Spatial Semantic-based Enhanced Address Parsing via Adaptive Weighted Learning

Huiling Qin
Beijing Normal University
Zhuhai, Guangdong, China
qinhuiling@bnu.edu.cn

Ming Wang
Xidian University
Xi'an, Shannxi, China

Yuanxun Li
King Soft
Zhuhai, Guangdong, China

Junbo Zhang
JD ICity, JD Technology
Beijing, China

Yu Zheng
JD ICity, JD Technology
Beijing, China

Abstract

Address parsing is an essential task that transforms natural language descriptions into standardized addresses, crucial for numerous urban applications. Existing methods struggle with ambiguous expressions, and even Large Language Models face challenges adapting to specialized domains with limited data. In this study, we focus on developing a robust framework to map diverse address descriptions into a unified semantic space of standardized addresses. We propose the Adaptive Weighted Learning-based Address Parsing (AWLAP) framework, which enhances parsing effectiveness through two key components: a multi-level constrained classifier that mines correlations between geographic entities across hierarchies, and an integrated discriminator that adaptively guides optimization based on parsing complexity. We evaluate the AWLAP using real data from JD Logistics and Point-of-Interest addresses. Extensive experiments comparing against state-of-the-art methods demonstrate AWLAP's effectiveness and robustness in address parsing. The proposed AWLAP framework has been successfully deployed as an address parsing service in practical applications.

CCS Concepts

• **Information systems** → *Structured text search.*

Keywords

Address parsing, Cooperation learning, Pretrain language model

ACM Reference Format:
Huiling Qin, Ming Wang, Yuanxun Li, Junbo Zhang, and Yu Zheng. 2025. Spatial Semantic-based Enhanced Address Parsing via Adaptive Weighted Learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746252.3761499

1 Introduction

Parsing natural language descriptions into standard addresses is crucial for many urban applications[7, 29]. For example, address standardization during the express delivery process can reduce losses caused by package misdelivery or returned mail due to inaccurately described addresses. Accurate address information is also crucial in government services such as population statistics, social services, and emergency response. Address parsing serves as the foundation for implementing these services. The automated parsing process can quickly respond to real-time address parsing demands, thereby reducing the cost of manual parsing. However, people express addresses in various ways in natural language, with significant differences in semantic information and grammatical structures. Existing methods typically focus on either using numerous rules for fuzzy address matching[3, 18], or training a model based on a large corpus applicable to address parsing scenarios[28, 29], enabling it to "understand" standard addresses referred to in natural language.

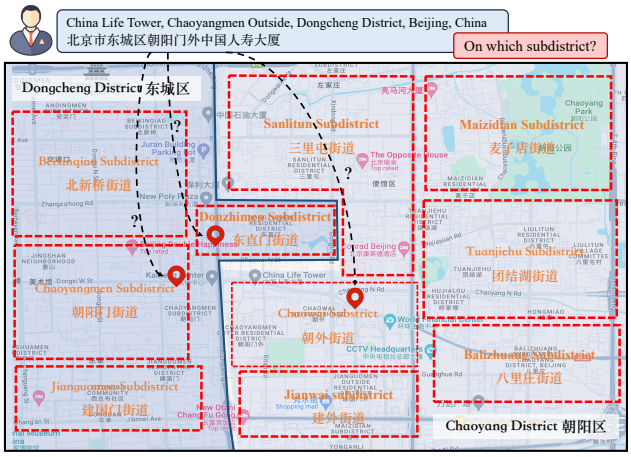


Figure 1: An illustration of the hierarchical parsing process for an ambiguous query address. Different shades of blue boxes represent different regions of the same hierarchy. And the red dashed boxes are sub-regions of the blue ones.

Despite some successful efforts in certain specific scenarios, accurate address parsing remains a challenging task due to the many difficulties presented by real-world, ever-changing geographical environments and inconsistent address description formats in different scenarios: **(1) Diversity of descriptions.** When describing the address of a geographic entity, there are typically multiple ways to do so. These may include descriptions based on geographic hierarchical relationships, the structure of the road network, or the

CIKM '25, November 10–14, 2025, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761499

surrounding points of interest. These different descriptive customary make natural language descriptions of addresses diverse and difficult to parse with a uniform rule. **(2) Unclear or incorrect address data descriptions.** Since addresses do not belong to a single individual, there is a lack of motivation for people to memorize standardized addresses. In addition, geographic location information is subject to change, making it even more difficult to remember. As a result, problems such as missing information or incorrect descriptions are common in everyday address descriptions. **(3) Conflicts in address hierarchy descriptions.** When describing geographic entities based on geographic affiliation or administrative jurisdiction, there is often a hierarchical constraint relationship. If different descriptive ways are used, the same entity may be attributed to different parent regions. In addition, the attribution of geographic entities may change as standards evolve. As a result, geographic hierarchy conflicts often arise when people describe addresses.

