Detecting geo-relation phrases from web texts for triplet extraction of geographic knowledge: a context-enhanced method

Peiyuan Qiu\textsuperscript{a}, Li Yu\textsuperscript{b}, Jialiang Gao\textsuperscript{a,c} and Feng Lu\textsuperscript{a,c,d,e}

\textsuperscript{a}State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; \textsuperscript{b}National Science Library, Chinese Academy of Sciences, Beijing, China; \textsuperscript{c}University of Chinese Academy of Sciences, Beijing, China; \textsuperscript{d}The Academy of Digital China, Fuzhou University, Fuzhou, China; \textsuperscript{e}Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China

\textbf{ABSTRACT}

As an effective organization form of geographic information, a geographic knowledge graph (GeoKG) facilitates numerous geography-related analyses and services. The completeness of triplets regarding geographic knowledge determines the quality of GeoKG, thus drawing considerable attention in the related domains. Mass unstructured geographic knowledge scattered in web texts has been regarded as a potential source for enriching the triplets in GeoKGs. The crux of triplet extraction from web texts lies in the detection of key phrases indicating the correct geo-relation between geo-entities. However, the current methods for key-phrase detection are ineffective because the sparseness of the terms in the web texts describing geo-relations results in an insufficient training corpus. In this study, an unsupervised context-enhanced method is proposed to detect geo-relation key phrases from web texts for extracting triplets. External semantic knowledge is introduced to relieve the influence of the sparseness of the geo-relation description terms in web texts. Specifically, the contexts of geo-entities are fused with category semantic knowledge and word semantic knowledge. Subsequently, an enhanced corpus is generated using frequency-based statistics. Finally, the geo-relation key phrases are detected from the enhanced contexts using the statistical lexical features from the enhanced corpus. Experiments are conducted with real web texts. In comparison with the well-known frequency-based methods, the proposed method improves the precision of detecting the key phrases of the geo-relation description by approximately 20\%. Moreover, compared with the well-defined geo-relation properties in DBpedia, the proposed method provides quintuple key-phrases for indicating the geo-relations between geo-entities, which facilitate the generation of new triplets from web texts.

\textbf{ARTICLE HISTORY}

Received 31 May 2019
Accepted 7 August 2019

\textbf{KEYWORDS}

Geographic knowledge graph; triplet extraction; geo-entity relation; key-phrase detection; context enhancement

\section*{1. Introduction}

A knowledge graph (KG) is a system that organizes entities (people, places, and things) and their relations in the form of graphs. The semantic difference of entities and relations is
expressed through the graph structure. This difference is more significant than that stored in a traditional knowledge base, where the semantic differences of entities are only reflected as the independent property values of the entities. Thus, a KG can facilitate special domains, such as medicine, education, and information science. As a special KG, a geographic KG (GeoKG) is an effective organization form of geographic information, especially extracted from the web texts in newswires, collaborative encyclopedias, social media, official or domain websites, etc. Profiting from the rich semantics of geographic entities (geo-entities) and geographic relations (geo-relations), GeoKGs play an important role in geographic question-answering (Mai, Yan, Janowicz, & Zhu, 2019) and geographic data sharing (Zhu et al., 2017a, 2017b).

Current general KGs, such as DBpedia¹ and Freebase², have maintained a large number of triplets regarding geographic knowledge. In these KGs, the geographic entities (geo-entities) and geographic relations (geo-relations) are represented as triplets in the form “(head entity, relation, tail entity)” or “(entity, property, value)” (both abbreviated as “(h, r, t)”). For example, the knowledge “The end point of Royal Canal is Dublin” is represented as (Royal Canal, end point, Dublin). Unfortunately, these triplets cannot satisfy the requirements of GeoKG applications. First, these triplets are still far from complete. For example, in English DBpedia, 53.86% of lake entities are missing their source of water (property “inflow”), and 85.80% of mountain entities have no fact describing the location of their parent peaks (property “parent mountain peak”). Second, most of these triplets express information in the form of “(entity, property, value)”. Thus, these triplets cannot provide geo-relation knowledge between geo-entities to make the links of the GeoKG dense, which is important to distinguish the semantic difference between geo-entities in applications. Third, the update frequency of these triplets does not adapt to the appearance of new geographic knowledge. Furthermore, vast unstructured geographic knowledge is scattered in web texts, which is a potential source of enriching triplets for GeoKGs.

The important step in triplet extraction from web texts is detecting the correct geo-relation key phrase. Key phrases are terms picked out from the context as indicators in the expressions of entities and relations, which provide important clues that describe the relations between entities (Chen, Ji, Tan, & Niu, 2005; Zhang, Sun, & Zhang, 2012). Knowledge engineering and supervised learning methods are common methods of key-phrase detection. However, these methods are ineffective for detecting geo-relation key phrases from web texts. First, as there are no open massive detection patterns and geographic annotated corpora, manual pattern generation and corpora annotation are indispensable pre-processing tasks. These pre-processing tasks determine the performance of the methods but are time-consuming. Second, the strong heterogeneity of web texts across various domains results in the poor portability of the detection patterns and trained models (Li, Goodchild, & Raskin, 2014). Third, and above all, the pre-defined patterns and pre-trained models can hardly capture the unexpected key-phrases of new types of geo-relations, which indicate the new geographic knowledge generated by the web constantly (Loglisci, lenco, Roche, Teisseire, & Malerba, 2012).

