International Journal of
Geo-Information

MDPI

Article

Knowledge Embedding with Geospatial Distance Restriction for Geographic Knowledge Graph Completion

Peiyuan Qiu [1], Jialiang Gao [1,2], Li Yu [3] and Feng Lu [1,2,4,5,*]

[1] State Key Laboratory of Resources and Environmental Information System,
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences,
Beijing 100101, China; qiupy@lreis.ac.cn (P.Q.); gaojl@lreis.ac.cn (J.G.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] National Science Library, Chinese Academy of Sciences, Beijing 100190, China; yul@mail.las.ac.cn
[4] Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou 350003, China
[5] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: luf@lreis.ac.cn; Tel.: +86-10-6488-8966

Received: 4 May 2019; Accepted: 26 May 2019; Published: 30 May 2019

Abstract: A Geographic Knowledge Graph (GeoKG) links geographic relation triplets into a large-scale semantic network utilizing the semantic of geo-entities and geo-relations. Unfortunately, the sparsity of geo-related information distribution on the web leads to a situation where information extraction systems can hardly detect enough references of geographic information in the massive web resource to be able to build relatively complete GeoKGs. This incompleteness, due to missing geo-entities or geo-relations in GeoKG fact triplets, seriously impacts the performance of GeoKG applications. In this paper, a method with geospatial distance restriction is presented to optimize knowledge embedding for GeoKG completion. This method aims to encode both the semantic information and geospatial distance restriction of geo-entities and geo-relations into a continuous, low-dimensional vector space. Then, the missing facts of the GeoKG can be supplemented through vector operations. Specifically, the geospatial distance restriction is realized as the weights of the objective functions of current translation knowledge embedding models. These optimized models output the optimized representations of geo-entities and geo-relations for the GeoKG's completion. The effects of the presented method are validated with a real GeoKG. Compared with the results of the original models, the presented method improves the metric Hits@10(Filter) by an average of 6.41% for geo-entity prediction, and the Hits@1(Filter) by an average of 31.92%, for geo-relation prediction. Furthermore, the capacity of the proposed method to predict the locations of unknown entities is validated. The results show the geospatial distance restriction reduced the average error distance of prediction by between 54.43% and 57.24%. All the results support the geospatial distance restriction hiding in the GeoKG contributing to refining the embedding representations of geo-entities and geo-relations, which plays a crucial role in improving the quality of GeoKG completion.

Keywords: geographic knowledge graph; geographic knowledge embedding; knowledge graph completion; geographic relation triplet

1. Introduction

A Knowledge Graph (KG) is a system that understands facts about people, places and things, and how these entities are all connected [1]. To emphasize the "connection", facts in a KG are represented as triplets with the form <head entity, relation, tail entity> or <entity, property, value>

(both abbreviated as <**h**, **r**, **t**>). For example, the knowledge "Gorgona is a major island of Elba" is represented as <Elba, major island, Gorgona>, and "Elba has seven islands" is represented as <Elba, number of islands, 7>. Then, the knowledge facts form a graph similar to the example illustrated in Figure 1. The semantic difference of entities and relations can be distinguished by mining the graph links of these entities and relations. This difference is more significant than that found in the traditional knowledge base, where semantic differences of entities are only expressed by the independent property values of the entities. Thus, KGs can facilitate artificial intelligence (AI) applications, such as semantic searching [2,3], question answering (QA) [4], and smart education [5]. As a special KG, a geographic KG (GeoKG) becomes the effective organization form of geographic information, especially the geographic relation triplets extracted from web texts existing in newswires, collaborative encyclopedias, social media, official or domain websites, and so on.

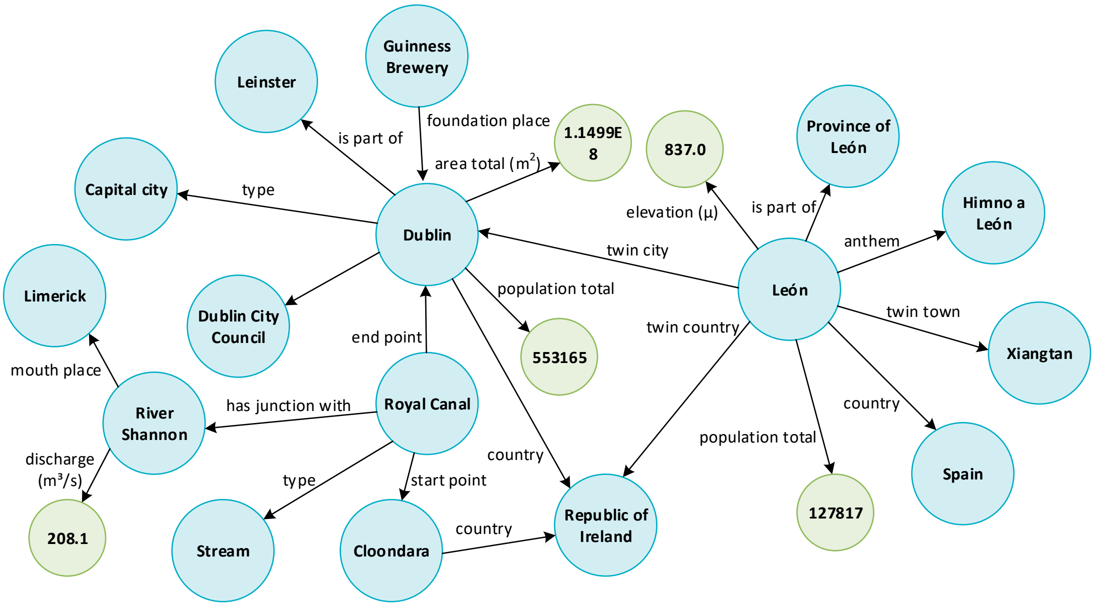


Figure 1. Example of part of a knowledge graph (the blue circles represent entities; the green circles represent values; edges represent different types of relations or properties between entities and values).

However, current GeoKGs are far from complete, due to missing geographic entities (geo-entities) or geographic relations (geo-relations) in their fact triplets. In English DBpedia (https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10), 53.86% of lake entities lack any source of water (property "inflow"), and 85.80% of mountain entities have no fact describing where their parent peaks are (property "parent mountain peak"). The sparsity of geographic information distribution on the web is the major reason for this data being unavailable. During the construction of a GeoKG, information extraction systems are often unable to detect enough references to geographic information from the massive web resource in a limited amount of time. Furthermore, the professionalism of geographic knowledge also reduces the likelihood that collected geographic information will be fully transformed into structured knowledge. Consequently, the generated GeoKGs will inevitably miss a large number of geographic knowledge facts. The incompleteness of the necessary information seriously impacts the performance of GeoKG applications, because GeoKGs only supply limited facts for a query or inference. Therefore, KG completion, used to supplement missing facts and guarantee basic data completion, has become an increasingly important task of KG research [6].

Translation knowledge embedding (translation KE) models are effective tools to complete KGs, and are able to achieve state-of-the-art completion performance [7,8]. These methods use the known entities and relations in a KG to fill in the missing facts by mining the potential semantics between the known entities. More concretely, the translation KE models encode (or "embed") the semantics of knowledge (both entities and relations) into a continuous, low-dimensional vector space [9]. Then, the missing entities or relations of facts will be predicted by a vector operation. For example, the vectors $\boldsymbol{l}_{\mathrm{h}}$ and $\boldsymbol{l}_{\mathrm{r}}$ of the entity **h** and the relation **r** are encoded by a translation KE model, then the missing **t** of the triplet <**h**, **r**, **?**> can be predicted with the operation $\boldsymbol{l}_{\mathrm{h}} + \boldsymbol{l}_{\mathrm{r}}$. The capacities of the translation KE methods have been verified in many studies, but these methods may perform poorly in the GeoKG completion. The translation KE models used in these methods assume that the entity should have distinct representations for different relations to improve the rationality of knowledge embedding. Unfortunately, the sparsity of geographic information distribution on the web results in sparse links in GeoKGs. The geo-entities in GeoKGs only connect a few other geo-entities, and one type of geo-relation contains a limited number of fact triplets. Thus, these models cannot obtain enough triplets of each type of geo-relation as training data to adequately represent geo-entities and geo-relations.

Meanwhile, the geospatial information hiding in a GeoKG can become the crucial additional information to enhance the representations of geographic knowledge of a link-sparse GeoKG. Intuitively, the symbolic geo-entity in a GeoKG must refer to a precise (or vague) geospatial location or scope. Then, the symbolic geo-relations between two geo-entities are actually the reflection of the geospatial distance restriction of these two geo-entities. That is, the geo-relations which can be used in describing two geo-entities are restricted by their geospatial pattern. However, current models fail to utilize these geospatial patterns. In this paper, a method with geospatial distance restriction is presented to optimize knowledge embedding for supplying the missing geographic facts of a link-sparse GeoKG. This method aims to optimize the training process of current translation KE models using the geospatial distance restriction. Then, the geo-entities with same distance in a geographic space will maintain a similar distance from each other in the embedding vector space. Finally, the missing geographic facts can be predicted by a vector operation according to both the semantic relationship and the geospatial distance feature between geo-entities in the GeoKG.

To summarize, our main contributions are as follows:

1. The presented method introduces a geospatial distance restriction to refine the embedding representations of geographic knowledge in a link-sparse GeoKG, which fuses geospatial information and semantic information into a low-dimensional vector space;
2. From the viewpoint of GIS (Geographic Information System), a novel task is designed to predict the geospatial locations of unknown geo-entities. This task is different from the tasks of the current translation KE research which only focuses on measuring the semantic relationship.

The rest of this paper is organized as follows. A brief review of translation KE model is introduced in Section 2. Section 3 proposes the GeoKG datasets used in this study, the optimization method with geospatial distance restriction for a translation KE model, and the workflow of GeoKG completion by a translation KE model. The experimental datasets, evaluation tasks and results, and result analysis are presented in Section 4. Section 5 is devoted to discussion, and Section 6 concludes this work.

2. Review of Translation KE Model

Translation KE is inspired by word embedding methods. For word embedding, the words in the corpus are encoded into a continuous low-dimensional semantic vector space, where each word is represented by a fixed dimensional real-valued vector [10,11]. If the distance between two words is close, these words have similar semantics or related semantics [12]. For example, the distance between "France" and "U.S.A" (or, "France" and "French") is less than the distance between "France" and "Mountain" in the vector space. Likewise, translation KE model aims to encode both entities and their relations in a KG into a continuous low-dimensional semantic vector space. In this space, the vector of the head entity projected ("translated") by the vector of the relationship should be similar to the vector of the tail entity. For example, as shown in Figure 2, after embedding the triplet "<Black Sea, inflow, Dniester>", the calculation result $\boldsymbol{l}_{Black\ Sea} + \boldsymbol{l}_{inflow} \approx \boldsymbol{l}_{Dniester}$ can be outputted, where $\boldsymbol{l}_{Black\ Sea}$, $\boldsymbol{l}_{inflow}$ and $\boldsymbol{l}_{Dniester}$ is the vector of "Black Sea", "inflow" and "Dniester".