To address the aforementioned challenges, we propose a new Adaptive Weighted Learning-based Address Parsing (AWLAP) framework to solve the problem of mapping from natural language descriptions to standard addresses. We focus on the robust parsing of diverse and unclear natural language addresses. (1) The AWLAP framework utilizes a pre-trained language model as an address encoder to uniformly encode diverse descriptions into address embeddings. By learning the spatial semantic information hidden in the address embeddings, a downstream classification task is constructed to map an infinite set of natural language descriptions to a finite set of standard addresses, thus adapting the model to the transition from natural language descriptions to standard addresses. (2) To alleviate the hierarchy conflicts problem in the parsing process, we design a multi-level constrained classifier that splits the classification task into multiple subtasks based on different geographic levels and sets hierarchical constraints between neighboring subtasks to improve the accuracy of the classifier for address resolution. (3) A discriminative task is introduced to optimize the learning of the difficult address (ambiguous description) in the classification task by using a simple multi-layer perceptron that receives feedback on the classification performance from samples with different addresses. Our main contributions are summarized as follows:

- We propose the AWLAP framework, which enhances address parsing through two key innovations: a specialized classifier built on pre-trained language models that mitigates noise by capturing geographic hierarchical relationships, and a discriminator that dynamically adjusts classification optimization weights to improve parsing effectiveness.
- We introduce an adaptive weighted learning strategy based on a discriminative task to help the classifier learn differentially under different address samples, which improves the accuracy and efficiency of address parsing.
- Our AWLAP-based address parsing service has been successfully deployed in multiple cities, significantly improving local government address management efficiency. The system seamlessly integrates with existing government infrastructure, reducing manual address verification workload, while providing public access through a WeChat mini-program.
- We conduct extensive experiments using a real-world address descriptions dataset. The results show that the proposed AWLAP achieves superior parsing accuracy over all the baseline methods and it's variants in different cases.

2 PRELIMINARIES

Definition 2.1. **Natural Language Description Address.** A natural language description address is a class of non-standard address. It may contain a variety of detailed or unclear information, but does not conform to a formal, standardized format. It is the input of address parsing and denoted as $X = \{x_1, x_2, \cdots, x_n\}$, n is the number of natural language description addresses.

Definition 2.2. **Standard Address.** A standard address is an address written according to a uniform rule and format. Its definition may vary in different countries, but it is typically organized based on the hierarchical relationships between geographic entities, such as *"country - province/state - city - district - town/street - community - building/landmark"*. We use standard addresses as the labels for training and testing, defined as $Y = \{y_1, \cdots, y_n | y_i = [y_{i1}, \cdots, y_{il}]\}$, l represents the different geographic levels.

Definition 2.3. **Match Mask.** We define a match mask M, with $M = \left[m_{x,y} | x \in X, y \in Y\right]$ denoting the matchability of the natural language description and standard addresses. We set $m_{x,y} = 1$ if y is the correct standard address referred to by x, and $m_{x,y} = 0$ if y is incorrect.

Our goal is to parse diverse and unclear natural language descriptions of addresses X into the standard address space, and return a standard address $\hat{Y}$ that matches X. We define the address parsing task as a mapping: $\mathcal{F}(X) \mapsto \hat{Y}$.

3 Adaptive Weighted Learning-based Address Parsing

We introduce the Adaptive Weighted Learning-based Address Parsing (AWLAP) framework for robust address parsing. After tokenizing natural language descriptions and standard addresses into token IDs $\tilde{X}$, AWLAP employs two key components: First, a pre-trained language model (PLM) captures spatial semantics of input data. To overcome the PLM's limitations in recognizing addresses with missing or conflicting information, we develop an end-to-end **parsing model** with a downstream classification task that maps addresses to their standard forms while fine-tuning the address encoder. Second, we incorporate a **discriminator** that enhances robustness by guiding the parsing model to adapt its training strategy for different address samples. This component significantly improves inference performance by helping the model differentiate between various address patterns. The framework is illustrated as Figure 2.

3.1 Spatial Semantic Learning in Address Parsing

Natural language address descriptions are a kind of textual information. However, due to the limited number of addresses and their unique expressions, training a specialized model solely on address data to achieve natural language parsing capabilities is challenging. In order to better mine the relationship between geographical entities expressed in diverse address descriptions and adapt address hierarchical structure, we use a pre-trained language model (PLM)