Unsupervised learning methods work independent of the data distribution and capture text features in real time with statistical techniques (Hasegawa, Sekine, & Grishman, 2004). The capacity of the unsupervised learning methods for exploring new relations makes them exclusively advantageous for mining unknown triplets for the continuous construction of a GeoKG. An unsupervised learning method usually regards key-phrase detection as
a frequency ranking task, i.e. identifying the top-ranked terms in a document with frequency statistics. These methods hypothesize that there are a large number of redundant terms indicating the relations for a specific entity pair. The frequency of terms uniquely distinguishes one term from the others. However, this hypothesis is inappropriate for key-phrase detection involving the description of geo-relations owing to the inherent characteristic of sparse distribution. First, the specific geo-entity pair rarely co-occurs in a sentence. The number of terms in the context of the specific geo-entity pair is also very limited. Second, the same meaning can be represented by different terms, exacerbating the problem of sparseness (Stefanakis, 2003). Thus, it is difficult to distinguish the key phrases for the sparse geo-entity relation descriptions using only frequency statistics. Thus, the performance of the current unsupervised learning methods for triplet extraction from web texts is not ideal.

This study focuses on the detection of geo-relation key phrases with sparse distribution in web texts to extract triplets. Two relevant problems are discussed: (1) How to relieve the influence of the sparseness of terms in context? Through introducing external semantic knowledge, the contexts of geo-entity pairs with the same type are merged into the enhanced contexts to increase the term frequency and mitigate the influence of the sparseness. (2) How to increase the distinctions between terms in context? A large-scale enhanced corpus is constructed automatically from the enhanced corpus using two frequency-based methods. Subsequently, the lexical features and their weights are statistically determined based on the enhanced corpus, and the features describing the terms from multiple perspectives are used to detect key phrases from the enhanced contexts.

The remainder of this paper is organized as follows. The related works are introduced in Section 2. A context-enhanced methodology for the detection of geo-relation key phrases is presented in Section 3. Section 4 describes the experiments and presents a discussion. Section 5 concludes this work.

2. Related works

Knowledge engineering and machine learning methods are the two most common methods of key-phrase detection in current studies.

Knowledge engineering manually builds geo-relation dictionaries and uses them to design relation patterns or construct feature templates for model training to extract triplets from web texts. Smole, Čeh, and Podobnikar (2011) manually defined 26 geo-relations between geo-entities to train a machine learning model. This method focuses on the five most frequent semantic geo-relations (is-a, is-located, has-purpose, is-result-of, has-parts) and manually annotates 1,308 definitions of geo-entities. Elia, Guglielmo, Maisto, and Pelosi (2013) retrieved spatial relations between multiple named entities by building grammars and a dictionary associated with 234 spatial verbs in Italian. Cao, Wang, and Jiang (2014) similarly used a handcrafted spatial dictionary and designed 493 geo-relation patterns for matching web pages. It is evident that manually gathering key phrases for describing geo-relations is not cost-effective. Moreover, it will inevitably lead to an incomplete coverage of geo-relations. To break the constraint of key-phrase dictionaries, Khan, Vasardani, and Winter (2013) detected spatial relations between places from degenerate locative expressions using prepositional phrases or prepositional clauses. However, their method cannot recognize semantic geo-relations described in texts without prepositions,
such as <Northern California, largest metropolitan area, San Francisco Bay Area> in the sentence "Northern California's largest metropolitan area is the San Francisco Bay Area."

As a statistical learning method, a machine learning method avoids the manual works of gathering a dictionary, patterns, or rules based on the corpus, and has the capacity of recognizing new geo-relations. Xiong, Mao, Duan, and Miao (2017) built a corpus of 37,431 characters to train deep neural networks for the detection of Chinese geo-relations. However, a massive annotated corpus for training a supervised learning model is rare and the annotation process requires a large amount of labor. To reduce the labor and time costs of annotation, distant supervision is introduced to extract a training corpus automatically based on the current KG for geo-relation detection (Jin, Zhao, & Wu, 2019; Mirrezaei, Martins, & Cruz, 2016). This method assumes that any sentence that contains a pair of entities existing in a known relation triplet of the KG is likely to express that relationship in some way (Mintz, Bills, Snow, & Jurafsky, 2009). However, distant supervision will bring a large amount of noise annotated corpus, which affects the performance of supervised learning models. Moreover, unsupervised learning methods use frequency statistics such as the term frequency and inverse document frequency (TFIDF) to detect geo-relation key phrases without the help of domain knowledge. These methods take the advantage of the redundancy of massive texts to weight each term and choose the top-ranked ones as key phrases. Considering the limitation that popular key-phrases for relation description would have relatively low IDF values, Mesquita (2012) used a new weight that accounts for the relatively discriminative power of a term within a given type of entity pair. Shen, Liu, and Huang (2012) determined term weights using the linear combinations of TFIDF and child concepts voting to improve the accuracy of key-phrase detection. Entropy is another popular frequency-based key-phrase detection method, which assumes that a term is irrelevant if its presence obscures the separability of the dataset (Dash, Choi, Scheuermann, & Liu, 2002). Therefore, a larger entropy of a term corresponds to higher importance. Chen et al. (2005) assessed the importance of all the terms in documents using the entropy criterion and selected a subset of important terms as the key phrases. Considering that processing all the terms would result in immense extraneous and incoherent information, Yan, Okazaki, Matsuo, Yang, and Ishizuka (2009) only dealt with verbs and nouns in documents. Similarly, Zhang et al. (2012) proved that the entropy method is effective for the detection of entity relations. However, frequency statistics techniques assume that relational terms will be frequently mentioned in large-scale corpora, which is not the case for geo-entity relations with sparse distribution.

