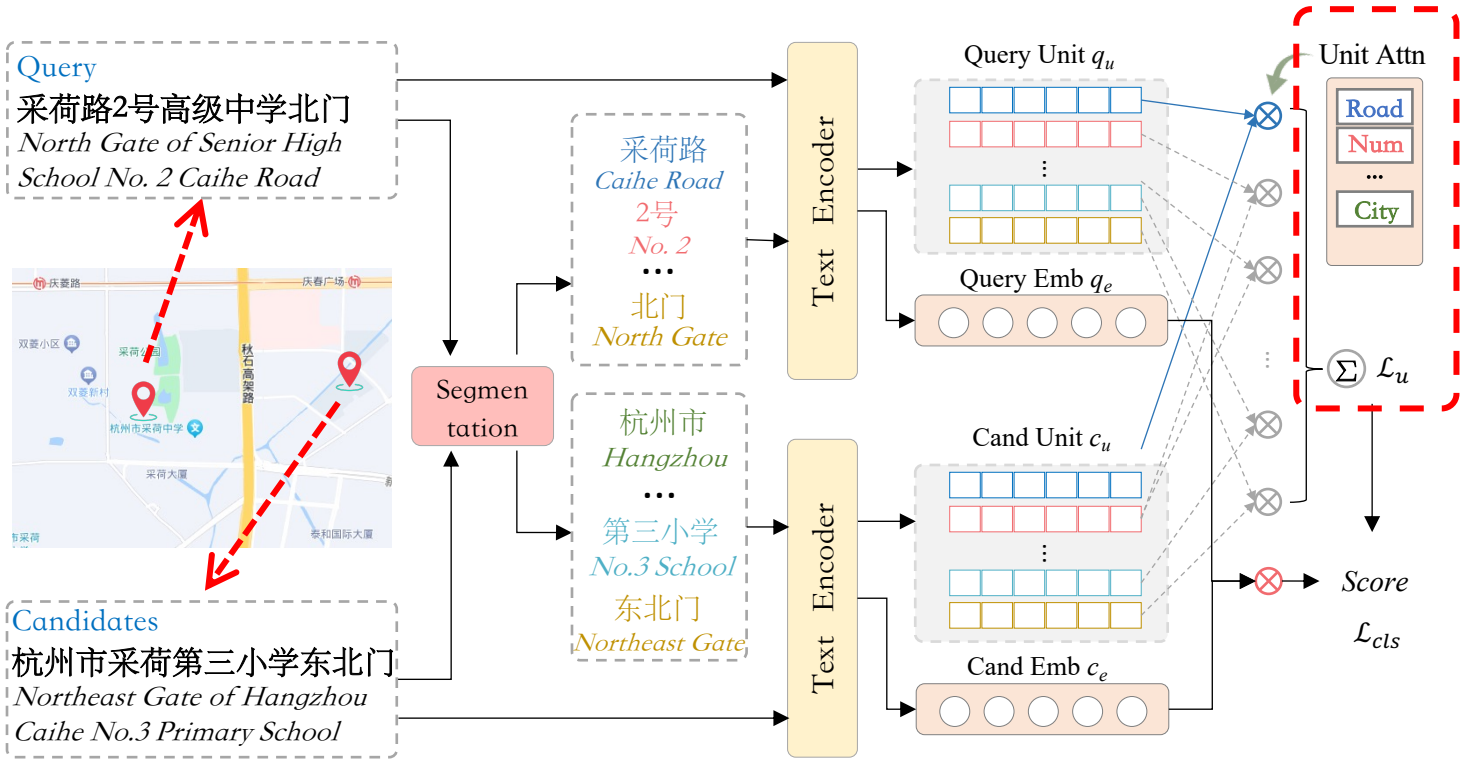
Methodology

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

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

- Asynchronous Update Mechanism

$$w'_u = w_u + \lambda \cdot \nabla w_u \cdot \gamma$$



Dataset

- GeoTES:
  - A widely recognized, large-scale benchmark dataset
- GeoIND
  - Our collected moderately-sized, real-world industry dataset

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

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.
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.

Evaluation Metrics

- Hit@K (k=1,3)

$$HR@K = \frac{NumberOfHits@K}{GT}$$

- NDCG@1 Normalized Discounted cumulative

$$CG_k = \sum_{i=1}^{k} rel(i) \qquad DCG_k = \sum_{i=1}^{k} \frac{rel(i)}{log_2(i+1)} \qquad NDCG = \frac{DCG}{IDCG}$$

- MRR@3 Mean Reciprocal Rank

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Main Results

- Our proposed approach achieves a remarkable state-of-the-art performance across all evaluated metrics.
- RoBERTa performs emerges as the superior candidate

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.