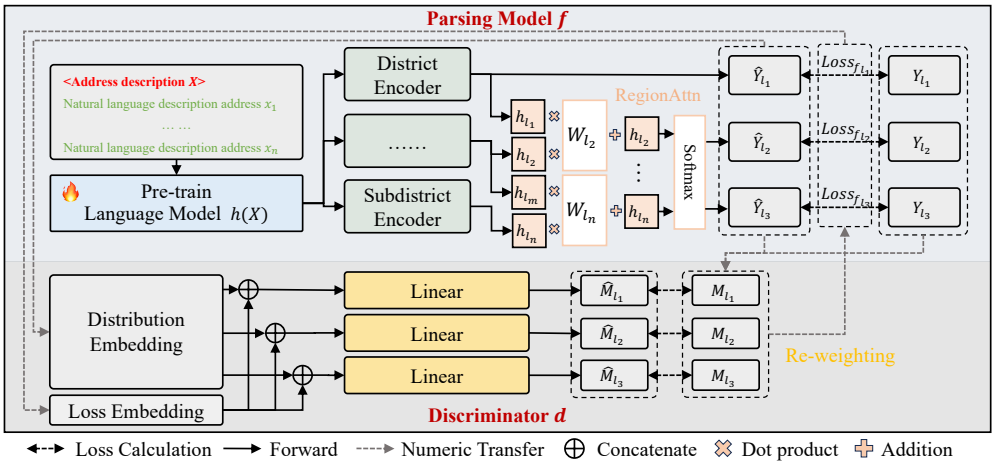


Figure 2: Overall Framework of Adaptive Weighted Learning-based Address Parsing

as an address encoder $h(\cdot)$ for generating address features. And a designed classifier is added downstream to infer the standard address corresponding to the address description by learning the mapping relationship between features and labels, thus obtaining a complete end-to-end address parsing model $f(\cdot)$.

In the AWLAP framework, we utilize the bge-small model [24] as an initialized address encoder. This is a pre-trained language model based on the Transformer architecture with an attention mechanism that allows the model to capture dependencies between different parts of a sequence. In the address parsing scenario, the Transformer allows the address encoder to concentrate on the contextual address information that is most relevant to the current geographic entity by using a self-attention layer. However, due to the specificity of address parsing scenarios, most of the language models lack the knowledge related to geographic entity description during the pre-training process. Therefore, we added a new classification task downstream to support the parsing of natural language descriptions of addresses. The address encoder is then fine-tuned by optimizing the inference results of the classification. This enables the model to gain the ability to mine correlations between geographic entities.

We obtain the address embedding of natural language descriptions through the address encoder, and divide the classification task into different levels, each of which only estimates the standard region of the current level (district, street or community). This allows the classifier to mine more detailed information about the address. Considering that there are mutual constraints among regions at different levels. Each subdistrict have only a parent district, but individuals may mistakenly fill in a wrong parent district when describing their address. To prevent the classifier from generating subordination mismatch problems in classification task, we design a region attention layer in the classifier to constraint the inference of different levels, Eq.1 shows the region attention layer in classifier:

$$z_l = Softmax\left(\frac{h_l(x) \cdot h_{l+1}(x)}{\sqrt{d}}\right) h_l(x) \tag{1}$$

The address parsing flow is shown in the parsing model of Figure 2, including the address encoder and the region attention module. The top district encoder takes in $h(x)$ to generate the hidden embedding of the corresponding level, and then estimates the standard region of the current level based on the hidden embedding. The subdistrict encoders accept both the $h(x)$ and the embedding of the parent district using region attention layer to generate the standard region. This improves the classification efficiency of the model while reducing misclassification due to hierarchical conflicts by using information from the parent district. The classifier's loss function is jointly constructed from the classification results of multiple levels:

$$\mathcal{L}_f = -\sum_{i\in 1,\cdots,n}\sum_{j\in 1,\cdots,l} y_{ij}\log(\hat{y}_{ij}) \tag{2}$$

Notice that, to increase the robustness of the model under incomplete address descriptions or misspellings, we randomly mask the characters (10% of the total characters) in the address descriptions during training, forcing the model to match the correct standard address even when receiving incomplete input.

3.2 Parsing Enhanced based on Discriminator-Weighted

Because of differences in expression habits, the same address may be described in various ways, resulting in varying geographic information. Some descriptions may be detailed and accurate, while others may be fuzzy or contain typos or hierarchical conflicts. Overly concise or erroneous addresses provide less effective information for address parsing and may even be misleading. However, the classifier f has limited ability to discriminate various samples and cannot learn differently for each one. If a suitable weight can be chosen to balance the learning degree of the classifier for different address samples, increasing the learning degree for the samples with large amount of information or difficult to learn, and decreasing the learning degree for the simple samples that can be well predicted, then the learning efficiency and accuracy of the classifier will be greatly improved.

The classification error is the most direct measurement of a classifier's ability to learn from different samples. The classification ability tends to be higher for larger errors and lower for smaller ones.

However, while it provides some insight into the degree of learning, it cannot be quantitatively calculated and cannot be used as a reference for model training feedback. Inspired by semi-supervised learning based on soft labels [22], we introduce an additional discriminator d that utilizes the classification information provided by f to guide the training of f.

The discriminator uses the classification error as input to estimate the correctness of the classification, thus determining how much the current sample has been learned. Additionally, the output distribution of the classifier is also added to the input, which provides the probability of natural language samples belonging to each class of standard addresses. This results in more comprehensive information about the inference of the classifier. During training, positive labels are those correctly classified by the classifier, while the negative are those misclassified. The discriminator is trained to gain the ability to quantify the difficulty of classifying samples from different addresses, shown in Eq. 3:

$$d(f(X), g(Y, f(X))) = P, \quad \mathcal{L}_d = -M \odot Y \log(P) + (1 - M) \odot Y \log(P) \tag{3}$$

We define the classification error function of the classifier f as $g(\cdot)$. M is the mask of correct and incorrect classification, and $\odot$ is the Hadamard product.

