

Article

# A Semantic-Spatial Aware Data Conflation Approach for Place Knowledge Graphs

Lianlian He <sup>1,\*</sup>, Hao Li <sup>2</sup> and Rui Zhang <sup>2</sup>

<sup>1</sup> School of Mathematics and Statistics, Hubei University of Education, No. 129 Second Gaoxin Road, East Lake Hi-Tech Zone, Wuhan 430205, China

<sup>2</sup> School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China; leehomm@whu.edu.cn (H.L.); rui.zhang@whu.edu.cn (R.Z.)

\* Correspondence: helianlian@hue.edu.cn

**Abstract:** Recent advances in knowledge graphs show great promise to link various data together to provide a semantic network. Place is an important part in the big picture of the knowledge graph since it serves as a powerful glue to link any data to its georeference. A key technical challenge in constructing knowledge graphs with location nodes as geographical references is the matching of place entities. Traditional methods typically rely on rule-based matching or machine-learning techniques to determine if two place names refer to the same location. However, these approaches are often limited in the feature selection of places for matching criteria, resulting in imbalanced consideration of spatial and semantic features. Deep feature-based methods such as deep learning methods show great promise for improved place data conflation. This paper introduces a Semantic-Spatial Aware Representation Learning Model (SSARLM) for Place Matching. SSARLM liberates the tedious manual feature extraction step inherent in traditional methods, enabling an end-to-end place entity matching pipeline. Furthermore, we introduce an embedding fusion module designed for the unified encoding of semantic and spatial information. In the experiment, we evaluate the approach to named places from Guangzhou and Shanghai cities in GeoNames, OpenStreetMap (OSM), and Baidu Map. The SSARLM is compared with several classical and commonly used binary classification machine learning models, and the state-of-the-art large language model, GPT-4. The results demonstrate the benefit of pre-trained models in data conflation of named places.

**Keywords:** knowledge graph; place entity matching; location-based service; place data; conflation



**Citation:** He, L.; Li, H.; Zhang, R. A Semantic-Spatial Aware Data Conflation Approach for Place Knowledge Graphs. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 106. <https://doi.org/10.3390/ijgi13040106>

Academic Editors: Wolfgang Kainz, Christos Chalkias, Marinos Kavouras, Margarita Kokla and Mara Nikolaidou

Received: 11 December 2023

Revised: 9 March 2024

Accepted: 19 March 2024

Published: 22 March 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

People often perceive the world using named places. Thus, the world of named places can be a kind of georeferencing knowledge. As science and technology continue to penetrate all levels of society, big data are continuously generated from various sources, such as sensor systems data, user-generated content in social networks, socio-economic data, and hybrid data sources, including linked data and synthetic data [1]. Most of these data involve named places, which effectively stimulates geospatial knowledge sharing among various sectors in cities [2]. From the ontological perspective, place entities can be formalized using ontological approaches. The Semantic Web and its best practice, Linked Data, help to publish place entities on the Web using ontologies and Resource Description Framework [3]. It is feasible to organize knowledge and details about place entities according to Linked Data principles and create a place knowledge graph on the Web [4]. The above ideas can result in a paradigm shift for Geographic Information Science [5–7]. From distributed databases accessed through Web services to knowledge represented as graphs, place nodes serve as a powerful “glue” to link any other data to its georeference, thus enabling integration of information across domains.

In the construction of multi-source data knowledge graphs that utilize location nodes as geographic references, place entity matching emerges as a pivotal technology [7–10].

Models for matching place entities can be categorized into two primary types, based on the nature of the data: textual matching and spatial matching. Textual matching processes largely depend on measuring similarities in the text-based attributes of entities, such as the congruence of place names and the likeness of categories, to determine if multiple source place entities refer to the same real-world location [11–13]. Conversely, spatial matching utilizes geometric measures of spatial coordinates, including spatial distances, as significant indicators of match suitability [14–16]. For the task of matching place entities, traditional rule-based methods necessitate the selection of matching factors for the computation of similarities, setting thresholds for each to ensure proper alignment. The most prevalent technique for amalgamating these similarities is the weighted sum approach within mathematical models, which assigns scores to each matching factor according to its relative weight. To diminish the level of manual intervention and circumvent the need for empirical weight configurations, recent advancements have been made in machine learning-based matching methods. Studies such as those by McKenzie et al. [11] and Santos et al. [17] have showcased models trained on manually annotated datasets, which serve as a substitute for manual parameter settings.

However, several challenging issues are hindering the advancement of geographical entity matching: (1) Rule-based methods are hampered by manual experiential interference and the tedious nature of feature selection and calculation, preventing the matching models from achieving uniformity across diverse, heterogeneous geographic data sources. (2) Existing machine learning approaches to geographical entity alignment generally depend on character distance similarity of toponyms, lacking the capability to capture deep semantic features, and thus limiting the further enhancement of place entities matching performance. (3) When multi-source place data lack attribute fields, geographic entity matching methods grounded in feature engineering may turn out to be ineffective. Therefore, the demands on matching models in place entity matching tasks are considerably higher, necessitating the ability of the models to rapidly, accessibly, and accurately adapt and align data from disparate sources for aggregation and unification [18].

With the advent of pre-trained models, new approaches to address these requirements and challenges have been introduced. Pre-trained models, exemplified by Bidirectional Encoder Representations from Transformers (BERT) [8] and Generative Pre-trained Transformer (GPT) [19], acquire a broad understanding of language through unsupervised or semi-supervised learning on extensive text corpora. These models, particularly leveraging the Transformer architecture, effectively capture the bidirectional contextual information of language. Extensive research has validated their exemplary performance across various natural language processing downstream tasks. However, the efficacy of these models in addressing the domain-specific needs of place entities matching remains unverified. Place entity matching differs significantly from general domain entity matching, primarily due to its emphasis on spatial characteristics [20]. A critical area of investigation is integrating spatial feature spaces with textual feature spaces within the model. This integration involves not only understanding the linguistic aspects, but also accurately interpreting and aligning the spatial dimensions of place entities. This would entail enhancing pre-trained models with spatial awareness and tailoring them to better accommodate the unique requirements of geographical data representation and alignment [21].

In this paper, we propose a semantic-spatial aware representation learning model for place matching, which is an end-to-end place entity matching approach based on the large-scale pre-trained model. This advanced approach transcends the cumbersome manual feature extraction steps, which are a staple in traditional machine learning models and, instead, implements a comprehensive place entity matching pipeline system. This system is adept at supporting swift and dynamic updates in geospatial data-intensive tasks.

Our investigation particularly scrutinizes the multifaceted challenges that emerge when fusing diverse datasets to construct a robust knowledge graph. These challenges include, but are not limited to, linguistic variations, orthographic disparities, geometric inconsistencies, temporal discrepancies, and categorical ambiguities. We meticulously

demonstrate the refinements and advancements that are brought about by incorporating rule-based models, machine learning models, as well as the more recent pre-trained large models and extensive language models when tackling these complexities.

The experimental data for this study is sourced from the authoritative Baidu Maps, the user-generated OSM, and the digital gazetteer GeoNames. Utilizing the methodology outlined herein, we publish our place entity linkage as an open data source for the place knowledge graph (PlaceKG).

Our contributions are threefold:

- We present a Semantic-Spatial Aware Representation Learning Model for Place Matching and Fusion based on a pre-trained large model that achieves a unified mapping of location feature spaces and textual feature spaces.
- We evaluate the capabilities of different types of models for the task of place matching, and present a granular showcase of the potential improvements offered by rule models, machine learning models, pre-trained large models, and large language models in response to nine types of case challenges.
- We construct the PlaceKG and validate its utility in the realm of location querying and Location-Based Services, furthering the field of Geographic Information Science.