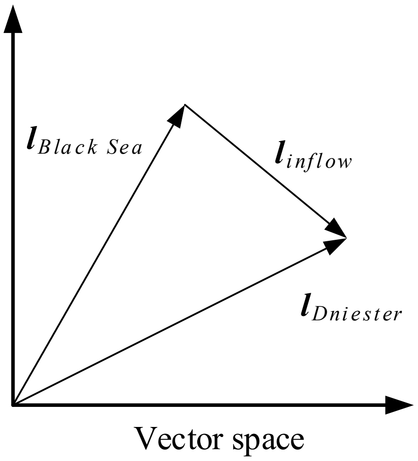


Figure 2. Simple illustration of a translation KE model.

A typical translation KE model consists of three steps [13]: (1) representing entities and relations, (2) defining a scoring function, and (3) learning entity and relation representations from observed facts in the current KG. Let us take the first translation KE model TransE [9] as an illustration:

Firstly, TransE represents entities and relations as vectors in the same space. The vectors of relations are used for translating the head vector to the tail vector.

Secondly, a scoring function is defined to measure the plausibility of representations as:

$$f_r(\boldsymbol{h}, \boldsymbol{t}) = -\|\boldsymbol{l}_h + \boldsymbol{l}_r - \boldsymbol{l}_t\|_{1/2} \tag{1}$$

where $\boldsymbol{l}_h$, $\boldsymbol{l}_t$ and $\boldsymbol{l}_r$ are the vectors of $\boldsymbol{h}$, $\boldsymbol{t}$ and $\boldsymbol{r}$ of a triplet <$\boldsymbol{h}$, $\boldsymbol{r}$, $\boldsymbol{t}$>. If a triplet <$\boldsymbol{h}$, $\boldsymbol{r}$, $\boldsymbol{t}$> appears in the KG, it is regarded as a positive triplet. $f_r(\boldsymbol{h}, \boldsymbol{t})$ of the positive triplet should be low, and is high otherwise.

Thirdly, TransE learns the representations of an entity and its relations through minimizing the following objective function, which is defined based on a margin-based ranking loss.

$$L = \sum_{\boldsymbol{h}.\boldsymbol{t}} \sum_{\boldsymbol{h}',\boldsymbol{t}'} \max(0, f_r(\boldsymbol{h}, \boldsymbol{t}) + \gamma - f_r(\boldsymbol{h}', \boldsymbol{t}')) \tag{2}$$

where $\boldsymbol{h}'$ and $\boldsymbol{t}'$ are the head entity and tail entity in negative triplet (the triplet does not appear in the KG), and $\gamma > 0$ is a margin hyperparameter separating the positive triplets and negative triplets. This objective function is designed to improve the gap between positive triplets and negative triplets. Generally, because there are no unreasonable triplets collected into the KG, the negative triplets are generated by replacing the head entity or tail entity of a positive triplet with a random entity. The optimization is realized by stochastic gradient descent (SGD) [14].

The extension models of TransE are proposed to improve the accuracy of completion based on more complex restrictions. These extension models optimize the effects of entity and relation representation through:

(1) mining the internal restrictions in KG fact triplets. For example, one head entity (tail entity) has multiple tail entities (head entities) under a relation by TransH [15]; one relation is used for different semantic of entities by TransR [7], CTransR [7], TransG [16], TransD [8], TransA [17]; the heterogeneity and imbalance that exists in the KG by TranSparse [18], and so forth;
(2) adding extra information aside from fact triplets, such as entities' types by TKRL [19], entities' attributes by TransEA [20], entities' textual descriptions by DKRL [21], relation paths by PTransE [22], graph structures by GAKE [23] and TCE [24], facts' temporal information [25], and so on.

All of the above translation KE models assume that the entity should have distinct representations for different relations to improve the rationality of knowledge embedding. Unfortunately, these models will achieve poor performance on the link-sparse GeoKG because the geo-entities in GeoKGs only connect a few other geo-entities, and one type of geo-relation contains a limited number of fact triplets.

3. Materials and Methods

First, we extract two GeoKG datasets from the general datasets. Next, we present the method to optimize the translation KE model with geospatial distance restriction. Then, we introduce the workflow for completing a GeoKG by translation KE model.

3.1. GeoKG Extraction

There are currently no mature and open GeoKG datasets available, so we extract two GeoKGs separately from open datasets DBpedia and GADM.

3.1.1. GeoKG from DBpedia

DBpedia is a KG project designed to extract structured information from Wikipedia (https://wiki.dbpedia.org/about), such as entity's category, entity type, properties, coordinates (geo-entities), abstract, context, and so on. DBpedia dumped its data by text file, where the raw structured information is organized as an n-triple format as "<http://dbpedia.org/resource/Lake_Erie http://dbpedia.org/property/inflow http://dbpedia.org/resource/Detroit_River>". This information can be easily parsed into a fact triplet form as "<Lake Erie, inflow, Detroit River>". The DBpedia as a general KG contains different types of entities, so its geo-entities can reflect the sparseness of geographic information on the web. We use the English DBpedia version 2016-10 to build a GeoKG. This version contains 6.6 million entities and 1.7 billion triplets by 57 dump files. The details of generation are as follows.

(1) Using "category" and "entities' types" files to save the triplets whose head entities and tail entities are both geo-entities. Here, the entities belonging to the DBpedia categories "Agent-Organization" and "Place" are designated as geo-entities;

(2) Using "entities' coordinates" files to filter the triplets whose head geo-entities and tail geo-entities both having geographic coordinates (the center points' longitude and latitude of geo-entities in DBpedia) information from the above results.

Finally, the GeoKG "GeoDBpedia" contains 44,819 geo-entities, 86 geo-relations, and 107,133 fact triples. The spatial distribution of geo-entities is shown in Figure 3 (blue points).

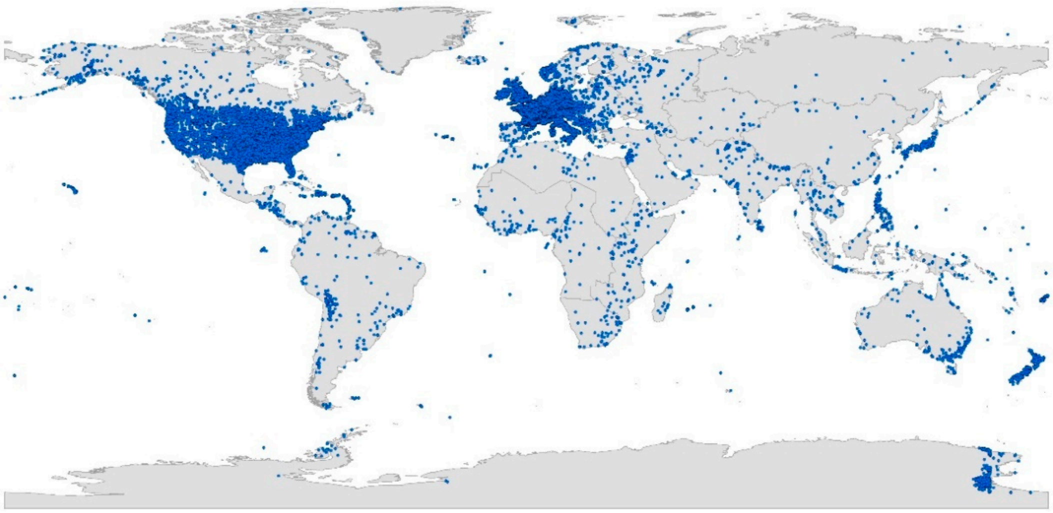


Figure 3. Spatial distribution of geo-entities in the dataset GeoDBpedia.

3.1.2. GeoKG from GADM

GADM, the Database of Global Administrative Areas, provides maps and spatial data for all countries and their sub-divisions (https://www.gadm.org/index.html). As a geographic domain database, the GADM contains abundant geo-entities (divisions). Although the explicit geo-relations of GADM are only the administrative relationships between geo-entities, these relationships are essential information for each geo-entity in this database. Thus, the GADM can be used to simulate an ideal link-dense GeoKG, while the GADM itself is not a KG dataset. We use the GADM version 3.6 and extract the data located in the range of France to generate a GeoKG dataset.

To enrich the types of geo-relations, we design nine geo-relations.

(1) Four types of geo-relations are extended from the administrative relationship: ispartof1, ispartof2, ispartof3, and ispartof4. An administrative relationship means a given geo-entity is a portion of a high-level geo-entity for the purpose of administration, so this relationship can be considered as a semantic relationship. The level of the given geo-entity may be 2, 3, 4, or 5 in GADM, and the level of its high-level geo-entity may be 1, 2, 3, or 4. Among above levels, level 1 is the top level, and level 5 is the lowest level. For example, <Ambronay, ispartof3, Belley > represents "the division 'Ambronay' is a part of the administrative level 3 division 'Belley'".

(2) Five types of geo-relations are generated from the adjacent geo-entities in the geographic space: adjoinwith1, adjoinwith2, adjoinwith3, adjoinwith4, and adjoinwith5. This relationship can be considered as a geographic relationship. These geo-relations mean the given geo-entity adjoins with geo-entity with an administrative level of 1, 2, 3, 4 or 5. The level of the given geo-entity may be 1, 2, 3, 4, or 5. For example, <Ambronay, adjoinwith 3, Bourg-en-Bresse > represents "the division 'Ambronay' adjoins with the level 3 division 'Bourg-en-Bresse'".

The details of generation are described in Appendix A. Finally, the GeoKG "GADM-KG-FRA" contains 40,799 geo-entities and 555,443 fact triples. The spatial distribution of geo-entities is shown in Figure 4 (blue points).