During the learning process of the discriminator d, f has appeared in the input of d in the form of a loss function, which means that the classification information of f will affect $\mathcal{L}_d$ and further influence the learning of d. Thus, both d and $\mathcal{L}_d$ become functionals of f, i.e., $d(X, Y, g(f(X)))$, $\mathcal{L}_d(d, g(f(X)))$. Therefore, in optimizing the discriminator, the effect of f on d becomes the basis for optimizing d for f. To explore how the variation of f impacts the information it provides to $\mathcal{L}_d$, we calculated the derivative of $\mathcal{L}_d$:

$$\begin{aligned} \frac{\partial \mathcal{L}_d}{\partial d} \cdot \frac{\partial d}{\partial g} \cdot \frac{\partial g}{\partial f} \cdot \frac{\partial f}{\partial \theta_f} &\overset{(i)}{=} \frac{\partial \mathcal{L}_d}{\partial d} \cdot \frac{\partial d}{\partial g} \cdot \nabla_{\theta_f} loss_A \\ &\overset{(ii)}{=} \frac{\partial d}{\partial g} \cdot \frac{\partial \mathcal{L}_d}{\partial d} \cdot \nabla_{\theta_f} loss_A = 0 \end{aligned} \tag{4}$$

Since we only concentrate on the effect of $\mathcal{L}_d$ on f, we consider a solution of Eq.(4) by letting $\frac{\partial \mathcal{L}_d}{\partial d} \cdot \nabla_{\theta_f} loss_A = 0$. And substituting the BCE loss into $\mathcal{L}_d$, the condition becomes:

$$-\sum_{i=1\cdots n} \sum_{j=1\cdots l} (\frac{1}{p_{ij}}) \cdot (y_{ij} \log(\hat{y}_{ij})) = 0 \tag{5}$$

p_{ij} is the output of $d(X, \hat{Y}, g(\tilde{Y}, \hat{Y}))$ with index ij, and it is also the confidence level given by discriminator d for each level of the parsing result based on the input address description and the classification error. We notice that the form of Eq. (5) is very similar to the loss function of f. And if we consider $\frac{1}{p_{ij}}$ as the reweighting factor of different address samples, then Eq. (5) represents the optimization direction of f given by the discriminator d after accepting the information of f.

Assuming that the discriminator d has been well-trained and can distinguish between difficult and easy samples during training, the confidence levels it generates are representative of the degree to which the current classifier has learned about different samples. If the confidence of a sample in d is low, it suggests that the classifier struggles to extract accurate geographic information and requires further training. Conversely, a high confidence for a sample in d indicates that the classifier can already accurately infer its standard address, allowing for a reduction in the classifier's learning degree for that sample. The classification loss function, weighted by the discriminator d, is as follows:

$$\mathcal{L}_{f_{new}} = -\sum_{i=1\cdots n} \sum_{j=1\cdots l} (1 + \frac{1}{p_{ij}}) \cdot (y_{ij} \log(\hat{y}_{ij})) \tag{6}$$

Above re-weighting factors induce very different behaviors on the optimization of different address sample X. As p_{ij} represents the judgment of the discriminator d on whether $\hat{y_{ij}}$ is corresponding to the x_{ij} or not. For each address sample, the classification errors are align by additional weight of $1/p_{ij}$. Note that if discriminator d judges an address $x_{ij}, m_{x_i y_i} = 1$ to be uncorrect or unreliable (confidence $p_i \to 0$), the classification error of this entry will get significantly boosted. This forces the parsing model f to pay more attention to classify the problematic address sample. For example, the inclusion of noise in the address description, or conflicting descriptions at different address levels, can make the parsing model more difficult to classify correctly. So in such a case, by introducing a reweighting factor in the loss function to make the parsing model pay more attention to the addresses that are difficult to classify, it can make it more robust in the classification of difficult samples.

Table 1: Basic information of the JDL and POI Datasets

	JDL	POI
Total Amount of Addresses	25,966,413	53,513
Average Length of Address (characters)	31.22	21.78
Average Count of Address per Community	8842.39	20.02
Average Count of Address per Neighborhood	3703.66	10.46

4 Experiments

4.1 Experimental Settings

Datasets We evaluate AWLAP's effectiveness using two real-world address datasets: POI addresses and JDL addresses. These datasets represent different common address types in everyday scenarios. JDL addresses, collected from customer delivery information, typically contain more irrelevant details and ambiguous descriptions. In contrast, POI addresses are officially verified and more standardized. Table 1 summarizes the basic statistics of both datasets.