## 2. Related Work

The place knowledge graph in this paper is a semantic format based on linked data principles, which is constructed by conflating and structuring existing named place datasets. This knowledge graph is oriented to the place field. It helps to solve many geospatial technology challenges, such as named entity recognition (NER) [22], toponym disambiguation [23], and POI recommendation [24]. For example, the NER task can recognize and predict the geographical coordinates of named place entities in text documents of web pages, blogs, encyclopedia articles, news stories, tweets, and travel notes, and is called geographic resolution or geographic coding. This work can connect the unstructured text with GIS structured entities [25,26]. Researchers claim that obtaining extensive gazetteer information is central for NER [27]. A place knowledge graph that fuses various named places on the network can help address this concern.

At the conceptual level, POI usually has the following attributes: name, current location, category, and identifier. As the essential requirement of spatial data infrastructure (SDI), POI is characterized by type and often uses names rather than locations to identify a place [28]. At the application level, POI can be used as a reference point in requesting location-based services, such as the destination of path navigation. The research on place data conflation can be traced back to the early works of digital map conflation and digital gazetteer conflation [29,30]. In the GIS context, POI data is usually the object of conflation. The term “conflation” describes integrating data from heterogeneous data sources, combining geographical information of different scales and precisions, and transferring or adding attributes from one dataset to another [30]. POI conflation aims to determine whether POIs from different data sources represent the same place in a physical world, resolve the ambiguity of attribute values such as names and geographical locations, and integrate matched POIs. It often involves the following steps [12,13,31]:

1. Pre-processing: unifying POI datasets into the same data structure and spatial coordinate system and mapping POI categories or types into a common taxonomy.
2. Candidate selection: using a set of conditions to select candidates from POI datasets.
3. Similarity measure: computing the attribute similarity of POI candidates, such as spatial similarity, name similarity, and type similarity.
4. Matching evaluation: aggregating different similarity measures to obtain an overall value that can rate the matching relevance and evaluate whether the POI candidates are matched.

5. Property conflation: conflating the property values from matched POI candidates. The properties can be either overlapped or complementary. In the former case, merging overlapping information often requires conflict-handling strategies. In the latter case, the final POI entity usually has various properties from different candidates.

Among these steps, the matching stage (Steps 3 and 4) is the most important in data conflation, and receives the most attention in existing efforts. For Step 3, i.e., similarity measures, various work has been conducted on spatial similarity, name similarity, as well as semantic similarity [11–13]. In Step 4, i.e., matching evaluation, some methods have been proposed, including the regression-based weighted model [11], the entropy-weighted model [12], the graph-based matching approach [13], and the Random Forest Classifier-based matching approach [17].

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

### 3. Data Sources for Named Places

Named place data are available through either authorities or geographical vendors. The former includes data provided by toponymic agencies or gazetteer services, and the latter includes POI data provided by crowdsourcing contributions such as OSM and WikiMapia. In this paper, we select one gazetteer, [GeoNames.org](https://www.geonames.org/) (accessed on 21 November 2023), which is one of the most famous accessible geographical databases in the world. For crowdsourcing data sources, we select POIs from OSM, one of the most representative sources of Volunteered Geographic Information (VGI). Furthermore, on the local scale, we select the POI data from one of the most significant Location Based Service (LBS) providers in China, Baidu Map. The empirical study using these datasets helps address the challenges in conflation caused by variability in language, spelling, historical changes, and feature types. Table 1 provides a comparison of the three data sources. Detailed descriptions are given as follows.

**Table 1.** A comparison of the three data sources.

	Feature Type	Fields	Alternative Names	Geometry	Language	Change	History
GeoNames	9 types/680 subtypes	administrative information	Yes	lat/long coordinates (WGS84)	multilingual	modification date	Yes
OpenStreet Map	24 types/free subtypes	creation and tag information	Yes	nodes, ways, and relations (WGS84)	multilingual	version and changeset	Yes
Baidu Map	23 types/153 subtypes	city and heat	Not Applicable (N/A)	lat/long coordinates (BD09)	Chinese	N/A	N/A

#### 3.1. Data Sources

##### 3.1.1. GeoNames

GeoNames is a free geographical database for named geographical features [32]. The dataset currently includes up to 13 million places, and the places are divided into nine categories and 680 subcategories. The fields in the dataset include some basic administrative information, such as administrative divisions, populations, and country codes. Some place entries in this dataset have aliases, such as names in different languages, short names, historical names, and colloquial or slang names. As for historical names, the start and end time stamps (valid dates) should be considered. All position coordinates are in World Geodetic System 1984 (WGS84).



### 3.1.2. OSM

OSM is a free online map built through crowdsourcing VGI [33]. The geometries of geographic features are represented by essential data elements, i.e., nodes, ways, and relations. A node is a specific point on Earth by its latitude and longitude using the WGS84 standard. A way is an ordered list of nodes. A relation records relationships between two or more data elements, such as a polygon relation. Each data element has some common attributes related to creation information, such as users modifying the element (user/uid), the time of the last modification (timestamp), the version number of the edit (version), and the changeset number for a group of changes (changeset). Feature attributes are represented using tags attached to the data elements. A tag is a key–value pair with free-format texts. The communities can agree on certain key and value combinations for the most commonly used tags as consensus-based informal standards, such as a classification tag “highway = footway”. Currently, 24 types of primary point features are defined using well-accepted tags [33]. The primary features used as keys in tags can act as types, and their values in tags can be regarded as subtypes. This paper selects 1511 frequently used subtypes. Although tags can be invented and used as needed, tags used by at least one wiki page can usually be used as types/subtypes. It is noted that historical changes are also tagged as significant features of a particular type. OSM also allows for additional properties to be defined using tags, such as “name”, “name:<lg>” for names in different languages, “alt\_name” for alternative names, and “old\_name” for historical names. The free tagging system makes OSM flexible, but hard to work with. Usually, these tags are used to serve as de facto standards.

### 3.1.3. Baidu Map

LBS services have been widely used in China. Baidu Inc., a Chinese-language internet search company, provides LBS services [34] and is rated as one of the top 10 LBS providers in the world. Baidu offers POI data in China in the Chinese language. Assigned by the governmental agency in China, most domestic online map services use the “GCJ02” coordinate system, and Baidu uses the encrypted coordinate system “BD09” converted from “GCJ02” to provide Chinese POI data, which is unique to Baidu. Currently, Baidu Chinese POI data provides 23 categories and 153 subtypes. Each piece of data includes the city where the POI is located, the popularity of searching (i.e., heat), and some optional detailed attribute information.

## 3.2. Challenges

When researching the three datasets, several challenging issues need to be faced. Taking named places entities in China as examples, our paper lists the issues that need to be solved during the named place matching and conflating process in Table 2. The differences are highlighted in the bold style, and the Chinese–English translation is provided in Appendix A.

Issue I1 concerns the multilingual problem of place names in different data sources. For example, the place Guangzhou East Railway Station is expressed as “Guangzhou Dong” in English in GeoNames and presented as “广州东站” in Chinese in both OSM and Baidu Map. Issues I2–I6 focus on the spelling changes of place names in different data sources, including case sensitivity (“Sultan Turkish Restaurant” vs. “SULTAN TURKISH RESTAURANT”), misspelling (“珞瑜路” vs. “珞喻路”), word order (“地铁浔峰岗站” vs. “浔峰岗地铁站”), abbreviation (“广州火车站” vs. “广州站”), and synonym (“从化区办证中心” vs. “从化区政务服务中心”). I7 and I8 are location variability issues in different data sources, such as the location values corresponding to different coordinate systems and how coordinate precisions will change. I9 shows the type variability that the same named places from different data sources can have different type names in different categories.