3. Methodology

The objective of this study is to detect key phrases for describing geo-relations from web texts. These key phrases are strongly associated with the spatial or semantic relations between geo-entities and facilitate the triplet extraction for the construction of a GeoKG. The formal definitions of this problem are provided in Section 3.1. The workflow is shown in Figure 1.

First, the crawled web texts are pre-processed, including sentence splitting, segmentation, part of speech (POS) tagging, and geo-entity recognition. Subsequently, the original contexts of geo-entity pairs are created. Second, the external ontology knowledge and semantic knowledge are introduced to enhance the contexts. Third, an enhanced corpus is generated from the enhanced contexts using the frequency-based statistical methods.
Finally, the terms that are the key phrases of each geo-entity pair are identified from the enhanced contexts with the statistical lexical features based on the enhanced corpus. The processes in the right frame of Figure 1 are described in detail below.

### 3.1. Problem definition

**Input:** Web texts crawled from websites. A sentence of the texts is shown below.

The Park Güell is a public park system composed of gardens and architectonic elements located on Carmel Hill, in Barcelona, Catalonia (Spain).

**Output:** a set of key phrases for geo-entity pairs.

This study only focuses on geo-entity pairs co-occurring in a sentence. Considering the above sentence as an example, the concepts used in this paper are defined as follows.

**Geo-entity pair** $p_{1,2} = (e_1, e_2)$: two geo-related entities co-occurring in a sentence. The first geo-entity appearing in a sentence is paired with other geo-entities in the same sentence. For example, (Park Güell, Carmel Hill), (Park Güell, Barcelona), and (Park Güell, Catalonia (Spain)) are geo-entity pairs in the example.

**Geo-entity relation** $r$: a state of connection between geo-entities, including spatial and semantic relations. A spatial relation consists of topological, directional, and distance relations, such as “adjacent to,” “south of,” and “10 kilometers away.” A semantic relation is used to describe the logical relationship or dependency relationship between geo-entities, such as “type of,” “part of,” and “equal to.” Both these types of relations can

![Figure 1. Workflow of key-phrase detection for triplet extraction.](image-url)
be represented as a set of facts in the form \((e_1, r, e_2)\), such as (Park Güell, within, Carmel Hill), (Park Güell, within, Barcelona), and (Park Güell, within, Catalonia (Spain)).

**Term** \(t\): a phrase or a word with a definite meaning in a sentence except for entities such as “is,” “a,” “public,” “park system,” and “composed of.”

**Context** \(c\): all the terms existing before, between, and after the specified geo-entity pair in a sentence except for other geo-entities in the same sentence with the stop words filtered. For example, the contexts of (Park Güell, Carmel Hill), (Park Güell, Barcelona), and (Park Güell, Catalonia (Spain)) are “public, park, system, composed, of, gardens, architectonic, elements, located, on,” “public, park, system, composed, of, gardens, architectonic, elements, located, on, in,” and “public, park, system, composed, of, gardens, architectonic, elements, located, on, in,” respectively.

**Key-phrase** \(k\): the terms selected from the context as indicators in the relation expressions. For example, the term “located on” selected from the context (public, park, system, composed, of, gardens, architectonic, elements, located, on) is a key phrase revealing the topological relation “within” for the geo-entity pair (Park Güell, Carmel Hill).

Among the spatial relations, directional and distance relations are generally expressed by specific words and limited forms. For instance, texts containing the terms “east,” “west,” “south,” or “north” indicate the directional relations, and distance relations are often described in the form of “digit + measure unit.” Therefore, key phrases with directional and distance terms detected from texts can be directly taken as the directional or distance relations between geo-entities. However, topological relations usually show various expressions in texts similar to semantic relations. For instance, the topological relation “within” can be described by the terms “be surrounded by,” “be contained by,” or “located in.” Furthermore, the terms “be known as,” “be called,” and “alias” all indicate the semantic relation “equal to.” Thus, for follow-up GeoKG construction or GeoKG completion, the detected key phrases, which indicate topological and semantic relations, require semantic generalization to form abstract concepts as the geo-relations between geo-entities.

After pre-processing, the geo-entity pairs \(P\) and context \(C\) are obtained where \(p \in P\) denotes a geo-entity pair, and \(c \in C\) denotes the context of a specified geo-entity pair.