- **Standard address.** Beijing's standard addresses follow a hierarchical structure from district to neighborhood level, encompassing 17 districts, 359 streets, and 7,639 communities. We enhanced this database by manually adding 9,134 neighborhoods with help from local community managers. These addresses follow a consistent format of <city-district-town-community-neighborhood> and serve as reference candidates for address matching.
- **JDL address.** JD Logistics address data combining user-selected geographic information with detailed address fields to create natural language descriptions. The dataset covers 7,446 Beijing neighborhoods and accommodates most address query scenarios.
- **POI address.** Point-of-interest data from Baidu Maps provides address descriptions, filtered to focus on "neighborhood" designations for training and testing. POI addresses typically combine

Table 2: Evaluation results of AWLAP and the baseline methods in JDL and POI datasets

Type	Methods	JDL				POI			
		acc@1	acc@5	acc@10	MAP	acc@1	acc@5	acc@10	MAP
PLM-based	text2vec-base-chinese	45.48%	67.59%	74.70%	55.72%	53.32%	77.70%	82.91%	64.09%
	gte-small-zh	49.35%	72.20%	79.50%	59.84%	59.95%	81.30%	86.28%	69.61%
	bge-small-zh-v1.5	52.63%	74.85%	80.40%	62.71%	64.72%	84.76%	87.82%	73.72%
	bge-m3	65.75%	85.63%	89.82%	74.70%	70.27%	88.74%	91.70%	78.61%
LLM with RAG	Baichuan-13B-Chat	47.90%	-	-	-	62.71%	-	-	-
	Qwen1.5-32B-Chat	55.44%	-	-	-	54.08%	-	-	-
	Llama-3-8B	62.32%	-	-	-	64.17%	-	-	-
Ranking-based	ColBERT	32.15%	53.61%	60.92%	-	35.11%	57.49%	64.66%	-
	MGeo	35.93%	56.66%	60.99%	-	43.57%	72.63%	79.66%	-
Variants	AWLAP-only PLM	81.98%	96.55%	97.90%	88.54%	75.43%	**91.88%**	**94.26%**	82.65%
	AWLAP-self weighting	88.98%	98.96%	99.43%	93.52%	69.84%	85.83%	88.68%	77.12%
	AWLAP-without weighting	88.29%	98.82%	99.30%	94.15%	69.72%	85.58%	88.49%	76.91%
Ours	**AWLAP**	**93.70%**	**99.45%**	**99.70%**	**96.31%**	**77.30%**	90.15%	92.29%	**83.12%**

administrative region information with road network details (e.g., <province-city-district> plus <road name-road number>).

Baselines We selected three different types of models, which are representative and have outstanding performance in the natural language processing field, as baselines in our experiments to evaluate the effectiveness of AWLAP.

- **PLM based methods.** We leverage pre-trained language models to encode addresses and calculate cosine similarity between natural language descriptions and standard addresses. We evaluated four Chinese-optimized models: *text2vec-base-chinese*[26], *gte-small-zh*[16], *bge-small-zh-v1.5*[24] and *bge-m3*[4], ranking standard addresses by correlation and returning the top match.
- **LLM with RAG.** We implement Retrieval-Augmented Generation[13] using bge-m3 to retrieve 50 high-similarity candidate addresses, then employ language models to generate the best match. We tested three models with Chinese language proficiency: *Baichuan-13B*[27], the larger *Qwen-32B*[2], and the recent *Llama-7B*[1].
- **Ranking based methods.** These compute similarity scores between descriptions and candidate addresses through more sophisticated token-level interactions than simple embedding distances. We compare two state-of-the-art text-matching models: *ColBERT*[11] with its efficient late interaction mechanism, and *MGeo*[7], a specialized address-focused model.
- **Variants of AWLAP.** (1) *AWLAP-only PLM* using fine-tuned bge-small-zh-v1 encoders; (2) *AWLAP-without weight*, which eliminates the re-weighting training component; and (3) *AWLAP-self weighting*, which uses classifier-generated probabilities instead of our proposed weighting method.

4.2 Results

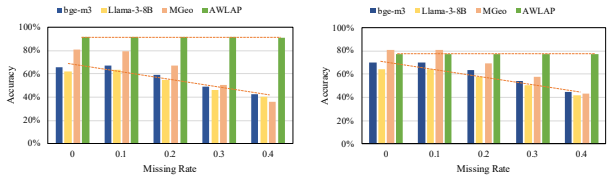
Parsing Result for Different Methods. The AWLAP framework outperforms all baselines in address parsing across both JDL and POI datasets, as shown in Table 2. JDL addresses follow geographic hierarchy, while POI addresses often use road network structures, creating significant distribution differences. JDL's larger training dataset yields higher accuracy, while POI's limited data (only 1% of JDL's volume) results in 15% lower top1 inference accuracy. Despite their strong embedding capabilities on diverse corpora or generative abilities, PLM-based, RAG-based, and ranking-based approaches fail to achieve practical real-world performance. This is because address descriptions differ semantically from standard natural language and contain spatial domain-specific relationships between geographic entities, making accurate parsing difficult without domain knowledge. Even MGeo, specialized for address matching, struggles with ambiguous addresses.