Table 2. Challenging issues of the conflation.






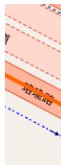








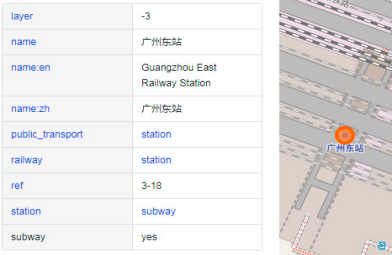

Challenging Issues	Examples																				
	GeoNames	OSM	Baidu Map																		
I1: names expressed in different languages	<div><p>Guangzhou Dong</p></div>	<div><table><tr><td>layer</td><td>-3</td></tr><tr><td>name</td><td>广州东站</td></tr><tr><td>name.en</td><td>Guangzhou East Railway Station</td></tr><tr><td>name.zh</td><td>广州东站</td></tr><tr><td>public_transport</td><td>station</td></tr><tr><td>railway</td><td>station</td></tr><tr><td>ref</td><td>3-18</td></tr><tr><td>station</td><td>subway</td></tr><tr><td>subway</td><td>yes</td></tr></table><p>广州东站</p></div>	layer	-3	name	广州东站	name.en	Guangzhou East Railway Station	name.zh	广州东站	public_transport	station	railway	station	ref	3-18	station	subway	subway	yes	<div><p>广州东站</p></div>
layer	-3																				
name	广州东站																				
name.en	Guangzhou East Railway Station																				
name.zh	广州东站																				
public_transport	station																				
railway	station																				
ref	3-18																				
station	subway																				
subway	yes																				
I2: names with different cases	<div>N/A</div>	<div><table><tr><td>amenity</td><td>restaurant</td></tr><tr><td>name</td><td>Sultan Turkish Restaurant</td></tr></table><p>Sultan Turkish Restaurant</p></div>	amenity	restaurant	name	Sultan Turkish Restaurant	<div><p>SULTAN TURKISH RESTAURANT</p></div>														
amenity	restaurant																				
name	Sultan Turkish Restaurant																				
I3: names with spelling errors	<div>N/A</div>	<div><table><tr><td>highway</td><td>trunk</td></tr><tr><td>lanes</td><td>3</td></tr><tr><td>name</td><td>珞瑜路</td></tr><tr><td>oneway</td><td>yes</td></tr><tr><td>surface</td><td>concrete</td></tr></table><p>珞瑜路</p></div>	highway	trunk	lanes	3	name	珞瑜路	oneway	yes	surface	concrete	<div><p>珞瑜路</p></div>								
highway	trunk																				
lanes	3																				
name	珞瑜路																				
oneway	yes																				
surface	concrete																				
I4: names with different word orders	<div>N/A</div>	<div><table><tr><td>bus</td><td>yes</td></tr><tr><td>highway</td><td>bus_stop</td></tr><tr><td>name</td><td>地铁浔峰岗站</td></tr><tr><td>name.en</td><td>Xunfenggang Metro Station</td></tr><tr><td>name.zh</td><td>地铁浔峰岗站</td></tr><tr><td>public_transport</td><td>platform</td></tr></table><p>地铁浔峰岗站</p></div>	bus	yes	highway	bus_stop	name	地铁浔峰岗站	name.en	Xunfenggang Metro Station	name.zh	地铁浔峰岗站	public_transport	platform	<div><p>浔峰岗地铁站</p></div>						
bus	yes																				
highway	bus_stop																				
name	地铁浔峰岗站																				
name.en	Xunfenggang Metro Station																				
name.zh	地铁浔峰岗站																				
public_transport	platform																				
I5: names with abbreviations	<div><p>广州火车站</p></div>	<div><table><tr><td>name</td><td>广州站</td></tr><tr><td>name.en</td><td>Guangzhou Railway Station</td></tr><tr><td>name.zh</td><td>广州站</td></tr><tr><td>operator</td><td>中国铁路广州局集团有限公司</td></tr><tr><td>public_transport</td><td>station</td></tr><tr><td>railway</td><td>station</td></tr></table><p>广州站</p></div>	name	广州站	name.en	Guangzhou Railway Station	name.zh	广州站	operator	中国铁路广州局集团有限公司	public_transport	station	railway	station	<div><p>广州火车站</p></div>						
name	广州站																				
name.en	Guangzhou Railway Station																				
name.zh	广州站																				
operator	中国铁路广州局集团有限公司																				
public_transport	station																				
railway	station																				
I6: names with synonyms	<div>N/A</div>	<div><table><tr><td>name</td><td>从化区办证中心</td></tr><tr><td>office</td><td>government</td></tr></table><p>从化区办证中心</p></div>	name	从化区办证中心	office	government	<div><p>从化区政务服务中心</p></div>														
name	从化区办证中心																				
office	government																				

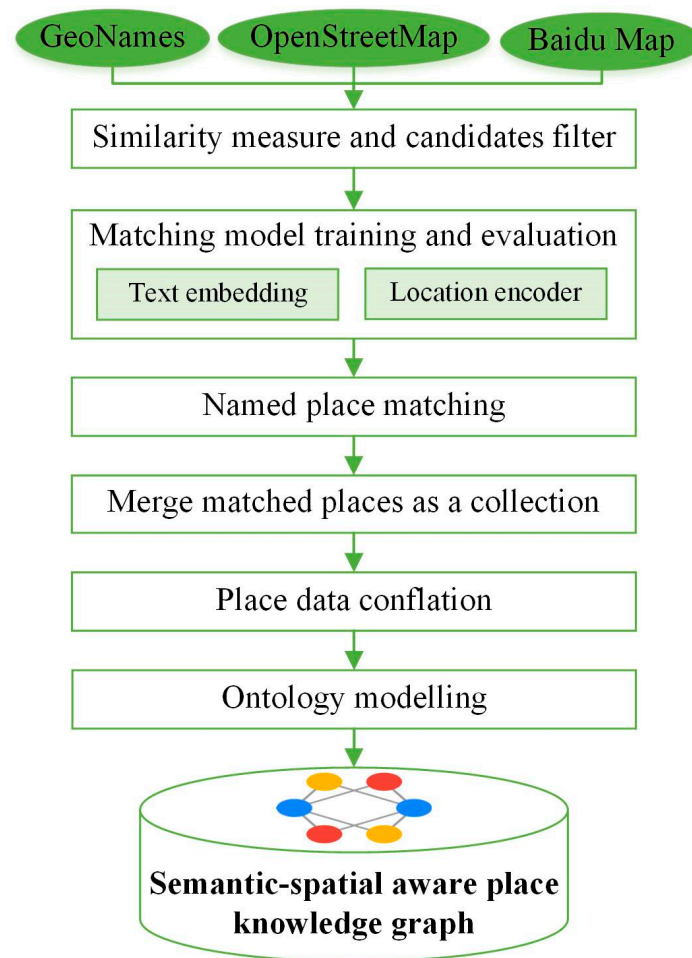
Table 2. Cont.