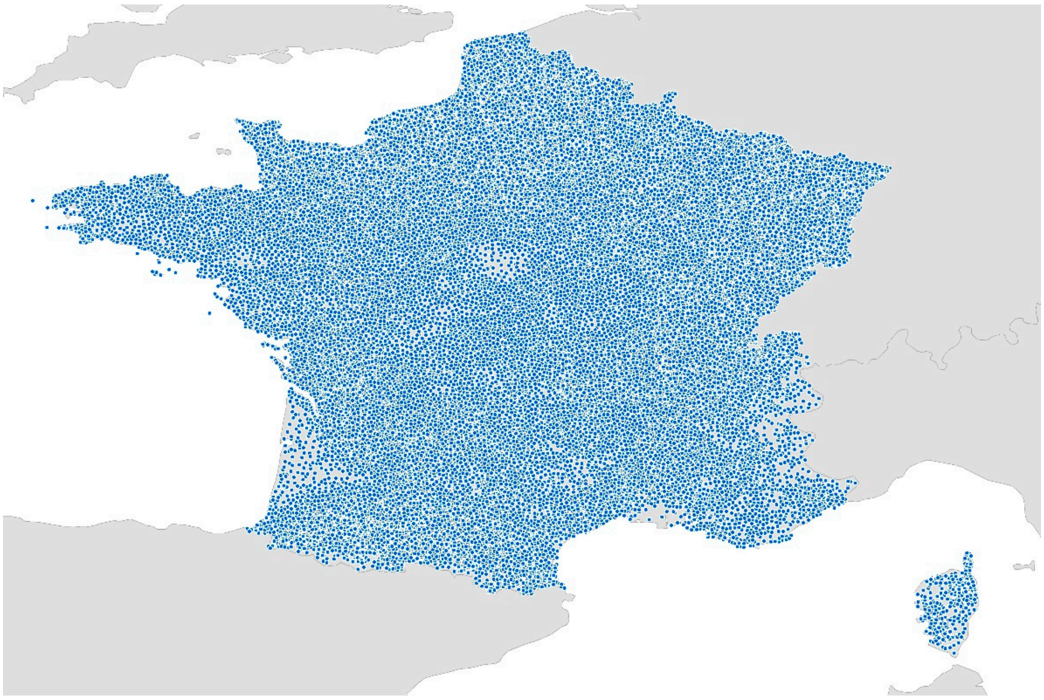


Figure 4. Spatial distribution of geo-entities in the dataset GADM-KG-FRA.

3.1.3. KG Node Degree

We use the node degree to explore the quantitative difference of our GeoKG datasets and general KG datasets on entity linking. The node degree is a measure indicating how many edges (relations) link with a node (entity) in graph theory [26]. A high degree means this entity connects to more other entities. WN18 and FB15K are two general KG datasets in many KG completion research publications. WN18 is a triplet dataset containing contains 40,943 entities and 18 types of relations, which is extracted from WordNet (a lexical knowledge base to support dictionary and thesaurus;). FB15K is a triplet dataset containing 14,951 entities and 1345 types of relations, which is extracted from Freebase (a fact knowledge base like DBpedia).

Figure 5 shows the cumulative frequency of degree 1 to 9 of each dataset. It is apparent from this figure that the percentage of entities with degree 1 in GeoDBpedia (70.14%) far exceeds GADM-KG-FRA (0%), WN18 (14.87%), and FB15K (2.18%). Moreover, the percentage of entities with degree ≤ 2 in GeoDBpedia is 86.52%, while this percentage for WN18 is just over half this amount. Consequently, GeoDBpedia, as a link-sparse GeoKG, cannot supply enough training data for current translation KE models to learn the entity's distinct representations of different relations. Thus, the performance of above translation KE models on the real GeoKG completion will be poor.

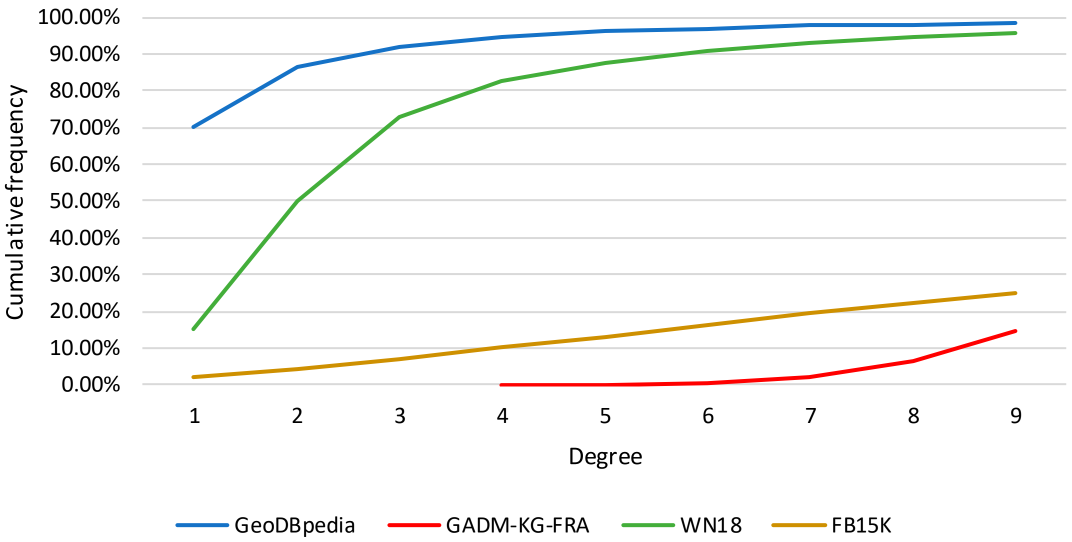


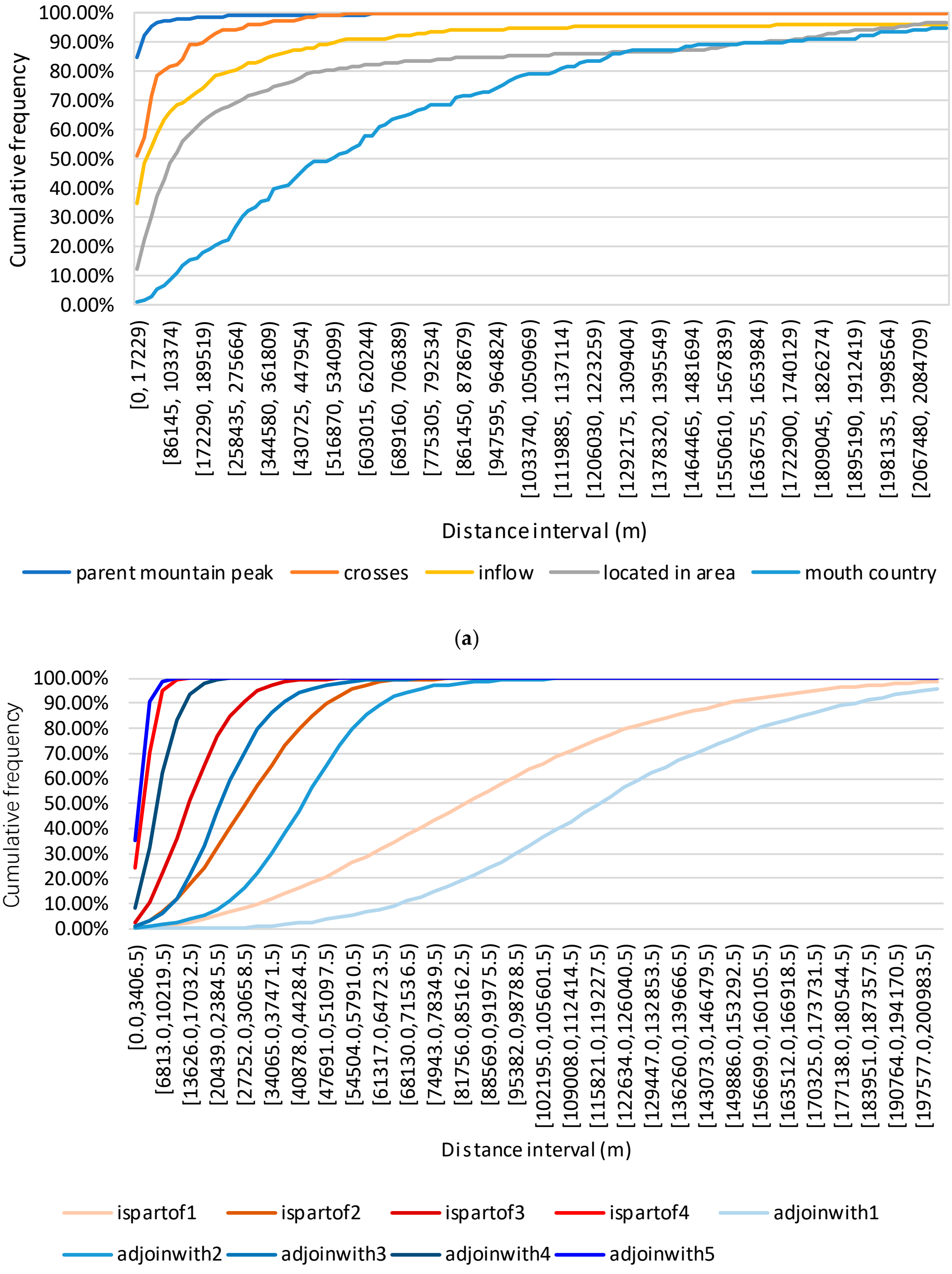
Figure 5. Cumulative frequency of entity degree 1 to 9 of different datasets.

3.2. *Translation KE with Geospatial Distance Restriction*

3.2.1. Geospatial Distance Restriction Hiding in GeoKG

The GeoKG supports abundant geospatial information, which will become the crucial additional information to optimize the representations of geographic knowledge. Each geo-entity of GeoKG must refer to a precise or vague geospatial location. Then, the symbolic geo-relations between two geo-entities can be regarded as the reflection of the geospatial distance of these two geo-entities. That is, the geo-relations which can be used in describing two geo-entities are restricted by their geospatial distance.

We assume that each geo-relation can only be used to describe the geo-entities which are located in a certain distance range. Figure 6 shows the cumulative frequency distributions of the Euclidean distances between the head geo-entity and tail geo-entity of fact triplets. Figure 6a is the result of five geo-relations' fact triplets in GeoDBpedia, and Figure 6b is that of nine geo-relations' fact triplets in GADM-KG-FRA. Here, the distance between two geo-entities is the geospatial distance between their center points. It is apparent that there are significant differences in the distance distributions of different geo-relations' fact triplets. As shown in Figure 6a, the number of "parent mountain peak" geo-relation triplets with a short distance is obviously larger than that of "mouth country" geo-relation triplets with a short distance. This means that if we have two geo-entities with a short distance, the most likely geo-relation between them is "parent mountain peak", otherwise the geo-relation is "mouth country". Inspired by above statistical results, we will introduce the geospatial distance restriction into a translation KE model.