### 3.2. Sparseness reduction

The sparse distribution of terms that describe the geo-entity relations in web texts renders the frequency-based methods ineffective for key-phrase detection, making the correct triplet extraction difficult. Therefore, we first increase the frequency of terms in the partial texts, namely reduce the sparseness of the terms, to make the frequency-based methods perform well in the follow-up step. We adopted two strategies to reduce the sparseness with external semantic knowledge: (1) Merging the contexts of geo-entity pairs of the same type in different web texts based on external category semantic knowledge, and (2) fusing the terms having a similar semantic in the merged contexts based on external word semantic knowledge. The process of sparseness reduction for words is shown in Figure 2.

(1) Category knowledge in many KGs provides the well-defined type information of entities such as DBpedia Ontology, Wikipedia Categories, and the label system of Baidu Baike. Thus, the fine type \(type_x\) of each recognized geo-entity \(e_x\) in web texts can be assigned from the GeoKG categories after aligning the geo-entity with the KG. For example, the type of geo-entity can be assigned from DBpedia Ontology using the
Figure 2. Sparseness reduction for terms in contexts.

Figure 3. Example of categories regarding geo-entities in DBpedia ontology.

tool Spotlight³. Figure 3 displays a part of categories regarding the geo-entities in DBpedia Ontology.

(2) The type $T_m = (type_x, type_y)$ of geo-entity pair $p_{x,y} = (e_x, e_y)$ can be determined. Subsequently, the original contexts of geo-entity pairs of the same type $T_m$ are merged. Thus, the frequencies of some terms increase in the merged contexts.

(3) The external word semantic knowledge supported by word embedding technology is used to evaluate the similarity between words to fuse the terms having similar semantics.

The word embedding technology encodes the words in the corpus into a continuous low-dimensional semantic vector space, where each word is represented by a fixed-dimensional real-valued vector (Bengio, Ducharme, Vincent, & Janvin, 2003; Mikolov, Chen, Corrado, & Dean, 2013). If the distance between two words is close, such words have similar semantics or related semantics (Liu et al., 2017). 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. Thus, the word embedding technology can effectively measure the semantic similarity between words. In this study, we introduce the pre-trained word embedding via the bidirectional encoder
representations from transformers (BERT) model to measure the semantic similarity. Compared with the classic word embedding model Word2Vec, the BERT model encodes the word semantic better through fusing both the left and right contexts of the word in the corpus (Devlin, Chang, Lee, & Toutanova, 2019).

After acquiring the vectors of terms in the merged contexts, the similarity $\text{ConSim}(t_i, t_j)$ between two terms $t_i$ and $t_j$ is measured by calculating the cosine similarity of their vectors:

$$
\text{ConSim}(t_i, t_j) = \frac{\sum_{n=1}^{N} \text{vec}_{in} \cdot \text{vec}_{jn}}{\sqrt{\sum_{n=1}^{N} \text{vec}_{in}^2} \sqrt{\sum_{n=1}^{N} \text{vec}_{jn}^2}}
$$

where $\text{vec}_i$ and $\text{vec}_j$ are the vectors of $t_i$ and $t_j$, respectively. $\text{vec}_{in}$ and $\text{vec}_{jn}$ are the components of $\text{vec}_i$ and $\text{vec}_j$, respectively. $N$ is the dimension of the vector. If the similarity between two terms in a merging context is equal to or greater than 0.95, these two terms are fused into one term. Thus, the frequency of these terms, which are represented by the same term, further increases.

Finally, the enhanced contexts of each geo-entity pair of the same type are generated through the above steps, e.g. the enhanced context $c'_1$ of $T_1$ is generated from the original contexts $c_1$, $c_5$, etc.

### 3.3. Corpus generation

A large-scale corpus is required for effective lexical feature statistics, which is crucial for key-phrase detection (Naughton, Stokes, & Carthy, 2010). In this study, the corpus is automatically generated with two well-known frequency-based statistical methods: domain frequency (DF) and entropy. The DF method extends the classic TFIDF using the frequency of the terms in the context of the type-specific entity pairs, which favors specific relational terms as opposed to generic ones. The entropy method converts the context to a vector of terms and assesses the discrimination of each term based on the information theory, which provides useful heuristic information for key-phrase detection.

DF is a global measure of the discriminating power of a term for the type-specific pairs of geo-entities. It is defined in Equation (2).

$$
DF_t = \frac{f_{t,T_i}}{\sum_{j=1}^{N} f_{t,T_j}}
$$

Here, $f_{t,T_i}$ denotes the frequency of the term $t$ appearing in the contexts of geo-entity pairs of the type $T_i \in TS$. $TS$ is the type set of all the geo-entity pair types with a size of $N$.

Entropy assesses the importance of terms on text classification and is calculated using Equations (3) and (4). $S_{ij}$ denotes the similarity between the contexts $p_i$ and $p_j$, which is measured using the average distance $D$ of all the contexts and the distance $D_{ij}$ between
and $p_i$ and $p_j$ after removing the term $t$ from all the contexts. Equation (4) denotes the entropy of the term $t$ measured using $S_{ij}$.

$$S_{ij} = \exp\left(\frac{\ln 0.5}{D} \times D_{ij}\right)$$  \hspace{1cm} (3)

$$\text{Entropy}_t = \begin{cases} - \sum_{i=1}^{N} \sum_{j=1}^{N} (S_{ij} \log S_{ij} + (1 - S_{ij}) \log(1 - S_{ij})), 0 < S_{ij} < 1 \\ 0, \quad \text{otherwise} \end{cases}$$  \hspace{1cm} (4)

The DF and entropy methods are, respectively, used for detecting key phrases from the crawled web texts. The intersection of the two key-phrase sets forms the enhanced corpus for the term assessment.