Challenging Issues	Examples		
	GeoNames	OSM	Baidu Map
I7: named places using different coordinate systems	WGS84	WGS84	BD09
I8: the same named places with different coordinates	 23.15344,113.31977	 23.15354,113.31906	 23.15582,113.33106
I9: the same named places using different type categories and type names.	 S: RSTN	 public_transport: station; railway: station	 交通设施: 火车站 (Transportation facilities: train station)

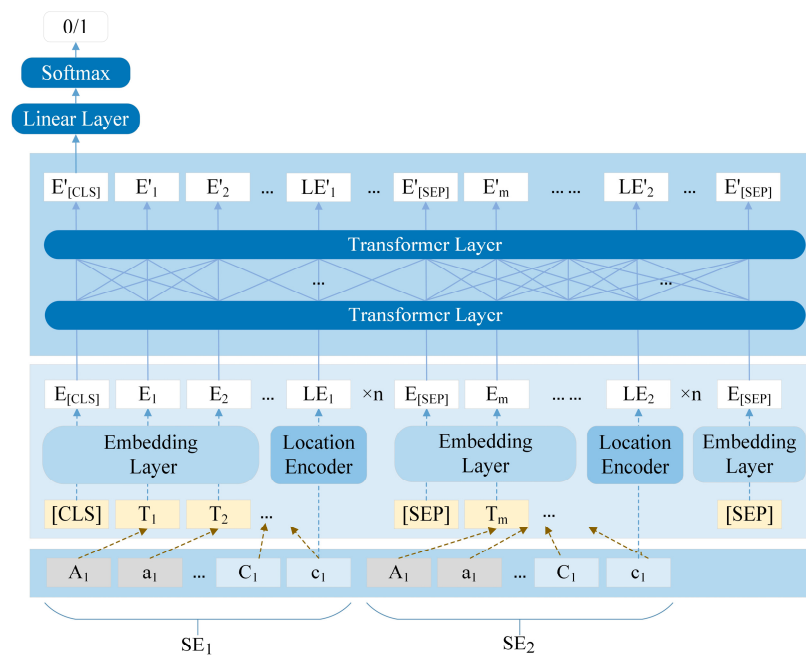
#### 4. Approach

Based on the preceding context, it is evident that the matching process necessitates addressing the disparities in names, locations, and types of named places in GeoNames, OSM, and Baidu Map. Our experiment primarily focuses on resolving such issues flexibly and conveniently.

The construction process of PlaceKG, as illustrated in Figure 1, begins with data cleansing to obtain preliminary data from these three sources. Given that there will be  $n \times m$  potential matches between two data sources containing  $n$  and  $m$  named locations, respectively, the sheer volume of data can make the precise matching process exceedingly laborious. To mitigate this, we initially filter these pairs based on threshold values to reduce the number of mismatched pairs. Then, this section introduces the SSARLM for data merging of named places from diverse data sources, to simplify and enhance the performance of similarity measurement. We have the spatial awareness of a large-scale pre-trained model with a location encoding strategy and utilize the candidate set data to fine-tune the model. The overall framework of the SSARLM model is depicted in Figure 2.



**Figure 1.** Process for semantic-spatial aware place knowledge graph construction.



**Figure 2.** SSARLM Model Architecture Diagram.

#### 4.1. Place Entity Serialization

In general, natural language tasks, pre-trained models accept token sequences (i.e., text) as input. Similarly, for named place entity data, serialization processing is required before feeding into the matching model, which facilitates subsequent encoding by the model. For multi-source heterogeneous geographic entity data, effective serialization is crucial for the model to optimally assimilate and process the data information. This step is fundamental in ensuring that the model accurately interprets and utilizes the geographic information embedded within these diverse data streams.

For data pairs composed of two places from two different data sources, such as their names, types, and location attributes, are transformed into token sequences. This transformation optimizes the model's encoding process, thereby enhancing its capacity to ascertain if the candidate pair signifies identical geographical features. In serializing the attributes and values of a place entity from a specific data source, a *[COL]* marker is used to denote each attribute and its value for an individual place. If *A*, *B*, *C* represent the attributes of name, type, and location, respectively, and their lowercase forms denote the corresponding attribute values, then a geographic entity can be represented as  $SE = [COL]Aa[COL]Bb[COL]Cc$ . This structured representation is instrumental in the model's processing and analysis of the geographical data.

For a pair of place entities constituting a data pair, upon joint input into the model, *[CLS]* and *[SEP]* tokens are generated for separation and subsequent classification computations. It is key to note that *[CLS]* and *[SEP]* are Bert model-specific notations for the start and separation of entities. The *[SEP]* token serves as a delimiter marking the boundary between the two entities in the data pair, with the beginning and end of the data pair marked by *[CLS]* and *[SEP]*, respectively. Once processed by the model, the vector generated at the *[CLS]* position is fed into a fully connected layer for classification calculation. Therefore, the serialization result of a data pair can be represented in the following format:

$$S = [CLS]SE_1[SEP]SE_2[SEP] \quad (1)$$

This structure ensures that the model effectively discerns the start, separation, and end of each data pair, facilitating accurate classification and analysis of the geographic entities.

#### 4.2. Fine-Tuning Pre-Trained Language Models

This study proposes an end-to-end approach for place entity matching models using pre-trained models, wherein the matching model is trained by fine-tuning these existing pre-trained architectures. Typical pre-trained models like BERT and GPT demonstrate robust performance across various Natural Language Processing (NLP) tasks. These models are usually composed of deep neural networks with multiple transformer layers, and are trained using unsupervised techniques on extensive text corpora, such as Wikipedia articles. During this pretraining phase, the models enhance their ability to understand sentence semantics by autonomously learning to predict missing tokens and subsequent sentences. This capability stems from the Transformer architecture's ability to generate token embeddings from all tokens in an input sequence, thereby producing highly contextualized embeddings that encapsulate both the semantic and contextual understanding of words. Consequently, these embeddings adeptly capture polysemy, recognizing that a word can have different meanings in different phrases. For example, place names such as "Wuhan Station" and "Wuhan Railway Station" may still acquire similar word embeddings, despite the former lacking the key phrase 'railway'. This similarity arises from training on extensive corpora, as the embeddings in pre-trained models are based on the semantic theories they have assimilated.



In this experiment, we utilized the case-sensitive pre-trained model DistilBERT provided by Hugging Face as the foundational model. The BERT pre-trained model is fundamentally developed through self-supervised training on a large-scale corpus, essentially constructed using a stacked architecture of multiple transformer layers. DistilBERT, as a distilled version of the BERT model, employs the same training corpus. However, it boasts a parameter size that is only 60% of that of BERT, while retaining 97% of its language understanding capability. Additionally, the model speed is enhanced by 60%. This efficient architecture of DistilBERT ensures a balance between computational resource requirements and the retention of substantial language processing proficiency, making it a suitable choice for our experiment's objectives.

#### 4.3. Semantic-Spatial Aware Representation Learning Model

It is well-known that locations are typically represented by latitude and longitude coordinate values. In multi-source data, issues such as coordinate drift and inconsistencies in decimal places among different data sources are common. When pre-trained models are used to directly encode these positional coordinates, a non-regular expression method based on an arithmetic foundation is adopted. This approach often fails to adequately address the aforementioned coordinate issues, consequently hindering its ability to effectively represent spatial location similarity. This limitation underscores the necessity for more sophisticated methodologies or preprocessing steps to ensure models accurately capture and reflect spatial relationships and similarities, particularly when dealing with diverse and inconsistent geographical data sources.

In addition, the use of location similarity calculation formulas presents challenges in integrating their computed outcomes with pre-trained models. This arises from the fact that similarity calculations require the separate extraction of coordinate values to obtain corresponding results, while pre-trained models typically process the entire data pair in a unified manner. Moreover, approaches based on location similarity calculations often neglect the original location information, relying solely on the outcomes derived from empirical formulas. This can lead to an over-reliance on empirical methods, potentially resulting in inaccuracies. Even small location deviations can inadvertently lead to the incorrect exclusion of candidate pairs. This highlights the need for a more balanced approach that integrates empirical calculations while retaining inherent location data.