(**a**)

(**b**)

Figure 6. Cumulative frequency distributions of the distances between the head geo-entity and tail geo-entity of fact triplets: (**a**) GeoDBpedia (parent mountain peak: a peak's parent as a particular peak in the higher terrain connected to the peak; crosses: where the bridge crosses a river; inflow: a source of the water in the body of water; located in area: where the entity is located in a place; mouth country: where the body of water flows into a country); (**b**) GADM-KG-FRA.

3.2.2. Model Optimization with Geospatial Distance Restriction

As mentioned in Section 2, the objective functions of translation KE models aim to widen the gap between positive triplets and negative triplets. Then, the training results will assign positive triplets the high score and negative triplets the low score. Thus, our method introduces geospatial distance restriction to translation KE model by modifying the objective function.

Firstly, we add a geospatial weight to the objective function to ease this gap if the distance between two geo-entities in negative triplets is unreasonable. The geospatial weight is defined as:

$$w_{geo} = \frac{1}{\left|\log_{10} \frac{dis(\boldsymbol{h},\boldsymbol{t})+\theta}{dis(\boldsymbol{h}',\boldsymbol{t}')+\theta}\right| + 1} \tag{3}$$

where $dis(\boldsymbol{h}, \boldsymbol{t})$ is the distance between the head geo-entity and tail geo-entity. θ is a compensation term to avoid having a denominator equal to zero. For simplicity, the locations of head geo-entity and tail geo-entity are both abstracted as points. Then the $dis(\boldsymbol{h}, \boldsymbol{t})$ is measured as:

$$dis(\boldsymbol{h}, \boldsymbol{t}) = \sqrt{(h_x - t_x)^2 + \left(h_y - t_y\right)^2} \tag{4}$$

where h_x and h_y are the longitude and latitude of the head geo-entity; t_x and t_y are the longitude and latitude of the tail geo-entity.

Next, the objective function becomes:

$$L = \sum_{h.t} \sum_{h',t'} \max\Big(0, f_r(\boldsymbol{h}, \boldsymbol{t}) + \gamma - w_{geo} f_r(\boldsymbol{h}', \boldsymbol{t}')\Big) \tag{5}$$

Specifically, the effects of two above depicted functions are: if the distance between $\boldsymbol{h}'$ and $\boldsymbol{t}'$ of a negative triplet is greater or less than the distance between $\boldsymbol{h}$ and $\boldsymbol{t}$ of the positive triplet, the w_{geo} will be less than 1, then the final score $w_{geo} f_r(\boldsymbol{h}', \boldsymbol{t}')$ of the negative triplet is turned lower. Thus, the model has to generate a lower $f_r(\boldsymbol{h}, \boldsymbol{t})$ or higher $f_r(\boldsymbol{h}', \boldsymbol{t}')$ to achieve the effects as before. A lower $f_r(\boldsymbol{h}, \boldsymbol{t})$ means that $\boldsymbol{h}$ and $\boldsymbol{t}$ need to be closer to each other in the vector space. A higher $f_r(\boldsymbol{h}', \boldsymbol{t}')$ will increase the distance between $\boldsymbol{h}'$ and $\boldsymbol{t}'$ in the vector space.

Most objective functions of translation KE models are the same as in Equation (2), so the proposed method has the ability to optimize all these translation KE models. Next, TransR and TransD will be optimized with geospatial distance restriction in the same way.

TransR [7] builds entity and relation embedding in separate entity spaces and relation spaces, to reflect the phenomenon that an entity may have multiple semantics and various relations which focus on the different semantics of entities. Its scoring function is:

$$f_r(\boldsymbol{h}, \boldsymbol{t}) = -\|\boldsymbol{M}_r \boldsymbol{l}_h + \boldsymbol{l}_r - \boldsymbol{M}_r \boldsymbol{l}_t\|_2^2 \tag{6}$$

where $\boldsymbol{M}_r$ is a projection matrix from the entity space to the relation space of $\boldsymbol{r}$.

TransD [8] uses two vectors to represent an entity (a relation) to further recognize the different meanings between the head entity and tail entity in one triplet. Its scoring function is:

$$f_r(\boldsymbol{h}, \boldsymbol{t}) = -\left\|\left(w_r w_h + I\right)\boldsymbol{l}_h + \boldsymbol{l}_r - \left(w_r w_t + I\right)\boldsymbol{l}_t\right\|_2^2 \tag{7}$$

where, w_h, w_t and w_r are the mapping vectors for the representations of $\boldsymbol{h}$, $\boldsymbol{t}$ and $\boldsymbol{r}$, and I is an identity matrix. Then, the vectors of $\boldsymbol{h}$ and $\boldsymbol{t}$ are projected by $\left(w_r w_h + I\right)$ and $\left(w_r w_t + I\right)$.

The objective functions of TransR and TransD both are the same as TransE, so the proposed geospatial weight can also be introduced by the same objective function as Equation (5) into TransR and TransD.

From the original models TransE, TransR and TransD, the optimized models by the proposed method are named as TransE-GDR, TransR-GDR and TransD-GDR. We implement the translation KE models based on the open-source package Fast-TransX (https://github.com/thunlp/Fast-TransX). This package includes TransE, TransR and TransD code. Then, TransE-GDR, TransR-GDR and TransD-GDR are obtained by modifying the code.

3.3. GeoKG Completion

After obtaining the trained translation KE models, the missing entities or relations of facts will be predicted with vector operations.

3.3.1. Entity Prediction and Relation Prediction

Given the vectors $\boldsymbol{l}_{\mathrm{h}}$ and $\boldsymbol{l}_{\mathrm{r}}$ of the head geo-entity **h** and the geo-relation **r**, the missing tail geo-entity **?** of the triplet <**h**, **r**, **?**> can be predicted by the operation $\boldsymbol{l}_{\mathrm{h}} + \boldsymbol{l}_{\mathrm{r}}$. The detailed procedure is as follows.

(1) the missing tail geo-entity "?" of the triplet <**h**, **r**, **?**> is replaced by all vectors of geo-entities in the known GeoKG. Then, the candidate triplets of triplet <**h**, **r**, **?**> are generated, whose number equals to the number of all geo-entities in GeoKG.

(2) the score of each candidate triplet <**h**, **r**, **x**> is calculated by a trained translation KE model. For example, if the translation KE model is TransE-GDR, the score of candidate triplet will obtained Equation (1).

(3) these candidate triplets can be ranked by scores in ascending order. The tail geo-entity **x** of the first candidate triplet becomes the missing tail entity of the triplet <**h**, **r**, **?**>, or the tail geo-entities of the top-n candidate triplets become the candidate missing tail geo-entities for subsequent inference with external information.

Similarly, the missing head geo-entity of triplet <**?**, **r**, **t**> or the missing geo-relation of triplet <**h**, **?**, **t**> can also be predicted through the above steps.

3.3.2. Location Prediction

Because the geo-entities of GeoKG have geospatial coordinates, the result of entity prediction can be used to predict the location of an unknown geo-entity. If the geospatial distance restriction is successfully embedded into the semantic vector space, the geo-entity which is closer to the correct geo-entity in the geographic space will be predicted at higher ranking positions on entity prediction. So, even the entity prediction does not give the correct geo-entity, the location of the missing geo-entity can be generated from the geospatial coordinates of the predicted candidate geo-entities: the coordinate of the first predicted geo-entity become the possible location of the missing geo-entity; or a polygon constructed by the coordinates of the top-n predicted geo-entities becomes an area which the missing geo-entity with high probability is located in.

4. Experiments and Results

Two kinds of tasks are used to evaluate the performance of various models in GeoKG completion: link prediction and location prediction. The former is a common task, and the latter is a novel task to explore the models' capacities in predicting the locations of unknown entities. First, we introduced the experimental datasets extracted from the above two GeoKGs. Next, we presented the metrics and results of the two evaluation tasks. Finally, we analyze these results.

4.1. Experimental Datasets

We generate the experimental dataset from GeoDBpedia and GADM-KG-FRA. Because some types of geo-relations in GeoDBpedia only have few fact triplets, which will influence the training effect, the geo-relations with a number of fact triplets exceeding 100 are selected from GeoDBpedia.

In addition, the triplets whose relations (properties) are sensitive to geospatial distance are reserved manually. As illustrated above, the geo-relation "parent mountain peak" may be applied to describe two mountains if these two geo-entities are close to each other in geospatial terms. The geo-relation "twinTown" is used to describe a legal or social agreement between towns, and is not related to their geospatial distance. This dataset is named as "GeoDBpedia21" and its types of geo-relations are listed in Table 1.

Table 1. Types of geo-relations (DBpedia properties) in the experiment.

Type	Explanation
department	which department the place belongs to (Department is one of the three levels of government in France)
located in area	where the entity is located in a place
source country	where the river originated from in a country
nearest city	the entities' nearest city in geospatial terms
mountain range	which mountain range the mountain belongs to (A mountain range is a series of mountains ranged in a line and connected by high ground)
mouth mountain	where the body of water flows into a mountain
mouth place	where the body of water flows into a place
parent mountain peak	a peak's parent as a particular peak in the higher terrain connected to the peak
outflow	a sink of the body of water
inflow	a source of the body of water
broadcast area	a place served by a radio station
river mouth	where the river flows into another river, a lake, a reservoir, a sea, or an ocean
river	a river located in or meets at the place
location city	where the organization is located in a city
mouth region	where the body of water flows into a region
crosses	where the bridge crosses a river
major island	which small major islands the island has
mouth country	where the body of water flows into a country
island	an island belongs to or contains the place
right tributary	a stream or river that flows into its right larger stream or main stem (or parent) river or a lake
left tributary	a stream or river that flows into its left larger stream or main stem (or parent) river or a lake

All fact triplets of GADM-KG-FRA are used as the experimental dataset, which is named as "GADM9" for simplicity.

Above two experimental datasets are divided into the training set, validation set and test set as 8:1:1. If the triplets' head geo-entities or tail geo-entities do not appear in the training set, these triplets will be removed from the validation set and test set. Table 2 gives the statistics of final GeoDBpedia21 and GADM9.

Table 2. Statistics of experimental datasets.

Dataset	#Relation	#Entity	#Training Set	#Validation Set	#Test Set
GeoDBpedia21	21	39,770	46,657	2560	2544
GADM9	9	40,799	444,345	55,549	55,549

4.2. Link Prediction and Results

The purpose of link prediction is to evaluate the model's performance on entity prediction and relation prediction. Link prediction is based on the results of KG completion as Section 3.3, where the translation KE model is trained by the fact triplets of training set. For entity prediction, the head entities (tail entities) are removed from the fact triplets of validation set or test set. The valid or test fact triplets become the incomplete fact triplets. After generating the ranked candidate triplets of each incomplete fact triplet through the steps of Section 3.3, the ranks of the correct head entities (geo-entities) are obtained. Two metrics are reported: (i) *MeanRank*, the mean of the correct entity rank, and (ii) *Hits@10*, the proportion of correct entities ranked in the top 10. The mean of the results of head entity prediction and tail entity prediction is the result of entity prediction.