### 3.4. Term assessment

After generating the corpus from the enhanced corpus, the terms, i.e. the key phrases of each geo-entity pair, are detected from the enhanced contexts according to the statistical lexical features of these terms in this enhanced corpus. Table 1 shows the lexical features used in this study, which are summarized from the existing literature (Blessing & Schutze, 2010; Pershina, Min, Xu, & Grishman, 2014; Zhang, Li, Hou, & Song, 2011).

The term assessment is performed considering the above lexical features as shown in Equation (5)–(8).

$$\text{wgt}^{(t)} = \theta_{\text{POS}} + \theta_{\text{LOC}} + \theta_{\text{DIS}}$$  \hspace{1cm} (5)

$$\theta_{\text{POS}} = p(t_{\text{POS}})$$  \hspace{1cm} (6)

$$\theta_{\text{LOC}} = \begin{cases} p(t_{\text{loc}} | \text{tp}(e_1)) \\ p(t_{\text{loc}} | \text{tn}(e_1)) \\ p(t_{\text{loc}} | \text{tp}(e_2)) \\ p(t_{\text{loc}} | \text{tn}(e_2)) \end{cases}$$  \hspace{1cm} (7)

| # | Lexical feature                                                                                   |
|---|--------------------------------------------------------------------------------------------------|
| 1 | POS of term (noun, verb, preposition, or others)                                                 |
| 2 | location of term (left of $e_1$, between $e_1$ and $e_2$, or right of $e_2$)                     |
| 3 | previous term just before $e_1$                                                                  |
| 4 | next term just after $e_1$                                                                      |
| 5 | previous term just before $e_2$                                                                  |
| 6 | next term just after $e_2$                                                                      |
| 7 | distance between the term and $e_1$                                                              |
| 8 | distance between the term and $e_2$                                                              |
| 9 | distance between the term and the head of the sentence                                           |
| 10| distance between the term and the tail of the sentence                                           |
In Equation (5), \(wgt_t\) denotes the weight of the term \(t\) for the specified geo-entity pair, considering the importance of the POS \(\theta_{\text{POS}}\), location \(\theta_{\text{LOC}}\), and distance \(\theta_{\text{DIS}}\).

Equation (6) denotes the weight of the POS, which is the probability of the event that the POS of \(t\), namely \(t_{\text{POS}}\), is equal to the specific POS.

Equation (7) denotes the weight of the relative location affected by the previous and next terms of the geo-entity. \(t_{\text{loc}}\) denotes the relative location of the term \(t\), which can be left, between, or right. \(t_{\text{loc}}(t_{\text{prev}})\) denotes the previous term of \(e_1\), and \(t_{\text{loc}}(t_{\text{next}})\) denotes the next term of \(e_1\). For example, \(p(t_{\text{loc}} = \text{between}|t_{\text{prev}}(e_1))\) denotes the probability that the term \(t\) located between \(e_1\) and \(e_2\) is the key phrase when the previous term of \(e_1\) is a specific term.

Equation (8) denotes the weight of the distance affected by the location of a term. \(dis(e_1)\) denotes the distance between \(t\) and \(e_1\). \(dis(e_2)\) denotes the distance between \(t\) and \(e_2\). \(dis(\text{head})\) denotes the distance between \(t\) and the head of the sentence. \(dis(\text{tail})\) denotes the distance between \(t\) and the tail of the sentence. For example, \(p(dis(e_1)|t_{\text{loc}} = \text{between})\) denotes the probability that the term \(t\) with a definite distance to \(e_1\) is the key phrase when \(t\) is located between \(e_1\) and \(e_2\).

All the terms in the contexts are assessed using Equation (5) and ranked in descending order. A local ordered list of terms is generated for each geo-entity pair, which indicates the decreasing importance of the terms for the expression of the geo-entity relation. The most important term is selected as the key phrase of the specified geo-entity pair.

4. Experiment & discussion

4.1. Dataset

All the articles referring to geo-entities are selected as the experimental corpus from the English Wikipedia. DBpedia is used to extract the geo-entities that belong to the “organization” (i.e., company, school, government agency, bank, etc.) and “place” (i.e., island, country, ocean, mountain, road, factory, hotel, etc.) categories of DBpedia Ontology. Finally, the size of the experimental corpus is 2.7 GB, and it contains 1,096,469 geo-entities and 26,018,455 sentences.

After pre-processing, the number of geo-entity pair types is 12,029. To ensure that there are sufficient original contexts to generate enhanced contexts, the geo-entity pair types whose number of original contexts is less than 100 are removed. Thus, the number of geo-entity pair types is 3,390 in this experiment.

4.2. Results

4.2.1. Key-phrase detection

The proposed method finally extracted 4,961,114 geographic facts from the experimental corpus. Examples of the detected key phrases referring to the geo-entity relation are shown in Table 2. The terms in the context of each geo-entity pair are arranged by the
decreasing order of their importance, and the term with the maximum weight is selected as the key phrase for these geo-entity pairs.