Therefore, this study introduces an embedding fusion module designed for the unified encoding of semantic and spatial embeddings. In the previous work by Mai et al. [35], a distributed location encoding approach is constructed to generate spatial embeddings. This approach employs methods such as unit vector inner product and periodic functions to transform two-dimensional positional coordinates into dense vectors of the same dimensionality as the pre-trained model. These vectors, when combined with other vectors encoded by the training model, are designed to convey data information more effectively, thereby enabling the training of a more precise overall similarity measure. Fundamentally, this approach can overcome the limitations of pre-trained models in location encoding, avoiding excessive reliance on empirical formulas, and thus enhancing the accuracy of place entities matching tasks. The specific encoding method is as follows:

$$LE_{s,j}(\mathbf{x}) = \left[ \cos\left(\frac{\langle \mathbf{x}, \mathbf{a}_j \rangle}{\lambda_{\min} \cdot (\lambda_{\max} / \lambda_{\min})^{s/5}}\right); \sin\left(\frac{\langle \mathbf{x}, \mathbf{a}_j \rangle}{\lambda_{\min} \cdot (\lambda_{\max} / \lambda_{\min})^{s/5}}\right) \right] \quad (2)$$

$$LE(\mathbf{x}) = [LE_0^{(t)}(\mathbf{x}); \dots; LE_s^{(t)}(\mathbf{x}); \dots; LE_{S-1}^{(t)}(\mathbf{x})] \forall s = 0, 1, 2, \dots, S-1 \quad (3)$$

For a positional coordinate  $\mathbf{x} = (x, y)$ , the process begins by assigning a specified number of unit vectors  $\mathbf{a}$  and then performing dot product calculations  $\langle \mathbf{x}, \mathbf{a} \rangle$ , resulting in a new spatial representation. Subsequently, each dot product value undergoes scaling to a specified dimension. This study adheres to the six-dimensional choice as utilized in the original work, leading to a periodic expansion through different frequency sine and cosine

functions. As illustrated in Equation (2), where  $s = 0, 1, \dots, 5$  represents the scale coefficients and  $\lambda$  is the scale value, this method effectively transforms the spatial coordinates into a more nuanced and dimensionally rich representation.

As depicted in Figure 2, the token sequence is processed through our location encoder and text embedding layer, formally generating the input for training within the transformer architecture. Finally, the output obtains a binary classification result of 0 or 1 through linear layers and softmax functions.

## 5. Experiment and Discussion

In this section, we first introduce how named place pairs are selected and labeled as experimental data, then evaluate the SSARLM model for named place pair matching. Based on the matched entities, a provenance-aware place knowledge graph is constructed, using named places in China from the three data sources in Section 2.

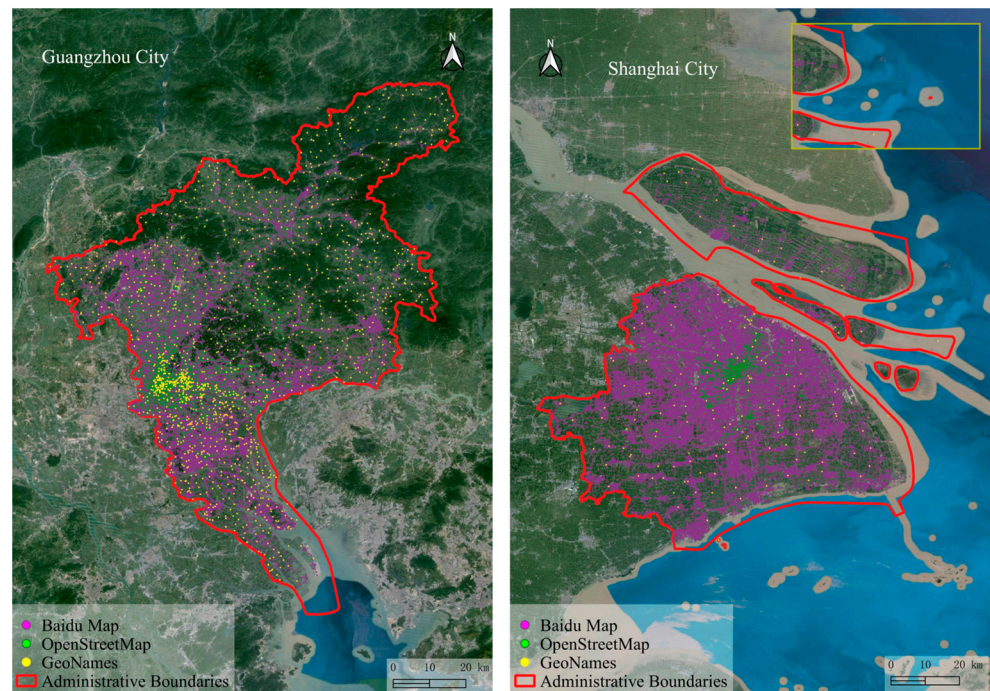
### 5.1. Data Preparation

The datasets used in this paper include 758,860 gazetteer entities from Chinese GeoNames, 287,182 POIs from Chinese OSM, and 1,127,582 POIs from Baidu Map covering six cities in China, i.e., Guangzhou, Hangzhou, Shanghai, Shenzhen, Zhuhai, and Wenzhou. Our goal is to establish equivalent associations between POIs from the three data sources. The basis of model training is whether the data is correlated or not. Therefore, each training data instance consists of two POIs from distinct sources (named place pair), along with a binary label (0/1) indicating whether they match. In theory, any two POIs from different sources can be used for model training. However, considering all possible combinations would result in excessively large datasets, not only significantly increasing the task of labeling annotations, but also imposing a burden on model training. Therefore, it is necessary to impose certain limiting conditions to ensure that we obtain relatively high-quality named place pairs.

To reduce the number of candidates matched named place pairs, our approach first filters mismatched named place pairs with a 1000 m distance threshold, a 0.4 place name similarity threshold, and a 0.5 place type similarity threshold. Then, for match prediction model training and test, this paper manually labeled matched and mismatched (similar but not matching) place pairs in Guangzhou and Shanghai from filtered candidates as positive and negative samples, as shown in Figure 3. Table 3 shows the statistics of matched and mismatched samples of place entity pairs from three data sources in two areas. It is observed that, despite the filtering process, the quantity of negative samples remains substantially large, with the ratio of positive to negative samples approaching 1:50. To mitigate the adverse effects caused by this imbalance, we extracted 15,000 data entries from the Guangzhou and Shanghai datasets, maintaining a ratio of approximately 1:3 between positive and negative samples.

**Table 3.** Statistics of the match and mismatched samples of named place pairs.

Area	Named Place Pair Source	Matched Sample Count	Mismatched Sample Count
Guangzhou	Baidu Map—OSM	1550	82,279
	Baidu Map—GeoNames	575	36,863
	OSM—GeoNames	106	812
Shanghai	Baidu Map—OSM	1767	79,579
	Baidu Map—GeoNames	900	51,937
	OSM—GeoNames	135	2730



**Figure 3.** Three data sources in Guangzhou and Shanghai.

## 5.2. Model Evaluation

The named place pair matching can be regarded as a classification problem with two possible outcomes (match or mismatch). We use Precision (P), Recall, and F1 as measurement indexes to evaluate the models' performance. They can be calculated using Formulas (4)–(6), where TP is the number of true positive results, FP is the number of false positive results, FN is the number of false negative results.

$$P = TP / (TP + FP) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

$$F1 = 2P / (P + \text{Recall}) \quad (6)$$

### 5.2.1. Overall Performance Evaluation

The named place matching model SSARLM in our paper is compared with several classical and commonly used binary classification machine learning models, including Support Vector Classifier (SVC), Random Forest Classifier (RFC), and Multilayer Perceptron (MLP). The prerequisite for utilizing these models is the computation of various similarity measures, including string similarity, phonetic similarity, and bag-of-words similarity, as well as location similarity and type similarity. It is important to highlight that type similarity relies on the type fusion system we have constructed. This system integrates types from three different data sources, referencing the national POI classification standards. Consequently, each type from every data source can find its corresponding representation within our system. Simultaneously, we conducted a comparative analysis with the MMCNN, which is a multi-layer neural network model specifically devised to place entity matching. The MMCNN consists of a two-layer architecture: the first layer is trained on features associated with the names of geographic entities, including string similarity, phonetic similarity, and bag-of-words similarity; the second layer is trained on a composite feature set that includes name similarity, type similarity, and location similarity. We contend that this model offers a comprehensive empirical representation of the pivotal attributes of geographic entities.