Relation prediction is similar to entity prediction; the relations in valid triplets or test triplets are removed and replaced with all relations in the training set. The metrics include *MeanRank* and *Hits@1* (the proportion of correct relations ranked in the top 1).

Note that, for some facts of a GeoKG, the correct geo-entities or geo-relations may not be unique. To avoid mistaking corrupted triplets as errors in the valid phase or test phase, these triplets will be removed from the training set, validation set, and test set before ranking. Thus, the above metrics can be further divided into the (*Raw*) part and (*Filter*) part.

Learning rate λ, margin γ, and vector space's dimension k are three important parameters of translation KE models. The learning rate λ is a parameter of SGD algorithm, which is used to control the rate of gradient descent for learning the representations of entities and relations. The margin γ is a margin hyperparameter of the margin-based ranking loss mentioned in Section 2. The dimension k is the dimension of the embedding vector space. First, we select λ for SGD among {0.1, 0.01, 0.001, 0.0001, 0.00001}, γ among {0.5, 1, 1.5, 2, 4, 6, 8, 10}, and k among {50, 100, 150, 200}. Next, the best configurations of GeoDBpedia21 and GADM9 are determined according to the metrics *MeanRank* and *Hits@10* of entity prediction with the validation sets. The optimal configurations of each translation KE model used in the test set are shown in Table 3.

Table 3. Parameter configurations of each model in the test sets.

Parameter	GeoDBpedia21			GADM9		
	λ	γ	k	λ	γ	k
TransE	0.001	10	100	0.0001	1.5	100
TransR	0.001	1	100	0.0001	1	100
TransD	0.001	1	100	0.0001	0.5	100
TransE-GDR	0.001	1.5	100	0.0001	0.5	100
TransR-GDR	0.0001	0.5	100	0.0001	1.5	100
TransD-GDR	0.001	1	100	0.00001	1	100

Tables 4 and 5 display the results on entity prediction and relation prediction. It can be seen that the optimized models (TransE-GDR, TransR-GDR and TransD-GDR) both outperform their originals (TransE, TransR and TransD) on GeoDBpedia21: *MeanRank(Filter)* is reduced by an average of 429.60, and the *Hits@10(Filter)* is improved by an average of 6.41% on entity prediction; *MeanRank(Filter)* is reduced by an average of 2.84, and the *Hits@10(Filter)* is improved by an average of 46.56% on relation prediction. As a contrast, the difference between the results of each model on GADM9 is not obvious.

Table 4. Experimental results on entity prediction.

Metric	**GeoDBpedia21**				**GADM9**			
	MeanRank		**Hits@10**		**MeanRank**		**Hits@10**	
	Raw	**Filter**	**Raw**	**Filter**	**Raw**	**Filter**	**Raw**	**Filter**
TransE	896.83	824.43	46.52%	56.98%	196.05	29.94	64.17%	80.44%
TransE-GDR	615.47	533.38	48.13%	60.36%	191.42	24.98	64.03%	81.12%
TransR	1084.86	995.50	45.17%	52.56%	191.37	22.42	62.70%	81.55%
TransR-GDR	611.86	527.38	48.21%	61.01%	166.04	**12.61**	60.34%	81.21%
TransD	1088.48	999.22	47.35%	54.62%	201.19	30.82	**66.78%**	82.28%
TransD-GDR	**556.07**	**469.58**	**49.35%**	**62.01%**	**159.34**	30.23	64.27%	**82.41%**

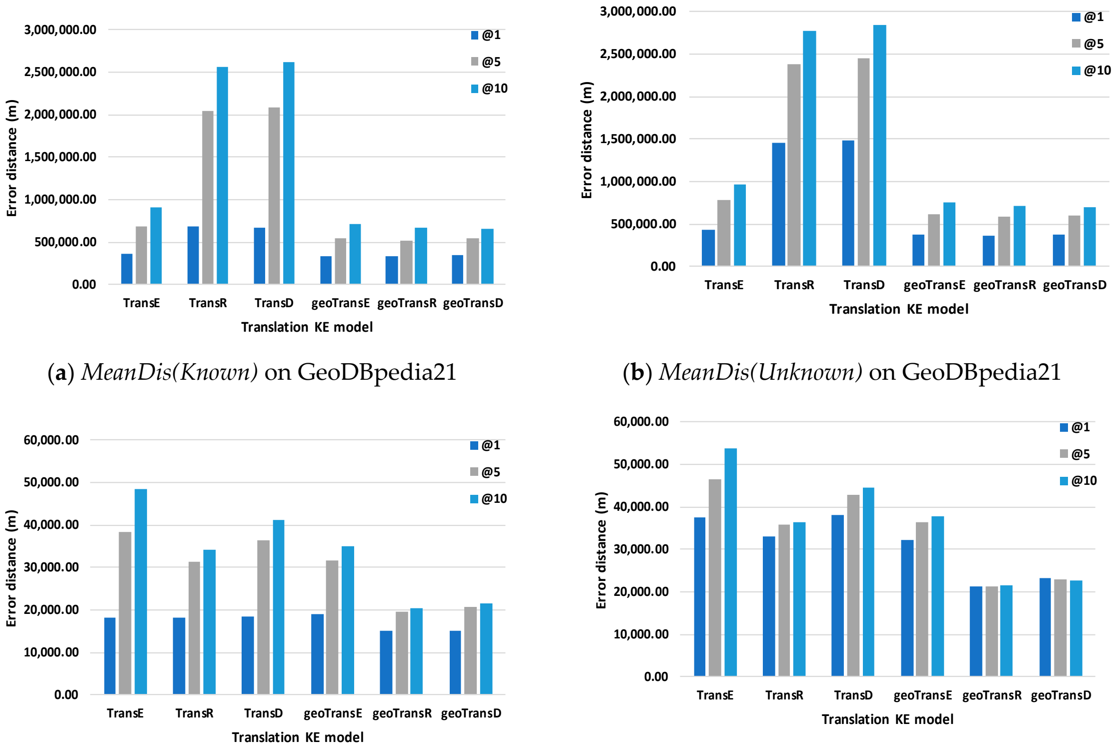
Table 5. Experimental results on relation prediction.

Metric	**GeoDBpedia21**				**GADM9**			
	MeanRank		**Hits@1**		**MeanRank**		**Hits@1**	
	Raw	**Filter**	**Raw**	**Filter**	**Raw**	**Filter**	**Raw**	**Filter**
TransE	1.56	1.41	75.20%	87.19%	1.47	1.34	72.93%	83.38%
TransE-GDR	**1.47**	**1.33**	**77.04%**	89.82%	1.53	1.41	**74.31%**	83.96%
TransR	4.79	4.65	33.96%	37.74%	1.54	1.41	73.22%	83.19%
TransR-GDR	1.48	**1.33**	76.65%	89.78%	1.30	1.17	72.67%	85.19%
TransD	3.84	3.70	45.52%	48.82%	1.48	1.35	**74.31%**	85.37%
TransD-GDR	1.49	1.34	76.42%	**89.90%**	**1.28**	**1.15**	73.26%	**85.85%**

4.3. Location Prediction and Results

Location prediction experiment is designed based on the results of entity prediction (*Filter*). The means of error distances between the correct geo-entity and the top 1, top 5 or top 10 predicted geo-entities will be reported, and labeled as *MeanDis@1*, *MeanDis@5*, and *MeanDis@10*. To simulate the application scenario of inferring the locations of unknown geo-entities which are not indexed by gazetteers (training set), the correct geo-entities will be excepted from the prediction results to calculate the metric *MeanDis*. Thus, the above metrics can be further divided into the (*Known*) part and (*Unknown*) part.

Figure 7 shows the error distance results of location prediction on GeoDBpedia21 and GADM9. Compared with their original models (TransE, TransR and TransD), the predicting error distances of all optimized models (TransE-GDR, TransR-GDR and TransD-GDR) decline on each metric. Concretely, on GeoDBpedia21, the average reducing rates of three (*Know*) metrics are −31.06%, −49.87% and −51.14%; of three (*Unknown*) metrics are −43.81%, −50.65% and −51.56%. On GADM9, the average reducing rates of three (*Know*) metrics are −10.34%, −32.82% and −38.52%; of three (*Unknown*) metrics are −29.56%, −36.40% and −39.93%.



(**a**) *MeanDis(Known)* on GeoDBpedia21

(**b**) *MeanDis(Unknown)* on GeoDBpedia21

(**c**) *MeanDis(Known)* on GADM9

(**d**) *MeanDis(Unknown)* on GADM9

Figure 7. The performance of location prediction.

4.4. Result Analysis

Comparing the results of GeoDBpedia21 and GADM9 (Tables 4 and 5), it can be seen that the difference between the results of each model on GADM9 is not more obvious than those obtained on GeoDBpedia21. The reason for this is that a link-dense GeoKG provides enough links of geo-entities, which can facilitate models to encode the geo-entities and geo-relations into the right positions in a low-dimensional vector space by multiple links. Thus, the link prediction results of the geographic distance restriction on GADM9 do not show a substantial improvement. Besides, it is apparent that TransR and TransD behave significantly worse than TransE on the link-sparse GeoDBpedia21. Meanwhile, the performance of TransR and TransD is close or better to that of TransE on the link-dense GADM9. Theoretically, TransR and TransD as the extension models of TransE, should achieve better prediction results than TransE [27]. Therefore, the above difference between two datasets indicates that the sparseness of links between geo-entities severely limits the performances of TransR and TransD on current GeoKG. Meanwhile, the presented method alleviates the influence of sparseness on geographic knowledge representation by capturing the geospatial distance restriction hiding in the GeoKG. Thus, TransR-GDR and TransD-GDR achieve performances that are close to or better than TransE-GDR on GeoDBpedia21.

The performance of location prediction is mainly affected by the entity prediction results on GeoDBpedia21. Because the entity prediction results of original models are significantly worse than their optimized models, the location prediction results of the original models are poor. As shown in Figure 7c,d, while the entity prediction results of all models are close to each other on GADM9, the optimized models show the better performance on location prediction than that of the original models. It means that more geo-entities near the correct location in a geographic space are predicted at higher ranking positions in entity prediction using our method.