Summarizing the detected key phrases from all the geo-entity pairs, the number of geo-entity pair types having at least one effective key-phrase is 1,784 out of 3,390. Figure 4 illustrates the detected key phrases of ten cases of geo-entity pair types. These key phrases are ordered by their frequency in the detected geographic facts of each geo-entity pair type. Note that there are some nouns with regard to geo-entity types as the key phrases such as “peak,” “school,” and “airport.” These key phrases describe the relationship “$e_1$ is a geo-entity with the specified geo-entity type of $e_2$,” such as “The

| #  | $e_1$                          | $e_1$                        | Terms in context                                                                 | Key-phrase   |
|----|--------------------------------|-----------------------------|----------------------------------------------------------------------------------|--------------|
| 1  | Russian Empire                 | Asia                        | began, to, expand, into, from, 17th, century                                      | expand       |
| 2  | University of Alabama          | Tuscaloosa                  | largest, single, campus, located, in, with, enrolled, fall                       | located      |
| 3  | Afghan Defense University      | Qargha                      | $200$, million, under, construction, near, in                                   | near         |
| 4  | Albania                        | European Union              | became, full, member, in, 2009, and, has, applied, to, join                     | join         |
| 5  | Aberdeen Harbour              | Scotland                    | important, largest, in, north, of, and, ferry, route, to, and                    | north        |
| 6  | Broadway                       | Amsterdam Avenue            | from, to, intersect, with, known, south, of                                      | intersect    |
| 7  | Teide                         | Spain                       | with, its, 3,718, metres, 12,198, feet, highest, peak, of, and, also, world, heritage, site | peak         |
| 8  | Burley                         | Idaho                       | city, in, and, counties, in, southern, portion, of, state, of                    | southern     |
| 9  | AAMI Park                      | Cox Architects              | designed, by, with, input, from, Waratah, and, former, Wallaby, prop, Al, Baxter | designed     |
| 10 | Oakham                         | Turners Hill                | area, of, situated, on, peak, of, highest, point, in                            | situated     |
| 11 | Astro Aruna                    | Astro Prima                 | replaced, by, from, 1, October, 2012, on, mio, TV, Channel, 602                 | replaced     |
| 12 | Russian Empire                 | Asia                        | began, to, expand, into, from, 17th, century                                     | expand       |

| Road,Road | Town,River | Mountain,Mountain | Mountain,River | Country,Island |
|-----------|------------|-------------------|---------------|----------------|
| intersect | attain     | hill              | mountain      | archipelago    |
| begin     | tributary  | peak              | sit           | war            |
| north     | confluence | north             | creek         | ship           |
| split     | bridge     | south             | affect        | control        |
| expand    | bank       | range             | brook         | occupy         |
| replace   | situate    | north             | drain         | fix            |
| follow    | mouth      | construct         | headwater     | invasion       |
| connect   | near       | east              | confluence    | conquest       |
| conjunct  | surrounding| beginning         | lagoon        | construct      |
| end       | estuary    | valley            | catchment     | territory      |

| Island,Country | School,City | University,Company | University,University | Airport,Settlement |
|----------------|-------------|--------------------|-----------------------|-------------------|
| archipelago    | school      | founder            | college               | airport           |
| war            | college     | unite              | collaborate           | construct         |
| country        | academy     | own                | rename                | airfield          |
| defeat         | located     | build              | change                | east              |
| return         | build       | collaborate         | depend                | northeast         |
| occupied       | north       | business           | john                  | located           |
| controlled     | south       | experiment         | hold                  | near              |
| north          | constructed | management         | learn                 | south             |
| south          | site        | award              | receive               | site              |
| east           | independent | change             | merge                 | serve             |

Figure 4. Example of the detected key phrases of geo-entity pair types.
Teide is the highest peak of Spain.” Further, the meanings of these key phrases can be abstracted as “be located in,” and hence, these key phrases still indicate the geo-relation “within.”

Some well-defined properties in the current GeoKGs describe the relationships between geo-entities, whereas these properties do not yet cover the geo-relation occurring in web texts. Thus, we compared the scale difference between the properties in DBpedia and the key phrases detected using the proposed method. The statistics on the properties of each geo-entity pair type originate from DBpedia Ontology.

There are 1,924 geo-entity pair types having at least one property or detected key phrase. Figure 5 shows the cumulative frequency of the key phrase (property) number of each geo-entity pair type. It is apparent that, compared with the well-defined properties in the GeoKGs, web text provides more key-phrases to describe the geo-relations for the same geo-entity pair types. Concretely, the total number of properties referring to the geo-relations is 1,571, and the total number of detected key phrases is 10,031. The average number of properties of each geo-entity pair type is 0.82, and the number of detected key phrases is 5.21. The geo-entity pair types whose number of detected key phrases is greater than that of the properties is 1,542 out of 1,924. Therefore, the key phrases detected from unstructured web texts have the potential to indicate new geo-relations that differ from the existing geo-relations in the current GeoKGs. Subsequently, the new geo-relations and geo-entities will compose new triplets, which can enrich the geographic knowledge of GeoKGs.