We utilized data from the Guangdong region to conduct quintuplicate experiments on each model. The data was partitioned into training, validation, and test sets in an 8:1:1 ratio. For the SSARLM model, we set the learning rate at  $5 \times 10^{-7}$ , with a batch size of 32, and epochs capped at 50. The maximum sequence length was configured to 256. Because the experiment's objective is to compare the optimal performance differences among various models, the hyperparameters for the other four models were not uniformly constrained.

Existing research has substantiated that large language models, exemplified by ChatGPT and GPT-4, achieve superior performance in a variety of natural language downstream tasks, even attaining state-of-the-art (SOTA) status [36,37]. In our study, we evaluated the performance of GPT-4 (zero-shot) and GPT-4 (20-shot) in the task of place entity matching. This approach enables direct testing on the test dataset without the necessity for training data.

For the GPT-4 model, we engineered a high-quality prompt specifically tailored to optimize its performance in the location entity matching task. In the case of GPT-4 (20-shot), we additionally selected examples from the training dataset that comprehensively covered all cases, employing these as in-context learning materials.

The final averaged evaluation metrics for each model are presented in Table 4. Notably, SSARLM emerges as the only model achieving an F1 score of 0.95, while the other four machine learning models consistently fall within the 0.93–0.94 range. Despite GPT-4's status as the most powerful current large language model, it demonstrates a relatively lower F1 score in the domain-specific task of entity matching. These results validate the efficacy and advancement of SSARLM. Given that the ultimate goal of our experiment is to achieve geographic entity matching across six cities and construct a knowledge graph, we further assessed the generalization performance of each model on data from other regions.

**Table 4.** Performance of different models in Guangzhou.

Model	Precision	Recall	F1
RFC	0.9338	0.9310	0.9323
SVC	0.9353	0.9434	0.9393
MLP	0.9353	0.9444	0.9398
MMCNN	0.9239	0.9414	0.9325
GPT-4 (zero-shot)	0.8820	0.5850	0.7034
GPT-4 (20-shot)	0.8916	0.6156	0.7284
SSARLM	0.9268	0.9829	0.9539

### 5.2.2. Generalization Performance Test

We have archived the training outcomes of each model based on the Guangdong region data and applied these models to data from the Shanghai region. This approach is employed to evaluate the generalization capabilities of each model. The results, presented in Table 5, are derived from averaging five test iterations for each model.

**Table 5.** Test results for different models in Shanghai.

Model	Precision	Recall	F1
RFC	0.8814	0.7876	0.8319
SVC	0.9039	0.8024	0.8501
MLP	0.8945	0.8062	0.8480
MMCNN	0.8926	0.8116	0.8501
SSARLM	0.8521	0.8633	0.8574

Upon a holistic consideration of various performance metrics, it is observed that our proposed SSARLM model exhibits enhanced overall effectiveness on the Shanghai dataset. It surpasses the F1 scores of the other four models, notably outperforming MMCNN, MLP,



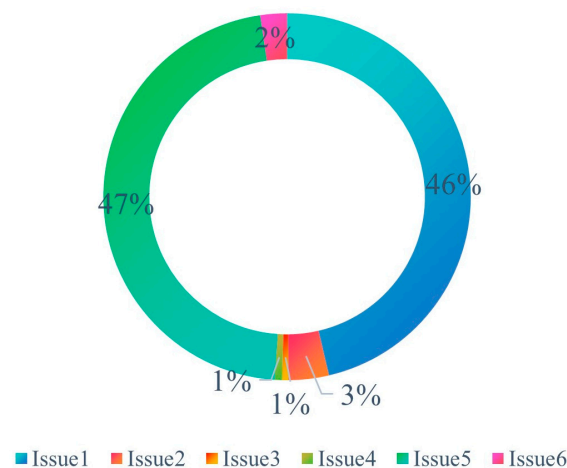
and SVC by 0.7 to 1 percentage points, and RFC by 2.6 percentage points. When comparing Precision and Recall values, the SSARLM model demonstrates a more balanced performance across all indicators. Furthermore, in the context of our entity matching task, the probability of accurately predicting positive samples emerges as a crucial metric.

Analysis of the results indicates that SSARLM effectively learns the common features of geographic entities. It robustly and comprehensively assesses the matching degree across data sources with regional differences, demonstrating strong generalization capabilities.

### 5.2.3. Statistics on the Resolution of Challenging Issues

Utilizing the predictive results from Shanghai, we statistically assessed the ability of five models to address the first six challenging issues described in Section 3.2. The remaining three issues, which are common to nearly all data pairs, will not be separately discussed further. Our analysis focused on original positive samples, with issue annotations for these samples obtained through both automated methods and manual verification.

Figure 4 presents the statistical distribution of various issues, revealing that Issues I1 and I5 have the highest proportions. While other issues have smaller shares, they still appear dozens of times in the positive samples from Shanghai and can be considered for evaluation purposes. Given that each model underwent five tests, to ensure reliability, we consider a model to accurately classify a data pair only if it predicts correctly in three or more tests.



**Figure 4.** Statistical results of issues for positive samples.

Table 6 presents the probability statistics of correctly classified positive samples by five models under various challenging issues. In the first six scenarios, SSARLM outperforms the other four models. Notably for I2, SSARLM shows a distinct advantage, exceeding the other models by nearly 40 percentage points. Furthermore, in case I4, SSARLM's performance is more than double that of the other models. In the other scenarios, SSARLM also consistently surpasses the others to varying degrees. Among the remaining four models, the performance differences are minimal, with MMCNN exhibiting relative superiority.

**Table 6.** Statistic results of challenging issues evaluation accuracy.

	SSARLM	MMCNN	MLP	SVC	RFC
I1: names expressed in different languages	0.7953	0.7776	0.7824	0.7671	0.7694
I2: names with different cases	0.5385	0.1077	0.1385	0.1077	0.1385
I3: names with spelling errors	0.9000	0.8000	0.7000	0.8000	0.7000
I4: names with different word orders	0.7500	0.3333	0.3333	0.1667	0.3333
I5: names with abbreviations	0.9602	0.8898	0.8759	0.8735	0.8607
I6: names with synonyms	0.9070	0.7907	0.7209	0.8140	0.7209



### 5.3. Discussion

The implementation results in Section 5.2 show that the similarity measure methods and the SSARLM model presented in this paper can help find similar named places from different data providers with high accuracy. The matched named places conflation and PlaceKG construction methods help to create a merged place dataset.

In Section 5.2.1, this paper compares the SSARLM based on a large-scale pre-trained model, various machine learning-based models, and models based on GPT-4. We initially conducted multiple training sessions for SSARLM and other machine learning models, using the training dataset from Guangzhou. In the case of the GPT-4 models, testing was directly performed using the test dataset. Subsequently, in Section 5.2.2, we conducted a generalization performance test for SSARLM and machine learning models using data from Shanghai.