Our ablation study demonstrates the effectiveness of each AWLAP component. The PLM model with address data fine-tuning (AWLAP-only PLM) improves top1 accuracy by 29% compared to unfine-tuned bge-small. Adding a customized classifier on the fine-tuned address encoder further increases accuracy by 8%. With these foundations plus our adaptive re-weighting strategy, AWLAP achieves over 90% parsing accuracy. We compared self-weighting (using classifier output probability) with our proposed method, finding that self-weighting fails to facilitate learning across different address samples because it cannot assess the overall sample distribution or accurately evaluate current sample learning levels.

Robustness of Parsing under Ambiguous Descriptions. To assess AWLAP's robustness, we compared it with baselines across two common ambiguous scenarios: description missing and hierarchical conflict (Table 3). Hierarchical conflict occurs when multiple different communities and streets appear in descriptions or entity attributions mismatch across levels. Description missing refers to addresses lacking community and street information. "Conflict" descriptions introduce additional noise compared to "easy" addresses with complete descriptions, significantly degrading baseline parsing accuracy. In the "missing" scenario, most baselines perform worse than in hierarchical conflict scenarios due to insufficient valid information. Ranking-based methods particularly suffer under these conditions due to their token-level interaction mechanism. AWLAP, however, maintains high accuracy even in complex scenarios by leveraging the pre-trained language model's fuzzy encoding capabilities and implementing discriminator-based reweighting to learn differentially from various samples.

Table 3: Evaluation results for hierarchical conflict and missing scenarios in JDL datasets

Methods	overall	conflict	missing	easy
text2vec-base-chinese	45.48%	38.35%	30.90%	57.17%
gte-small-zh	49.35%	44.76%	39.61%	57.02%
bge-small-zh-v1.5	52.63%	44.59%	27.51%	69.42%
bge-m3	65.75%	57.54%	52.95%	77.55%
Baichuan-13B-Chat	47.90%	39.05%	45.06%	55.68%
Qwen1.5-32B-Chat	55.44%	42.21%	54.66%	65.92%
Llama-3-8B	62.32%	49.11%	59.51%	73.95%
ColBERT	32.15%	33.62%	21.23%	35.54%
MGeo	35.93%	30.81%	32.27%	41.24%
AWLAP-only PLM	81.98%	80.45%	79.83%	84.08%
AWLAP-without weighting	88.29%	87.82%	85.14%	90.99%
AWLAP-self weighting	89.01%	88.23%	86.77%	90.46%
AWLAP	**93.70%**	**92.80%**	**91.94%**	**94.88%**



(a) JDL (b) POI

Figure 3: Parsing accuracy of AWLAP and baselines in JDL and POI datasets under different missing rate

We further evaluated robustness against abbreviations and spelling errors by randomly masking portions of neighborhood descriptions, increasing the masking rate from 0% to 40% (Figure 3). At low masking rates (about 0.1), 1-2 missing words minimally impact deep learning models. However, as masking increases, PLM, LLM, and Ranking models struggle to extract valid information, causing significant accuracy decline. Notably, the ranking model MGeo can match or exceed AWLAP at low missing rates due to its deep interaction mechanism and pre-training on extensive POI address data. However, its accuracy drops dramatically at higher missing rates (>0.3), revealing limited robustness with ambiguous descriptions. AWLAP maintains high accuracy even at high missing rates through random dropout during training and adaptive re-weighting, enabling better knowledge extraction from address descriptions.

Accuracy in different region scale. As Beijing is a very large city with an area of 16,410.54 square kilometers. In order to evaluate the scalability of the proposed method at different scales, we consider the main urban areas (small-scale with 5 district and 8,310 candidate standard addresses), the main urban areas and new districts (medium-scale with 10 district and 12,942 candidate standard addresses), and the main and new urban areas plus the suburban areas (large-scale with 17 district and 17,136 candidate standard addresses) of Beijing as three different geospatial scales to determine the accuracy of the address parsing with different numbers of standard address candidate sets. From Table. 4, it can be seen that

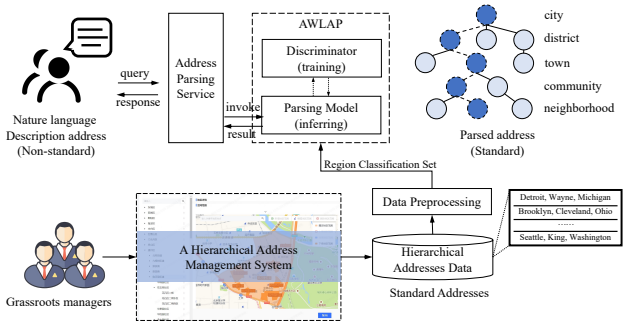


Figure 4: AWLAP Applied in Grassroots Governance.

the number of candidate addresses increases with the increase of city size, and with sufficient training data, the increase of candidate addresses will not confuse the judgment of the model, maintained over 93% accuracy. Meanwhile, the deep learning-based classification model avoids the similarity calculation with the candidate addresses, which also makes the inference efficiency of the model not decrease with the increase of the size of the candidate addresses.