Ablation Study

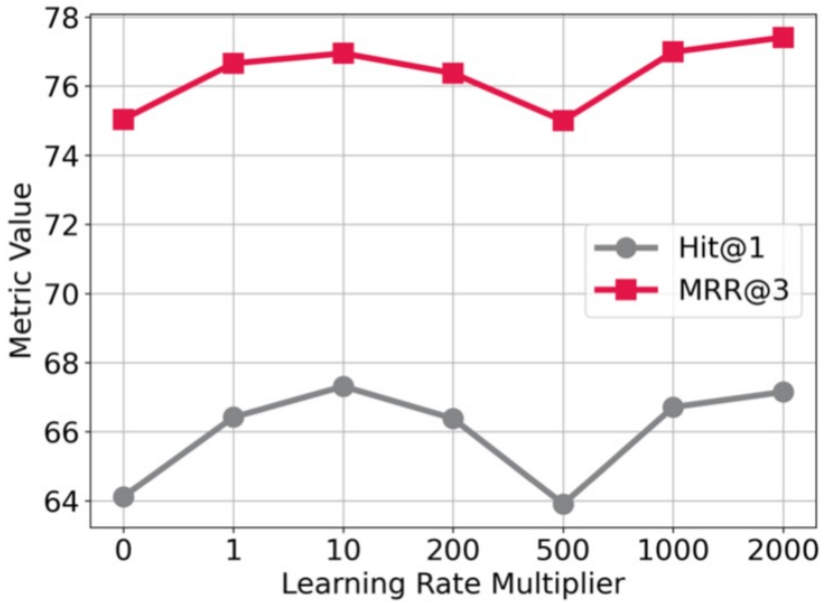
- Fix Contribution vs. Learning Weight
- Geo Chunking vs. General Chunking

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**

Ablation Study

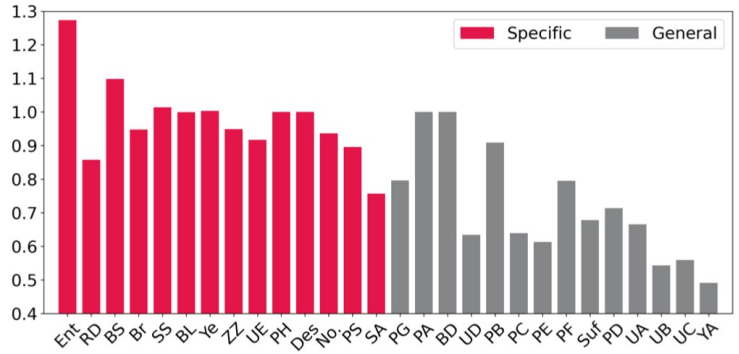
- Sub-optimal nature of synchronously updating metrics with model parameters.

- Inference times remain remarkably similar

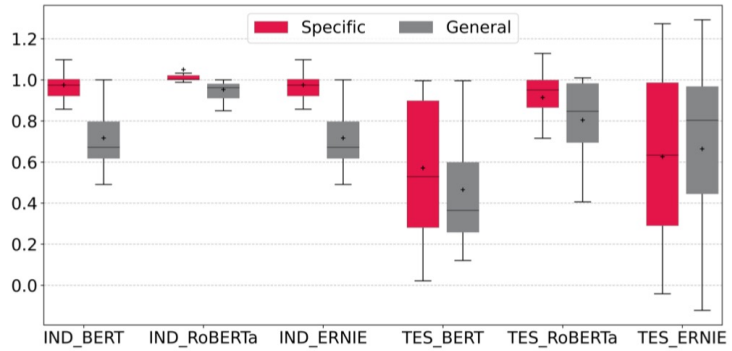


Method	GeoTES		GeoIND	
	Training (hour)	Inference (ms/case)	Training (hour)	Inference (ms/case)
Word2vec	–	5.9	–	3.5
Augment-Encoder	6.24	32.7	1.52	15.8
MEGO-BERT	4.50	33.8	0.92	18.9
Geo-Encoder	5.94	35.6	1.25	19.5

Chunking Weight Distribution



(a) BERT chunk attention weights on GeoIND dataset



(b) Statistical distribution of attention matrix

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	–

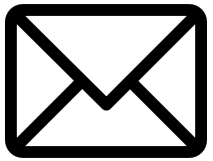
Model	IndBERT	IndRoBERTa	IndERNIE
TesBERT	0.614*	0.409*	0.501*
TesRoBERTa	0.713*	0.634*	0.672*
TesERNIE	0.253	0.035	0.175

Conclusion

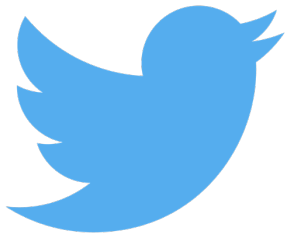
- A novel framework called Geo-Encoder for Chinese geographic reranking task
- Chunks Contribution Learning
- Asynchronous update mechanism
- A real-world CGR dataset

Thanks! Feel free to reach out.

yongcaoplus.github.io



yongcao@di.ku.dk



@YongCaoPlus

Yong Cao