In summary, the optimized models (TransE-GDR, TransR-GDR and TransD-GDR) perform well on both link-dense GeoKG (GADM9) and link-sparse GeoKG (GeoDBpedia21). Thus, these optimized models are the better options for GeoKG completion, especially when the sparsity of GeoKG is unknown. Among these optimized models, although TransR-GDR and TransD-GDR achieve similar performance on the above tasks, which are both better than that of TransE-GDR, TransD-GDR can be the preferred model in GeoKG completion due to TransR-GDR requiring a longer training time.

Next, we explore the models' performance of geo-entity prediction by different geo-relation types, and the impact of training set scale on geo-entity prediction effect. Then, we give a specific case to illustrate the results of geo-entity prediction.

4.4.1. Geo-Entity Prediction by Different Geo-Relation Types

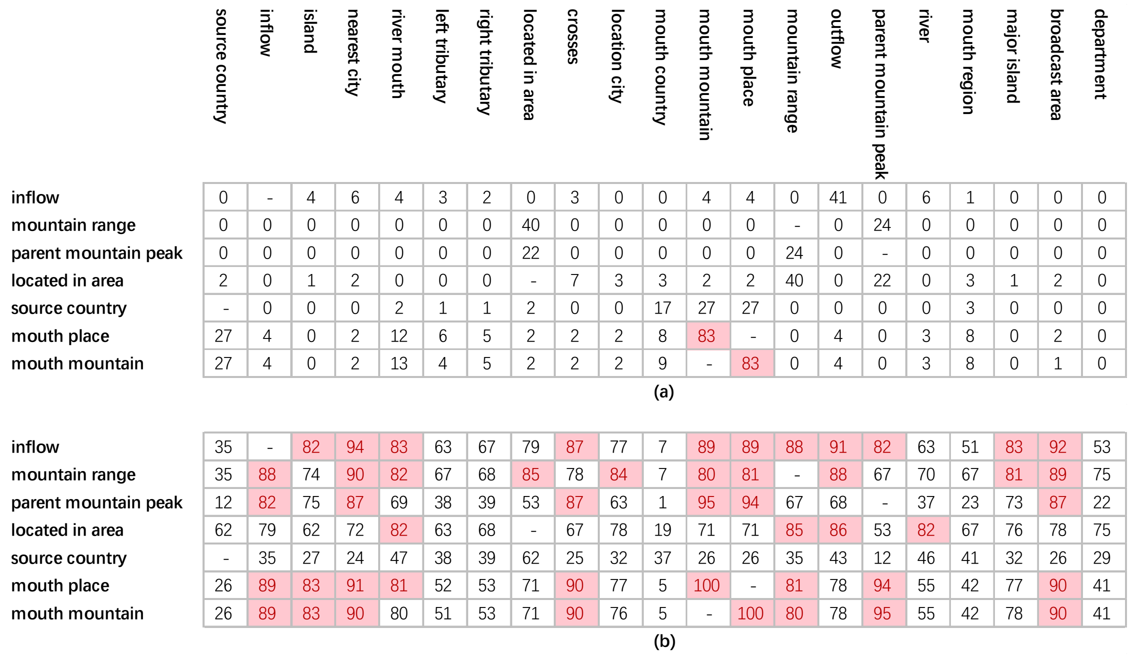
We compare the models' performance of entity prediction on GeoDBpedia21 by different geo-relation types. The results are shown in Table 6. Note that, the result of geo-relation "river" is not reported, because that the head geo-entities or tail geo-entities of triplets with geo-relation "river" in test set do not appear in the training set, and all "river" triplets are removed from test set finally. The geo-relations which contain more than 50 test triplets are selected for further analysis to ensure that the data is representative. These geo-relations are "inflow", "mountain range", "parent mountain peak", "located in area", "source country", "mouth place" and "mouth mountain". As mentioned before, for translation KE models, the semantic difference of relations is distinguished by mining the graph links of KGs. And the proposed method uses the geographical distance restriction from GeoKG to enable the models to distinguish the semantic difference better. Thus, if a geo-relation has (a) special semantic reflecting by the graph links of GeoKGs, and (b) more candidate geo-entities which located in the certain distance range, this geo-relation may achieve better prediction performance. Then geo-entity similarity and distance distribution similarity are designed to explore the geo-relations' difference on geo-entity prediction:

1. The geo-entity similarity is a similarity of geo-relations based on whether the geo-entities become the head or tail geo-entities of the geo-relations' fact triplets. A geo-relation can be represented as a vector: dimension is the number of head geo-entities plus that of tail geo-entities in the training set, and the value is 1 when a geo-entity as the head (tail) geo-entity of the geo-relation triplet. The geo-entity similarity between two geo-relations is higher, the geo-entities of these geo-relations' fact triplets are more consistent, namely the semantics of these geo-relations are more similar.
2. The distance distribution similarity is a similarity of geo-relations based on the frequency distributions of the distances between the head geo-entities and tail geo-entities of fact triplets. A geo-relation can be represented as a vector: dimension is the number of distance range, and value is the number of geo-entities in a distance range. Then the distance distribution similarity between two geo-relations is higher, the distance distributions of these geo-relations' geo-entities are more similar.

Figure 8 shows the two similarity results of above 7 geo-relations. It can be seen that (1) the geo-relations ("inflow" and "mountain range"), whose geo-entity prediction is significantly improved by the proposed method, have more other geo-relations with similar distance distributions (similarity ≥ 0.8) and no other geo-relations with same geo-entities (similarity ≥ 0.8). (2) the geo-relations ("parent mountain peak", "located in area" and "source country"), whose geo-entity prediction slightly improved by the proposed method, have less other geo-relations with similar distance distribution and no other geo-relations with same geo-entities. (3) the geo-relations ("mouth place" and "mouth mountain"), which achieve worse prediction performances, have less other geo-relations with similar distance distributions but one other geo-relations with same geo-entities. To summarize, if a geo-relation can obtain more candidate geo-entities which located in its certain distance range from GeoKG, but these candidate geo-entities are different from the known geo-entities of this geo-relation's fact triplets, the prediction performance of this geo-relation will be better.

Table 6. Entity prediction results on GeoDBpedia21 by geo-relation types.

Geo-Relation	#Test Set	Hits@10 (Raw)						Improvement			
		TransE	TransE-GDR	TransR	TransR-GDR	TransD	TransD-GDR	TransE-GDR	TransR-GDR	TransD-GDR	Mean
location city	5	40.00%	40.00%	20.00%	40.00%	20.00%	40.00%	0.00%	20.00%	20.00%	13.33%
mouth region	8	6.25%	12.50%	0.00%	12.50%	0.00%	18.75%	6.25%	12.50%	18.75%	12.50%
river mouth	35	48.57%	45.71%	30.00%	47.14%	31.43%	44.29%	−2.86%	17.14%	12.86%	9.05%
inflow	59	68.64%	66.95%	53.39%	68.64%	55.93%	69.49%	−1.69%	15.25%	13.56%	9.04%
crosses	8	31.25%	37.50%	25.00%	37.50%	25.00%	31.25%	6.25%	12.50%	6.25%	8.33%
mountain range	267	46.44%	49.25%	40.07%	48.69%	41.39%	49.44%	2.81%	8.61%	8.05%	6.49%
nearest city	49	22.45%	24.49%	14.29%	23.47%	15.31%	23.47%	2.04%	9.18%	8.16%	6.46%
outflow	46	77.17%	79.35%	72.83%	80.43%	70.65%	78.26%	2.17%	7.61%	7.61%	5.80%
right tributary	13	65.38%	65.38%	57.69%	65.38%	65.38%	69.23%	0.00%	7.69%	3.85%	3.85%
mouth country	18	50.00%	50.00%	44.44%	50.00%	47.22%	52.78%	0.00%	5.56%	5.56%	3.70%
parent mountain peak	131	12.60%	20.99%	17.56%	20.23%	18.32%	16.41%	8.40%	2.67%	−1.91%	3.05%
broadcast area	20	15.00%	20.00%	20.00%	20.00%	17.50%	20.00%	5.00%	0.00%	2.50%	2.50%
located in area	1119	36.06%	36.10%	34.09%	37.71%	38.11%	39.10%	0.04%	3.62%	0.98%	1.55%
source country	211	36.73%	38.39%	37.44%	37.44%	38.86%	40.28%	1.66%	0.00%	1.42%	1.03%
mouth place	268	79.29%	81.53%	81.72%	79.85%	81.34%	81.90%	2.24%	−1.87%	0.56%	0.31%
major island	4	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	0.00%	0.00%	0.00%	0.00%
department	3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
mouth mountain	265	78.87%	82.83%	83.96%	80.75%	84.53%	83.40%	3.96%	−3.21%	−1.13%	−0.13%
left tributary	12	75.00%	75.00%	75.00%	70.83%	79.17%	70.83%	0.00%	−4.17%	−8.33%	−4.17%
island	3	0.00%	0.00%	33.33%	0.00%	16.67%	16.67%	0.00%	−33.33%	0.00%	−11.11%



	source country	inflow	island	nearest city	river mouth	left tributary	right tributary	located in area	crosses	location city	mouth country	mouth mountain	mouth place	mountain range	outflow	parent mountain peak	river	mouth region	major island	broadcast area	department
inflow	0	-	4	6	4	3	2	0	3	0	0	4	4	0	41	0	6	1	0	0	0
mountain range	0	0	0	0	0	0	0	40	0	0	0	0	0	-	0	24	0	0	0	0	0
parent mountain peak	0	0	0	0	0	0	0	22	0	0	0	0	0	24	0	-	0	0	0	0	0
located in area	2	0	1	2	0	0	0	-	7	3	3	2	2	40	0	22	0	3	1	2	0
source country	-	0	0	0	2	1	1	2	0	0	17	27	27	0	0	0	0	3	0	0	0
mouth place	27	4	0	2	12	6	5	2	2	2	8	83	-	0	4	0	3	8	0	2	0
mouth mountain	27	4	0	2	13	4	5	2	2	2	9	-	83	0	4	0	3	8	0	1	0

(a)

	source country	inflow	island	nearest city	river mouth	left tributary	right tributary	located in area	crosses	location city	mouth country	mouth mountain	mouth place	mountain range	outflow	parent mountain peak	river	mouth region	major island	broadcast area	department
inflow	35	-	82	94	83	63	67	79	87	77	7	89	89	88	91	82	63	51	83	92	53
mountain range	35	88	74	90	82	67	68	85	78	84	7	80	81	-	88	67	70	67	81	89	75
parent mountain peak	12	82	75	87	69	38	39	53	87	63	1	95	94	67	68	-	37	23	73	87	22
located in area	62	79	62	72	82	63	68	-	67	78	19	71	71	85	86	53	82	67	76	78	75
source country	-	35	27	24	47	38	39	62	25	32	37	26	26	35	43	12	46	41	32	26	29
mouth place	26	89	83	91	81	52	53	71	90	77	5	100	-	81	78	94	55	42	77	90	41
mouth mountain	26	89	83	90	80	51	53	71	90	76	5	-	100	80	78	95	55	42	78	90	41

(b)

Figure 8. (**a**) Geo-entity similarity of geo-relations (%); (**b**) distance distribution similarity of geo-relations (%).