Table 4: Parsing accuracy and average parsing time per address for AWLAP at different area scales

Area scale	districts	standard addresses	accuracy	time
small-scale	5	2,285	93.67%	1.8ms
medium-scale	10	4,505	93.71%	2.2ms
large-scale	17	7,637	93.70%	1.8ms

4.3 Address Parsing System.

In real-world implementations, AWLAP-based address parsing service has been successfully deployed across multiple cities including Beijing, Datong, and Xinyu, significantly enhancing local government address management efficiency. As illustrated in Figure 4, the deployment architecture integrates seamlessly with existing government systems: municipal administrators maintain authoritative address databases through the Hierarchical Address Management System, while AWLAP processes incoming non-standard addresses after data preprocessing (cleaning and format standardization), using the official address repository as its classification reference set. The deployment configuration is remarkably lightweight and efficient, requiring only the parsing model component while omitting the discriminator module in production environments. This streamlined architecture enables high-throughput processing while maintaining minimal server resource requirements, with average parsing times under 10ms per address on standard hardware configurations. The system employs containerized deployment using Docker with Kubernetes orchestration, ensuring scalability during peak usage periods and facilitating seamless updates without service interruption. Local governments leverage this technology to automatically route address-associated resident submissions to appropriate community offices, effectively replacing previously manual assignment workflows. The robust handling of ambiguous

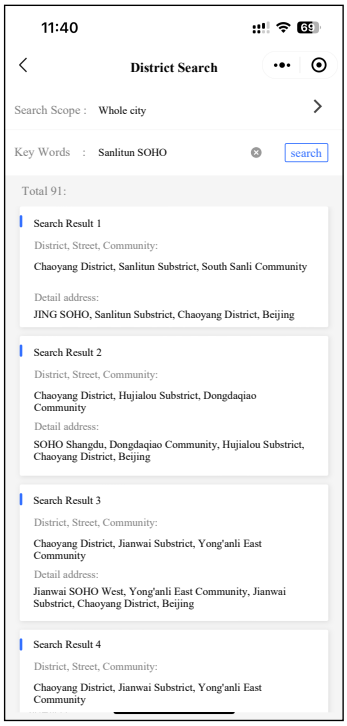


Figure 5: AWLAP in WeChat Applet

addresses has proven particularly valuable in densely populated urban districts where address descriptions often contain colloquial references or incomplete information. System monitoring indicates a 78% reduction in manual address verification workload across deployed municipalities, with address parsing accuracy consistently exceeding 92% in production environments.

Beyond its core governmental applications, we've extended the AWLAP framework to public users through a WeChat mini-program interface, shown in Figure 5. This consumer-facing deployment utilizes edge computing principles to distribute processing load, with the model deployed across regional cloud nodes to minimize latency. The user-friendly interface allows residents to discover standardized addresses by entering descriptive keywords, with the added functionality of geographic scope selection to improve search precision. The system returns multiple candidate addresses ranked by confidence score, through which users can access official street and community information. Load testing confirms the system can handle up to 10,000 concurrent users with negligible performance degradation, making it suitable for citywide public adoption.

5 Related Work

Nature Language Understanding in Geospatial Scenario Address parsing is a critical natural language understanding task in urban scenarios. Within geospatial contexts, effectively capturing implicit spatial relationships is essential, particularly when handling geo-entities like addresses and points of interest (POIs). Pretrained Language Models (PLMs) have revolutionized natural language processing by providing robust representations across various tasks. While models like BERT [6, 12, 25] and GPT [19, 21] excel in general NLP tasks, their application to geospatial data remains nascent. To address the lack of correlation priors in spatial semantic data, researchers have adapted PLMs for geo-entity representation. Li et al. [15] introduced SPABERT, a spatial language model providing general-purpose geo-entity representation based on neighboring entities in geospatial data. [10, 23] construct heterogeneous graphs based on geographical correlations to better capture spatial relationships and enhance representations for POIs and address geo-entities. For relevance models in search ranking, approaches typically calculate similarity scores between geo-entities or perform binary classification to determine matches. Previous studies [8] enhanced geographic representation of address text using contrastive learning to strengthen similarity between proximate addresses. [7, 30] augment ranking models with geographical information for multimodal (textual and spatial) representation learning in query-POI scenarios. Recently, LLM-based methods such as RAG [13] have been applied to numerous NLP tasks. Huang et al. [9] proposed an LLM agent framework for address standardization.

Re-weighting based Robust Learning Natural language descriptions of addresses are often partially noisy, making it difficult for models to mine them for meaningful spatio-temporal semantic information. And different address descriptions have large differences, when training with the same learning intensity, some samples may have overfitted, while some samples still do not learn useful information. Therefore, robust learning methods based on cost-sensitive learning or reweighting are applied to such scenarios. Cost-sensitive reweighting methods assign different weights to samples to adjust their importance. Commonly used methods include reweighting samples inversely proportional to the number of classes. For example, CB [5] re-weights the loss value to be inverse to the effective number of samples per class. FL [17] down-weights the loss values assigned to the majority classes and the well-classified examples. GHM [14] re-weights the loss values of samples based on their gradients per iteration. And to avoid overfitting in important samples, Park et al. [20] propose a balancing training method to address the problems of unbalanced data learning. They derive a new loss by adding a class-wise reweighting term to balance the training phase, which mitigates the influence of samples that cause an overfitted decision boundary.