The machine learning models, including MMCNN and MLP, realize alignments automatically between sources and targets. However, implementing such methods requires the use of empirical formulas for calculating various features, undoubtedly leading to a relatively substantial workload. At the same time, in our practice, a mature and complete place type category with full semantic information is required for type alignments of all named place datasets. Due to the limited types in the three data sources, manual creation and alignment is feasible. Unfortunately, one typical drawback of the manual method is that it may cause some extension problems when more named place data sources are included. Furthermore, manually assigning types to them requires a certain standard as the basis and relies on the builder's extensive experience. Each type of mapping needs to undergo careful verification, resulting in a heavy workload.

Our proposed SSARLM model develops an end-to-end model based on the large-scale pre-trained model to replace manual feature selection. This provides a more streamlined and efficient approach, eliminating the need for separate feature engineering steps. Furthermore, whether tested within Guangzhou or using the trained model on Shanghai data, our model exhibits a higher F1 score performance compared to others. It demonstrates a significantly higher Recall than competing models, indicating its superior accuracy in identifying matching geographical entities.

Both the SSARLM model and other machine learning models require high-quality sample data. Despite demonstrating the model's generalization capabilities in location matching across different regions in Section 5.2.2, we remain curious about whether it can replace training samples to reduce the resources needed for location matching. Several existing studies have shown that large models like ChatGPT and GPT-4 can be directly applied to downstream natural language tasks without any fine-tuning or training to update model parameters [38]. Large language models are more robust than large-scale pre-trained models, possessing greater parameter volume and advanced language comprehension capabilities. In our experiments, we did not use training datasets; instead, we obtained location-matching results by engaging GPT-4 in dialogue using high-quality prompts. The experiment results show that the GPT-4-based model (zero-shot) achieved high Precision, but lower Recall. This suggests that GPT-4 might not effectively recognize that two names refer to the same location in certain specific scenarios or subtle differences, indicating that high-quality training data is still a necessary option for location entity matching tasks. Based on the concept of domain-specific fine-tuning for large models, substantial improvements in performance on specific tasks and enhanced tracking of human instructions can be achieved through extensive fine-tuning with a large volume of high-quality, location-matching-related training data. Ultimately, fine-tuning techniques for large models could enable an out-of-the-box place entity matching functionality based on natural human language instructions.

In Section 5.2.3, our statistical analysis of the six challenging issues presented demonstrates a distinct advantage of SSARLM in addressing complex geographical entity matching problems. Specifically, for names with different cases (I2) and different word orders (I4), SSARLM's metrics significantly exceed those of other models several times. These outcomes are likely attributed to the strengths of the pre-trained model DistilBERT. We selected a case-sensitive version of the model, thereby enhancing its focus on the variation in the capitalization of place names and their relevance to geographical entity matching during the training phase. Furthermore, BERT-based models utilize an attention mechanism to capture context without emphasizing the necessity of word order. This feature is beneficial for place name matching, as the inversion of word order typically does not alter the description of a location. Such reversals are common across various data sources. Overall, SSARLM demonstrates greater flexibility in representing place names, showing a clear performance advantage.

#### 5.4. PlaceKG Construction

Using the SSARLM model, we can obtain matched named places through the three datasets. The statistical results are shown in Table 7. In total, 1365 similar named places appear simultaneously in all three datasets, and 7264, 8663, and 97,453 matched named places separately in the pairwise matching of the three datasets. OSM has the most matched named places with other datasets, where 36.9 percent of POIs from OSM can find matched named places in GeoNames and Baidu Map. Only 1.5 percent of POIs from Baidu Map can find matched entities in other datasets. Since POIs from Baidu Map only cover six cities, and Baidu Map is a well-localized company, it has stronger capabilities in collecting POIs in China than other data providers. It is then necessary to conflate the matched named places and resolve conflicts when creating a place knowledge graph.

**Table 7.** Statistic results of entities matched among three datasets.

Named Place Pair Source	Matched Named Place Pairs Count
Baidu Map—OSM	7264
Baidu Map—GeoNames	8663
OSM—GeoNames	97,453
Baidu Map—OSM—GeoNames	1365

As a result, a provenance-aware place knowledge graph named PlaceKG is developed using the conflation strategies and the provenance model recording source entities. The PlaceKG contains 2,076,693 PlaceEntity instances in China and their provenance information. It is encoded using RDF format and currently includes 57,801,364 statements.

The PlaceKG can be further published on the SPARQL server and queried by the SPARQL. Figure 5 provides examples of SPARQL queries on the PlaceKG with Web UI. The examples show how entities in the PlaceKG be searched and how their provenance information can be tracked for analysis. In Figure 5a, a PlaceEntity instance is queried by its Chinese name. In Figure 5b, matched named places from three data sources that generate the PlaceEntity instance can be traced; In Figure 5c, the strategies exploited for handling conflicts when conflating place names can also be explored. Finally, in Figure 5d, other properties, like the location of the PlaceEntity instance, can be obtained from the PlaceKG.



**Author Contributions:** Lianlian He: Methodology, supervision of the project, data analysis, figure preparation, and writing of the original draft. Hao Li: Methodology, writing of the original draft, experimental design. Rui Zhang: Figure preparation, data analysis, and literature review. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work was supported by the Educational Commission of Hubei Province of China.

**Data Availability Statement:** The data that support the findings of the present study are available on Figshare at <https://figshare.com/s/c11efc6ca73fb4306b3e> (accessed on 1 December2023).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Chinese-English Translation

Chinese	English
咖啡厅	Cafe
茶座	Teahouse
从化区办证中心	Conghua District Government Licence Application Center
从化区政务服务中心	Conghua District Government Service Center
地铁浔峰岗站	Xunfenggang Metro Station
火车站	Railway Station
珞瑜路	Luoyu Road
珞喻路	Luoyu Road
公交车站	Bus Station
购物	Shopping
广州东站	Guangzhou East Railway Station
广州火车站	Guangzhou Railway Station
广州站	Guangzhou Railway Station
交通设施	Traffic Facilities
美食	Fine Food
三林镇	Sanlin Town
商铺	Store
收费站	Toll Booth
新华书店	Xinhua Bookstore
新华书店(北京路一店)	Xinhua Bookstore, Beijing Road No. 1 Store
浔峰岗地铁站	Xunfenggang Metro Station
越秀公园A	Exit A, Yuexiu Park Subway Station
越秀公园-A口	Exit A, Yuexiu Park Subway Station
越秀公园C	Exit C, Yuexiu Park Subway Station
越秀公园-C口	Exit C, Yuexiu Park Subway Station
中餐厅	Chinese Restaurant
地址	Address
电话	Telephone
坐标	Coordinate
途径地铁	By Subway
广州市天河区东路1号	No.1 Dongzhan Road, Tianhe District, Guangzhou
广东省广州市越秀区环市东路367号白云宾馆	Baiyun Hotel, No. 367 Huanshi East Road, Yuexiu District, Guangzhou City, Guangdong Province
湖北省武汉市洪山区	Hongshan District, Wuhan City, Hubei Province
地铁6号线	Subway Line 6
中国铁路广州局集团有限公司	China Railway Guangzhou Group Co., Ltd
广州市越秀区环市西路159号	159 Huanshi West Road, Yuexiu District, Guangzhou City
广州市从化区河滨北路128号城晖大厦	Chenghui Building, No. 128 Hebin North Road, Conghua District, Guangzhou City

## References

1. Thakuriah, P.; Tilahun, N.Y.; Zellner, M. *Seeing Cities through Big Data: Research, Methods and Applications in Urban Informatics*; Springer: Cham, Switzerland, 2017; p. 554.
2. Manville, C.; Cochrane, G.; Jonathan, C.A.V.E.; Millard, J.; Pederson, J.K.; Thaarup, R.K.; WiK, J.K.; WiK, M.W. *Mapping Smart Cities in the EU*; European Parliamentary Research Service: Brussels, Belgium, 2014; p. 200.