4.2.2. Precision

The quality of a key-phrase detection method correlates with how reliably the positive or negative key phrases can be detected. As the number of key phrases among the entire experiment data is unknown, we define the precision as Equation (9). \( \text{Cnt(right set)} \) denotes how many of the detected key phrases are correct. \( \text{Cnt(result set)} \) denotes the total number of key phrases in the results.

![Figure 5](image-url)

Figure 5. Cumulative frequency of the key phrase (property) number of each geo-entity pair type.
\[
\text{Precision} = \frac{\text{Cnt(right set)}}{\text{Cnt(result set)}}
\]  

\[ (9) \]

In addition, we compare the proposed method with the DF and entropy methods to validate the effect of sparseness reduction on detecting geo-relation key phrases from web texts. The DF and entropy methods are directly applied to the original contexts without the merging of contexts.

Table 3 shows the number of extracted geo-entity pair types and the detected key phrases of the three methods in the experimental dataset. It can be observed that the DF method detected more key-phrases than the proposed method, whereas the entropy method missed some key-phrases.

We randomly sample 1,000 geographic facts from the extracted results and manually annotate whether the detected key-phrase in each fact is correct. Table 4 displays the precision of detection of the proposed, DF, and entropy methods. The results show that the proposed method performs better than the DF and entropy methods on web texts where the distribution of geo-relation key phrases is sparse. The precision of the proposed method surpasses that of the DF and entropy methods by 21.80% and 18.20%, respectively. Thus, the key phrases detected by the proposed method have higher reliability and can be used to extract high-quality triplets.

We further analyzed the results and found that compared with DF and Entropy methods, the proposed method performs better when the frequency of the key phrase is equal to or less than that of the other terms in the original context. For example, (1) the frequency of the terms “central” and “park” in the original context is 46,884 and 43,424, respectively. The DF method detected “central” as the key phrase of the geo-entities pair (Plomb du Cantal, Monts du Cantal) from the text “The department is named for the Plomb du Cantal, the central peak of the bare and rugged Monts du Cantal mountain chain which traverses the area”; (2) The frequency of the terms “defeated,” “final,” and “season” in the original context is 28,037, 31,794, and 69,223, respectively. The DF and entropy methods detected “final” and “season” as the key phrases of (Sydney, Collingwood) from the text “Sydney finished the 2007 home and away season in 7th place, and advanced to the finals, where they faced and were defeated by Collingwood by 38 points in the elimination final.”

The above examples show that the sparse distribution of the key-phrase terms in web texts seriously affected the capacities of the DF and entropy methods. Furthermore, the proposed

| Method   | #Extracted geo-entity pair type | #Detected key phrases |
|----------|--------------------------------|-----------------------|
| DF       | 1,896                          | 12,753                |
| Entropy  | 1,709                          | 8,626                 |
| Proposed | 1,784                          | 10,031                |

**Table 3. Statistics of key-phrase detections.**

| Method       | Cnt(right set) | Precision(%) |
|--------------|----------------|--------------|
| DF           | 509             | 50.90%       |
| Entropy      | 545             | 54.50%       |
| Proposed     | 727             | 72.70%       |

**Table 4. Precision of key-phrase detection.**
method increases the frequency of the key-phrase terms in the enhanced contexts to improve the detection performance by merging the contexts of the same geo-entity pair types.

### 4.3. Discussion

As mentioned in Section 2, the frequency-based statistic methods for key-phrase detection are derived from TFIDF and entropy. TFIDF is based on the premise that entity relations appear frequently in massive texts, and entropy is dependent on the hypothesis that the relational terms used to describe a specific relation appear more often than others. TFIDF and entropy assess the importance of terms using frequency statistics. Unfortunately, there are typically no significant differences between the frequencies of key phrases describing spatial relations and other terms because of sparse distribution. Thus, it is difficult to distinguish the key phrases from contexts using frequency-based statistic methods for recognizing geo-entity relations. Therefore, the TFIDF (including DF) and entropy methods do not effectively detect key phrases for geo-relations in web texts where the distribution of terms is sparse, especially for new types of key phrases. In contrast, we detect the key phrases from the corpus generated from the enhanced contexts by introducing external semantic knowledge, which reduces the sparsity of terms by merging the contexts of the same types of geo-entity pairs and fusing the terms with similar semantics. Thus, the proposed method improves the performance of the DF and entropy methods.

Furthermore, the balance between reliability and coverage is maintained by combining the type of geo-entity with the lexical features of the term and semantic fusion, which produces massive key phrases with higher quality than the other two statistic methods when dealing with a sparse geo-entity relation in web texts. Particularly, this advantage is prominent when the description patterns of key phrases in texts are relatively homogeneous. In this case, incorrect detections appear with a very low probability. More importantly, our method has a strong ability to discover new key phrases, making it possible to make up for the limitation of supervised learning methods on relation extraction, which can only recognize predefined types of relations.

However, there are two kinds of cases we could not effectively handle yet.