4.4.2. Impact of Training Set Scale on Geo-Entity Prediction

We extract sub-datasets from GeoDBpedia21 by proportion 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% to analyze the impact of training set scale on geo-entity prediction performance. The triplet number of training set of each sub-dataset is 13,990, 18,654, 23,326, 27,987, 32,651, 37,317, 41,984 and 46,657. In order to ensure the comparability of the results, the validation sets and test sets of each sub-data set are the same: the triplet number of validation set and test set is 810 and 785. Figure 9 shows the geo-entity prediction results of the model TransD and TransD-GDR on each sub-data set. The orange line and values represent the improvements of the performance of TransD-GDR compared with that of TransD. It can be seen that, as the training set scale increases, the overall trend of *Hits@10(Filter)* decreases. To further analyze the reason for this result, we calculate the node degrees of geo-entities of the test set in different training sets of sub-data sets. Figure 10 shows the cumulative frequency of degree 1 to 9 of each sub-dataset. It is apparent that as the training set scale increases, the node degree of geo-entities also increases. More concretely, the geo-entities will connect more other geo-entities in a training set with a large number of triplets. Then the original model TransD cannot obtain enough training triplets to refine the representation of geo-entities and geo-relations, so the improvement of TransD-GDR decreases. Thus, the impact of the training set scale on the performance of the proposed method depends on whether the training set supplies enough triplets to increase the connections between geo-entities or not.

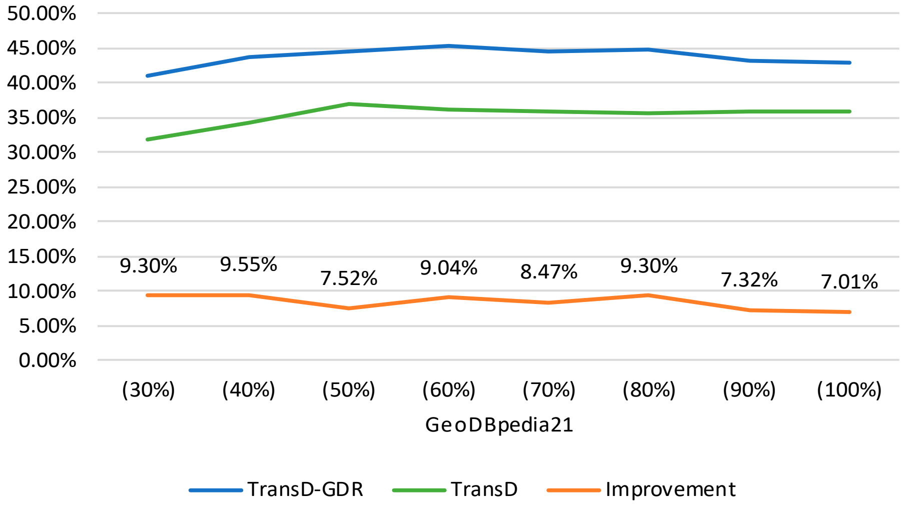


Figure 9. The performance of geo-entity prediction on different sub-datasets.

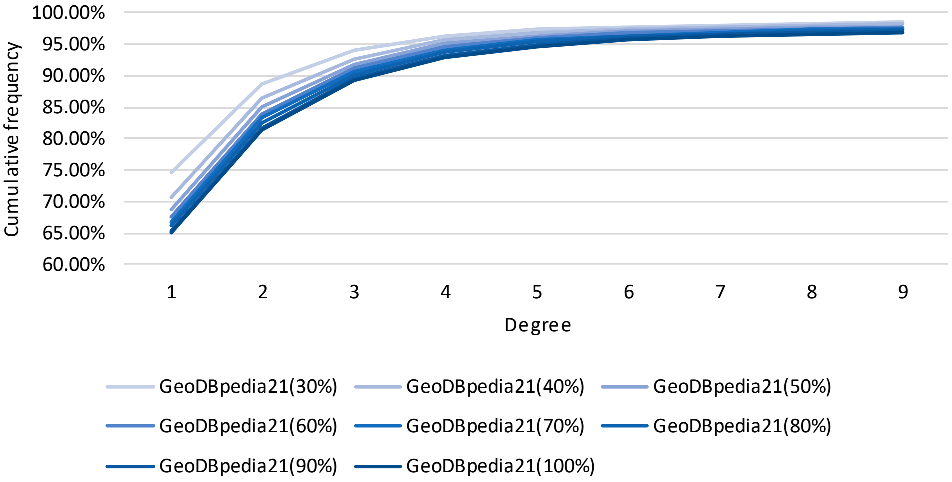


Figure 10. Cumulative frequency of entity degree 1 to 9 of sub-datasets.

4.4.3. Geo-Entity Prediction Case

We use a specific case to demonstrate the detail results of entity prediction: "which mountain range the Monte Acero belongs to". This fact is not supported by the DBpedia entity "Monte Acero" (http://dbpedia.org/page/Monte_Acero). The above presented case is converted into an incomplete triplet in the form of DBpedia:

<http://dbpedia.org/resource/Monte_Acero, http://mappings.dbpedia.org/server/ontology/classes/Mountain, ?> (abbreviated as <Monte Acero, mountain range, ?> for simplicity)

The missing tail geo-entity **?** is predicted through the steps of Section 3.3, where TransD and TransD-GDR trained on GeoDBpedia21 are selected in this case.

Table 7 shows the top 10 predicted tail geo-entities of the case based on TransD and TransD-GDR. The **Apennine Mountains** is proved to be the correct tail geo-entity of <Monte Acero, mountain range, "?"> through a special website (https://www.mountain-forecast.com/peaks/Monte-Acero/forecasts/736). The results indicate that the correct geo-entity predicted by TransD-GDR has a higher rank than that generated by TransD. Thus, compared with the original translation KE model, the presented method enhances the chances of correctly predicting the missing geo-entities.

Table 7. Tail geo-entity prediction results of the incomplete triplet <Monte Acero, mountain range, ?>.

	TransD	TransD-GDR
1	http://dbpedia.org/resource/Gennargentu	http://dbpedia.org/resource/Apennine_Mountains
2	http://dbpedia.org/resource/Pala_group	http://dbpedia.org/resource/Rieserferner_Group
3	http://dbpedia.org/resource/Apennine_Mountains	http://dbpedia.org/resource/Pala_group
4	http://dbpedia.org/resource/Rieserferner_Group	http://dbpedia.org/resource/Zillertal_Alps
5	http://dbpedia.org/resource/Dolomites	http://dbpedia.org/resource/Gennargentu
6	http://dbpedia.org/resource/Ligurian_Alps	http://dbpedia.org/resource/Dolomites
7	http://dbpedia.org/resource/Bernina_Range	http://dbpedia.org/resource/Maritime_Alps
8	http://dbpedia.org/resource/Zillertal_Alps	http://dbpedia.org/resource/Ligurian_Alps
9	http://dbpedia.org/resource/Maritime_Alps	http://dbpedia.org/resource/Campania
10	http://dbpedia.org/resource/Pennine_Alps	http://dbpedia.org/resource/Province_of_Benevento

5. Discussion

Current translation KE models assume that the entity should have distinct representations for different relations to improve the rationality of knowledge embedding. However, because that the geo-entities in GeoKG only connect a few other geo-entities, and one type of geo-relation contains a limited number of fact triplets, these current models will achieve poor performance on this link-sparse GeoKG. We address the shortcomings of such models by introducing the geospatial distance restriction hiding in the GeoKG. Through the proposed method, the translation KE method has the ability to capture and embed the geospatial distance restriction with the semantic information of GeoKG into a vector space. In this vector space, the geo-entities with same distance in a geographic space will maintain a similar distance from each other. Then, the optimized model outputs the refined representations of geo-entities and geo-relations, which improves the completion performance on the link-sparse GeoKG. Concretely, on the one hand, the geospatial distance restriction drives the geo-entities near the correct location to be given higher ranking positions in entity prediction. Then, the correct geo-entity as one of geo-entities near the correct location also appears higher in the ranking, which increases the possibility of predicting the missing entity correctly. On the other hand, the geospatial distance restriction is merged into the semantic vector spaces of the optimized models, which facilitates relation prediction to find the best geo-relation for two geo-entities based on their geospatial distance. Therefore, the presented method alleviates the influence of link-sparseness on geographic knowledge representation by capturing the geospatial distance restriction hiding in the GeoKG.

However, there are still some issues that require further investigation.

(1) In this study, the geospatial distance between two geo-entities is simplified to the distance of these geo-entities' center points, which leads to that the geospatial distance between two geo-entities with some relation types, such as the adjacency of two administrative divisions, or the cross of a river and a bridge, is not equal to zero. While this simplicity is not completely rigorous from the geographic perspective, it is still reasonable. For example, although the geospatial distance between two geo-entities with adjacency relation is not equal to zero under this simplicity, this geospatial distance is not unlimited. If the adjacency relation is of two administrative divisions, their geospatial distance will not exceed the maximum length of all divisions. Thus, the geospatial distance is still a valuable restriction for the above relation types. Because high quality geometric data of different types of point geo-entities, linear geo-entities and area geo-entities is scattered across different datasets yet (like GADM only contains the polygons of administrative divisions), we will attempt to introduce the geospatial distance restriction, topology restriction and orientation restriction of the geo-entities with difference geometric types into translation KE models after fusing current GeoKGs and geographical databases based on aligning technology [28,29] in subsequent research.

(2) The translation KE model is a data-driven model for representing geo-entities and geo-relations of GeoKGs. On one hand, the model understands the semantic difference of geo-relations by mining the graph links of the detailed GeoKGs but not by the conceptual level of geo-relation. So, if the fact triplets of two geo-relations have the same geo-entities in the training set, the proposed method

will treat these two geo-relations as the same geo-relations. This is because the semantic and the geographical distance restrictions of these two geo-relations are the same, as geo-relation "mouth place" and "mouth mountain" stated in Section 4.4.1. On the other hand, the error distance of location prediction is influenced by the distribution of geo-entities in GeoKGs. For example, the average error distance of location prediction is 369,725.24 m (*MeanDis@1(Unknown)*) with GeoDBpedia21 because the geospatial information of the GeoDBpedia21 is scarce at a global scale, so the geospatial distance between geo-entities is large. Thus, aligning different GeoKGs and geographical databases to supply more fact triplets is a feasible way for the translation KE model to further distinguish the semantic of geo-relations and reduce the error distance of location prediction. Besides, merging the similar graph links of geo-relations before model training can also improve the prediction performance of models.