3. Allemang, D.; Hendler, J. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*; Elsevier: Amsterdam, The Netherlands, 2011.
4. Liu, J.; Guo, D.; Liu, G.; Zhao, Y.; Yang, W.; Tang, L. Construction Method of City-Level Geographic Knowledge Graph Based on Geographic Entity. In Proceedings of the International Conference on Geoinformatics and Data Analysis, ICGDA 2022, Paris, France, 21–23 January 2022; pp. 133–142.
5. Kuhn, W.; Kauppinen, T.; Janowicz, K. Linked data—a paradigm shift for geographic information science. In Proceedings of the Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, 24–26 September 2014; pp. 173–186.
6. Mai, G.; Janowicz, K.; Cai, L.; Zhu, R.; Regalia, B.; Yan, B.; Shi, M.; Lao, N. SE-KGE: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Trans. GIS* **2020**, *24*, 623–655. [\[CrossRef\]](#)
7. Du, J.; Wang, S.; Ye, X.; Sinton, D.S.; Kemp, K. GIS-KG: Building a large-scale hierarchical knowledge graph for geographic information science. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 873–897. [\[CrossRef\]](#)
8. Chen, J.; Deng, S.; Chen, H. Crowdgeokg: Crowdsourced geo-knowledge graph. In Proceedings of the Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence: Second China Conference, CCKS 2017, Chengdu, China, 26–29 August 2017; pp. 165–172.
9. Sun, K.; Zhu, Y.; Song, J. Progress and challenges on entity alignment of geographic knowledge bases. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 77. [\[CrossRef\]](#)
10. Ma, X. Knowledge graph construction and application in geosciences: A review. *Comput. Geosci.* **2022**, *161*, 105082. [\[CrossRef\]](#)
11. McKenzie, G.; Janowicz, K.; Adams, B. Weighted multi-attribute matching of user-generated points of interest. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA, 5 November 2013; pp. 440–443.
12. Li, L.; Xing, X.; Xia, H.; Huang, X. Entropy-weighted instance matching between different sourcing points of interest. *Entropy* **2016**, *18*, 45. [\[CrossRef\]](#)
13. Novack, T.; Peters, R.; Zipf, A. Graph-based matching of points-of-interest from collaborative geo-datasets. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 117. [\[CrossRef\]](#)
14. Zhou, C.; Zhao, J.; Zhang, X.; Ren, C. Entity alignment method of points of interest for internet location-based services. *J. Adv. Comput. Intell. Inform.* **2020**, *24*, 837–845. [\[CrossRef\]](#)
15. Khodizadeh-Nahari, M.; Ghadiri, N.; Baraani-Dastjerdi, A.; Sack, J.R. A novel similarity measure for spatial entity resolution based on data granularity model: Managing inconsistencies in place descriptions. *Appl. Intell.* **2021**, *51*, 6104–6123. [\[CrossRef\]](#)
16. Zhou, Y.; Wang, M.; Zhang, C.; Ren, F.; Ma, X.; Du, Q. A points of interest matching method using a multivariate weighting function with gradient descent optimization. *Trans. GIS* **2021**, *25*, 359–381. [\[CrossRef\]](#)
17. Santos, R.; Murrieta-Flores, P.; Calado, P.; Martins, B. Toponym matching through deep neural networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 324–348. [\[CrossRef\]](#)
18. Zhang, W.; Wang, C. A machine learning approach to improve place name matching through user-generated content. *Cartogr. Geogr. Inf. Sci.* **2019**, *46*, 229–242.
19. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Available online: <https://openai.com/research/language-unsupervised> (accessed on 22 December 2023).
20. Zhao, L.; Deng, H.; Qiu, L.; Li, S.; Hou, Z.; Sun, H.; Chen, Y. Urban multi-source spatio-temporal data analysis aware knowledge graph embedding. *Symmetry* **2022**, *12*, 199. [\[CrossRef\]](#)
21. Chen, Q.; Zhuo, Z.; Wang, W. Bert for joint intent classification and slot filling. *arXiv* **2019**, arXiv:1902.10909.
22. Gritta, M.; Pilehvar, M.T.; Limsopatham, N.; Collier, N. What’s missing in geographical parsing? *Lang. Resour. Eval.* **2018**, *52*, 603–623. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Melo, F.; Martins, B. Automated geocoding of textual documents: A survey of current approaches. *Trans. GIS* **2017**, *21*, 3–38. [\[CrossRef\]](#)
24. Li, H.; Yue, P.; Li, S.; Zhang, C.; Yang, C. Spatio-temporal intention learning for recommendation of next point-of-interest. *Geo-Spat. Inf. Sci.* **2023**. [\[CrossRef\]](#)
25. Leidner, J.L.; Lieberman, M.D. Detecting geographical references in the form of place names and associated spatial natural language. *Sigspatial Spec.* **2011**, *3*, 5–11. [\[CrossRef\]](#)
26. Han, B.; Cook, P.; Baldwin, T. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.* **2014**, *49*, 451–500. [\[CrossRef\]](#)
27. Cucchiarelli, A.; Luzi, D.; Velardi, P. Automatic semantic tagging of unknown proper names. In Proceedings of the COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, Montreal, QC, Canada, 10 August 1998; pp. 286–292.
28. Points of Interest Core. Available online: <https://www.w3.org/2010/POI/documents/Core/core-20111216.html> (accessed on 10 December 2021).
29. Saalfeld, A. Conflation automated map compilation. *Int. J. Geogr. Inf. Sci.* **1988**, *2*, 217–228. [\[CrossRef\]](#)
30. Ruiz, J.J.; Ariza, F.J.; Urena, M.A.; Blázquez, E.B. Digital map conflation: A review of the process and a proposal for classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1439–1466. [\[CrossRef\]](#)
31. Low, R.; Tekler, Z.D.; Cheah, L. An end-to-end point of interest (POI) conflation framework. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 779. [\[CrossRef\]](#)



32. The GeoNames Geographical Database. Available online: <https://www.geonames.org/about.html> (accessed on 21 November 2023).
33. Map Features. Available online: [https://wiki.OSM.org/wiki/Map\\_Features](https://wiki.OSM.org/wiki/Map_Features) (accessed on 21 November 2023).
34. Baidu LBS Cloud Service. Available online: <http://lbsyun.baidu.com/index.php?title=lbscloud/poitags> (accessed on 21 November 2023).
35. Mai, G.; Janowicz, K.; Yan, B.; Zhu, R.; Cai, L.; Lao, N. Multi-scale representation learning for spatial feature distributions using grid cells. *arXiv* **2020**, arXiv:2003.00824.
36. Kalyan, K.S. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *arXiv* **2023**, arXiv:2310.12321. [[CrossRef](#)]
37. Thapa, S.; Naseem, U.; Nasim, M. From humans to machines: Can ChatGPT-like LLMs effectively replace human annotators in NLP tasks. In Proceedings of the Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media, Limassol, Cyprus, 5 June 2023.
38. Chang, E.Y. Examining GPT-4: Capabilities, Implications and Future Directions. In Proceedings of the 10th International Conference on Computational Science and Computational Intelligence, London, UK, 14–16 July 2023.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.