6 Conclusion

In this paper, we propose the AWLAP framework to solve the address parsing problem, which maps natural language descriptions of addresses to standard address sets. We construct a downstream multi-classification task based on a pre-trained language model and design a new multi-level constrained classifier that performs mapping inference from natural language to standard addresses by learning correlations between geographic entities under geographic hierarchy constraints. To adapt the model to address data with varying degrees of noise and to extract as much information as possible, we added an additional discriminator that is used to dynamically generate optimized weights for the classifier during training. We evaluated the framework on large address datasets of JDL and POI. Extensive experiments against SOTA baselines demonstrate the effectiveness and robustness of the approach.

References

[1] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[3] Abhranil Chatterjee, Janit Anjaria, Sourav Roy, Arnab Ganguli, and Krishanu Seal. 2016. SAGEL: Smart Address Geocoding Engine for Supply-Chain Logistics. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, Burlingame California, 1–10. https://doi.org/10.1145/2996913.2996917

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 9268–9277. https://doi.org/10.1109/CVPR.2019.00949

[6] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[7] Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. 2023. Mgeo: Multi-modal geographic language model pre-training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 185–194.

[8] Govind and Sohoney Saurabh. 2022. Learning Geolocations for Cold-Start and Hard-to-Resolve Addresses via Deep Metric Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Abu Dhabi, UAE, 322–331. https://doi.org/10.18653/v1/2022.emnlp-industry.33

[9] Chenghua Huang, Shisong Chen, Zhixu Li, Jianfeng Qu, Yanghua Xiao, Jiaxin Liu, and Zhigang Chen. 2024. GeoAgent: To Empower LLMs using Geospatial Tools for Address Standardization. In *Findings of the Association for Computational Linguistics ACL 2024*. 6048–6063.

[10] 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*. ACM, Washington DC USA, 3029–3039. https://doi.org/10.1145/3534678.3539021

[11] 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*. 39–48.

[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/V1/2020.ACL-MAIN.703

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[14] Buyu Li, Yu Liu, and Xiaogang Wang. 2019. Gradient Harmonized Single-Stage Detector. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 8577–8584. https://doi.org/10.1609/AAAI.V33I01.33018577

[15] Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. 2022. SpaBERT: A Pre-trained Language Model from Geographic Data for Geo-Entity Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 2757–2769. https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.200

[16] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281* (2023).

[17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2 (2020), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

[18] Qian Liu, Heyan Huang, Junyu Xuan, Guangquan Zhang, Yang Gao, and Jie Lu. 2020. A fuzzy word similarity measure for selecting top-k similar words in query expansion. *IEEE Transactions on Fuzzy Systems* 29, 8 (2020), 2132–2144.

[19] Niklas Muennighoff. 2022. SGPT: GPT Sentence Embeddings for Semantic Search. *CoRR* abs/2202.08904 (2022). arXiv:2202.08904 https://arxiv.org/abs/2202.08904

[20] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. 2021. Influence-Balanced Loss for Imbalanced Visual Classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 715–724. https://doi.org/10.1109/ICCV48922.2021.00077

[21] Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2023. Bidirectional Language Models Are Also Few-shot Learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=wCFB37bzud4

[22] Huiling Qin, Xianyuan Zhan, Yuanxun Li, and Yu Zheng. 2023. FlexSSL: A Generic and Efficient Framework for Semi-Supervised Learning. *arXiv preprint arXiv:2312.16892* (2023).

[23] Lixia Wu, Jianlin Liu, Junhong Lou, Minhui Deng, Jianbin Zheng, Haomin Wen, Chao Song, and Shu He. 2024. G2PTL: A Geography-Graph Pre-trained Model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 4991–4999. https://doi.org/10.1145/3627673.3680023

[24] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597* (2023).

[25] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. *CoRR* abs/2309.07597 (2023). https://doi.org/10.48550/ARXIV.2309.07597 arXiv:2309.07597

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

[27] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305* (2023).

[28] Zixuan Yuan, Hao Liu, Junming Liu, Yanchi Liu, Yang Yang, Renjun Hu, and Hui Xiong. 2021. Incremental spatio-temporal graph learning for online query-poi matching. In *Proceedings of the Web Conference 2021*. 1586–1597.

[29] 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*. 629–638.

[30] 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*. ACM, Virtual Event China, 629–638. https://doi.org/10.1145/3397271.3401159

7 GenAI Usage Disclosure

Generative AI tools were used solely for editing and improving the quality of existing text, including grammar correction, clarity enhancement, and stylistic improvements, similar to standard writing assistance tools.