(3) Translation KE model is a tool to take full advantage of known geo-entities and geo-relations that exist in the GeoKG to complete the GeoKG. However, this method cannot complete the missing geo-entities or geo-relations which have never appeared in a GeoKG. Meanwhile, information extraction is an effective way to extract the missing unknown facts from external structured, semi-structured or unstructured texts [30,31]. Thus, integration of different completion results may be worth researching as a way to improve the quality of GeoKGs completion.

6. Conclusions

In this paper, a method with geospatial distance restriction is presented to optimize knowledge embedding for supplying the missing geographic facts of a link-sparse GeoKG. Specifically, by adding the geospatial distance restriction to the objective function of translation KE models, these models output the optimized representations of geo-entities and geo-relations of a link-sparse GeoKG. Then, the missing geo-entities or geo-relations can be predicted by a vector operation according to both the semantic relationship and the geospatial distance feature between geo-entities in the GeoKG.

The effects of the presented method are validated on a real GeoKG extracted from DBpedia. Compared with the results of the original models, the presented method improves the metric *Hits@10(Filter)* by an average of 6.41% for geo-entity prediction, and the *Hits@1(Filter)* by an average of 31.92% for geo-relation prediction. Furthermore, a set of novel experiments explored the models' capacities for predicting the locations of unknown geo-entities. The results show that the geospatial restriction reduced the average geospatial distance error of prediction by between 54.43% and 57.24%. All results indicate that the presented method successfully captures the geospatial distance restriction hiding in the GeoKG to refine the representations of geo-entities and geo-relations in the link-sparse GeoKG. In addition, more experimental results indicate that the completion performance of the proposed method is influenced by the graph structure of GeoKG: the proposed method will improve the completion performance better on (1) the GeoKGs whose links of geo-entities are more sparse, and (2) the geo-relation with more candidate geo-entities differed from the known geo-entities of this geo-relation's fact triplets within the distance restriction.

Optimizing the embedding representations of geographic knowledge in the GeoKG is the next step: (1) adding the connections between geo-entities and non-geo-entities to methods for learning the difference between geo-entities; and (2) fusing multiple geospatial restriction, which include not only distance relations but also topological relations and orientation relations. Additionally, introducing these embedding representations of geographic knowledge into other GeoKG applications, such as geographic knowledge extraction and geographic QA, is a promising direction for further research.

Author Contributions: Conceptualization, P.Q. and F.L.; Methodology, P.Q.; Validation, P.Q., L.Y. and J.G.; Formal Analysis, P.Q.; Investigation, P.Q. and J.G.; Resources, P.Q. and L.Y.; Data Curation, P.Q. and L.Y.; Writing-Original Draft Preparation, P.Q.; Writing-Review & Editing, F.L., L.Y. and J.G.; Supervision, F.L.; Project Administration, F.L.; Funding Acquisition, F.L.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No.41631177, Grant No.41801320). Their supports are gratefully acknowledged. We also thank the anonymous referees for their helpful comments and suggestions.

Acknowledgments: The authors would like to thank the four anonymous reviewers for their valuable suggestions, which significantly improve the quality of this article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This appendix gives the details of generating dataset GADM-KG-FRA.

A geo-entity (division) is recorded in GADM as (some fields are removed here for simplicity):

GID_0	NAME_0	GID_1	NAME_1	GID_2	NAME_2	GID_3	NAME_3	GID_4	NAME_4	GID_5	NAME_5
FRA	France	FRA.1_1	Auvergne-Rhône-Alpes	FRA.1.1_1	Ain	FRA.1.1.1_1	Belley	FRA.1.1.1.1_1	Ambérieu-en-Bugey	FRA.1.1.1.1.2_1	Ambronay

This record refers to the geo-entity "Ambronay", which is an administrative level 5 division. The upper level divisions of "Ambronay" are "Ambérieu-en-Bugey" (level 4), "Belley" (level 3), "Ain" (level 2), "Auvergne-Rhône-Alpes" (level 1), and "France" (level 0). GID_x is the preferred unique ID at level x.

The processing flow of triplet extraction as follows.

(1) Based on the administrative information described in records, four types of geo-relation triplets can be extracted: ispartof1, ispartof2, ispartof3, and ispartof4. For example, the triplets obtained from the above record "Ambronay" are:

<Ambronay, ispartof4, Ambérieu-en-Bugey>;
<Ambronay, ispartof3, Belley>;
<Ambronay, ispartof2, Ain>;
<Ambronay, ispartof1, Auvergne-Rhône-Alpes>.

If a record is about a geo-entity with administrative level 4. only three types of geo-relation triplets can be extracted: ispartof1, ispartof2, ispartof3, and so on. The number of extracted fact triplets is 158,428.

(2) Using ArcGIS tool "Generate Near Table" to detect whether geo-entity A (level X) adjoins to geo-entity B (level Y). If these two geo-entities are adjacent, the five types of geo-relation triplets can be generated: adjoinwith1, adjoinwith2, adjoinwith3, adjoinwith4 and adjoinwith5, as <geo-entity A, adjoinwithY, geo-entity B>.

(3) In the GDAM, the spatial information of geo-entities (divisions) is geospatial polygons. Thus, the polygon's center point is extracted to be the geospatial location of a division.

References

1. Introducing the Knowledge Graph: Things, Not Strings. Available online: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html (accessed on 22 July 2018).
2. Uyar, A.; Aliyu, F.M. Evaluating search features of Google Knowledge Graph and Bing Satori: Entity types, list searches and query interfaces. *Online Inf. Rev.* **2015**, *39*, 197–213. [CrossRef]
3. Xiong, C.; Power, R.; Callan, J. Explicit Semantic Ranking for academic search via knowledge graph embedding. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2017; pp. 1271–1279.
4. Dong, L.; Wei, F.; Zhou, M.; Xu, K. Question answering over freebase with multi-column convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 260–269.
5. Chi, Y.; Qin, Y.; Song, R.; Xu, H. Knowledge graph in smart education: A case study of entrepreneurship scientific publication management. *Sustainability* **2018**, *10*, 995. [CrossRef]
6. Wang, Q.; Wang, B.; Guo, L. Knowledge base completion using embeddings and rules. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, 25–31 July 2015.

7. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15), Austin, TX, USA, 25–30 January 2015; AAAI Press: Menlo Park, CA, USA, 2015; pp. 2181–2187.
8. Ji, G.; He, S.; Xu, L.; Liu, K.; Zhao, J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 687–696.
9. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Dutchess County, NY, USA, 2013; pp. 2787–2795.
10. Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
11. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the Workshop at International Conference on Learning Representations 2013, Scottsdale, AZ, USA, 2–4 May 2013; pp. 1–12.
12. Liu, K.; Gao, S.; Qiu, P.; Liu, X.; Yan, B.; Lu, F. Road2Vec: Measuring traffic interactions in urban road system from massive travel routes. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 321. [CrossRef]
13. Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2724–2743. [CrossRef]
14. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **1985**, *22*, 400–407. [CrossRef]
15. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14), Québec City, QC, Canada, 27–31 July 2014; AAAI Press: Menlo Park, CA, USA, 2014; pp. 1112–1119.
16. Xiao, H.; Huang, M.; Zhu, X. TransG: A generative model for knowledge graph embedding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 2316–2325.
17. Xiao, H.; Huang, M.; Hao, Y.; Zhu, X. TransA: An adaptive approach for knowledge graph embedding. *arXiv* **2015**, arXiv:150905490.
18. Ji, G.; Liu, K.; He, S.; Zhao, J. Knowledge graph completion with adaptive sparse transfer matrix. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; AAAI Press: Menlo Park, CA, USA, 2016; pp. 985–991.
19. Xie, R.; Liu, Z.; Sun, M. Representation learning of knowledge graphs with hierarchical types. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16), New York, NY, USA, 9–15 July 2016; AAAI Press: Menlo Park, CA, USA, 2016; pp. 2965–2971.
20. Wu, Y.; Wang, Z. Knowledge graph embedding with numeric attributes of entities. In Proceedings of the 3rd Workshop on Representation Learning for NLP, Melbourne, Australia, 20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 132–136.
21. Xie, R.; Liu, Z.; Jia, J.; Luan, H.; Sun, M. Representation learning of knowledge graphs with entity descriptions. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16), Phoenix, AZ, USA, 12–17 February 2016; AAAI Press: Menlo Park, CA, USA, 2016; pp. 2659–2665.
22. Lin, Y.; Liu, Z.; Luan, H.; Sun, M.; Rao, S.; Liu, S. Modeling relation paths for representation learning of knowledge bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP' 15), Lisbon, Portugal, 17–21 September 2015; pp. 705–714.
23. Feng, J.; Huang, M.; Yang, Y.; Zhu, X. GAKE: Graph aware knowledge embedding. In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 641–651.
24. Shi, J.; Gao, H.; Qi, G.; Zhou, Z. Knowledge graph embedding with triple context. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; ACM: New York, NY, USA, 2017; pp. 2299–2302.

25. Jiang, T.; Liu, T.; Ge, T.; Sha, L.; Li, S.; Chang, B.; Sui, Z. Encoding temporal information for time-aware link prediction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 2350–2354.
26. Diestel, R. *Graph Theory*, 5th ed.; Springer Publishing Company: New York, NY, USA, 2018; ISBN 978-3-662-57560-4.
27. Liu, H.; Wu, Y.; Yang, Y. Analogical inference for multi-relational embeddings. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 11 August 2017; pp. 2168–2178.
28. Galárraga, L.A.; Teflioudi, C.; Hose, K.; Suchanek, F. AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; ACM: New York, NY, USA, 2013; pp. 413–422.
29. Yu, L.; Qiu, P.; Liu, X.; Lu, F.; Wan, B. A holistic approach to aligning geospatial data with multidimensional similarity measuring. *Int. J. Digit. Earth* **2018**, *11*, 845–862. [CrossRef]
30. Khan, A.; Vasardani, M.; Winter, S. Extracting spatial information from place descriptions. In Proceedings of the First ACM SIGSPATIAL International Workshop on Computational Models of Place, Orlando FL, USA, 5–8 November 2013; ACM: New York, NY, USA, 2013; pp. 62:62–62:69.
31. West, R.; Gabrilovich, E.; Murphy, K.; Sun, S.; Gupta, R.; Lin, D. Knowledge base completion via search-based question answering. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14), Seoul, Korea, 7–11 April 2014; ACM: New York, NY, USA, 2014; pp. 515–526.