1. **Key phrase with semantic constraint.** Sometimes, relations are dependent on semantic constraints, which no longer satisfy the given format of a triplet in a GeoKG. Although these semantic constraints can be reflected by the POS of the term, they are more complex because they have no significant frequency and lexical features. For example, the statement “Mount Everest, known in Nepali as Sagarmatha and in Tibetan as Chomolungma” expresses two facts: Mount Everest, known in Nepali as Sagarmatha and in Tibetan as Chomolungma. The proposed method can detect the key phrases “known,” but misses the semantic constraint of different regions. More features should be considered when dealing with key phrases with semantic constraints, such as the grammatical structure and semantic coherence. Moreover, dependency parsing is an effective solution for generating special expressions to recognize semantic constraints (Schmitz, Bart, Soderland, & Etzioni, 2012).

2. **Implicit key phrase.** A sentence indicates a kind of relation between two geo-entities, whereas the key phrases describing this kind of relation do not appear in the sentence. For example, the sentence “Zijin had 49.28 tons of the gold output and the gold produced from mining reached 20.70 tons, respectively accounting for 20.53% of China’s total gold production” indicates a topological relation (Zijin, within, China), but there are no terms meaning...
“within” in the sentence. In this case, the term “output” detected by our method is irrelevant to the relation of “within.” Other implicit key phrases are often described in certain syntax statements – for instance, “Jaén (Spain),” where the parenthesis indicates that “Jaén” is within “Spain.” Pattern mining and linguistic rules may be helpful to understand this type of implicit relation (Quan, Wang, & Ren, 2014).

Note that the proposed method should calculate the similarities between terms in each merging context (by Equation (1)) and the distance between enhanced contexts in the entropy calculation of each term (by Equation (3)) based on word embedding. Therefore, the computation complexity of the proposed method is higher than those of the DF and entropy methods. However, the computation complexity does not increase significantly in practice. First, as the number of terms in each merging context is limited (for example, the mean size of the word bag of the merging context in the experimental dataset is 3,461; the size of the word bag of the entire experimental dataset is 172,045), the calculation amount of term similarity is small. Second, in the entropy calculation of each term, when removing \( t \) to recalculate the vectors of each enhanced context, we can pre-compute the vectors of each enhanced context with all the terms and then subtract the vectors of \( t \) from the pre-computed vector to avoid the large real-time addition operation of retained terms.

Moreover, the triplets with the detected key phrases are not directly used to construct or complete GeoKGs. The reason is that there still exist some redundant key phrases with the same semantic in the detected results, but the similar geographic relationship semantic should be defined as a unified geo-relation in the GeoKGs. For example, the detected key phrases “located (in),” “situated (in),” and “sit (at)” express the geo-relation “within.” Thus, semantic clustering or semantic aligning is an essential follow-up processing to generate new geo-relations from key phrases or fuse key phrases to the existing geo-relations in GeoKGs.

5. Conclusions

This paper proposed a context-enhanced method to detect geo-relation key phrases for the extraction of triplets from web texts where the geo-relation description is sparsely distributed. The main idea of the proposed method is introducing external semantic knowledge to alleviate the sparseness of geo-relation description terms in web texts. Specifically, the contexts of geo-entities are fused into enhanced contexts with category semantic knowledge and word semantic knowledge. Subsequently, the frequency-based statistical methods and lexical features can perform better on corpus generation and key-phrase extraction.

In comparison with the direct deployment of well-known frequency-based methods, the proposed method improves the precision of detecting the geo-relation key phrases from real web texts by approximately 20%. Moreover, compared with the well-defined geo-relation properties in DBpedia, the proposed method provides quintuple key-phrases for indicating the geo-relation between geo-entities. It is argued that the proposed method can efficiently enhance the ability to discover key phrases representing geo-entity relations with sparse distribution, as well as to detect massive new key phrases, which are beneficial for generating new triplets for the construction of a GeoKG from web texts.

Future studies will aim to (1) extend the form of triplets to represent the semantic constraint hiding in geo-relations better; (2) introduce deep dependency parsing into key-phrase detection to deal with the complex linguistic phenomena of geo-relation
description in web texts; and (3) apply the proposed method to large-scale web texts and generalize the detected key phrases to unified geo-relations to construct a GeoKG.

Notes
1. https://wiki.dbpedia.org/.
2. https://developers.google.com/freebase/.
3. https://www.dbpedia-spotlight.org/.

Data availability statement
The data referred to in this paper is not publicly available at the current time.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This research was supported by the National Natural Science Foundation of China [41631177, 41801320].

References
Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155.
Blessing, A., & Schütze, H. (2010). Self-annotation for fine-grained geospatial relation extraction. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 80–88). Beijing, China.
Cao, C., Wang, S., & Jiang, L. (2014). A practical approach to extracting names of geographical entities and their relations from the web. In *International conference on knowledge science, engineering and management* (pp. 210–221). Sibiu, Romania.
Chen, J., Ji, D., Tan, C. L., & Niu, Z. (2005). Unsupervised feature selection for relation extraction. In *Proceedings of the 2nd international joint conference on natural language processing* (pp. 262–267). Jeju Island, Korea.
Dash, M., Choi, K., Scheuermann, P., & Liu, H. (2002). Feature selection for clustering-a filter solution. In *Proceedings of international conference on data mining* (pp. 115–122). Maebashi, Japan.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 4171–4186), Minneapolis, MN.
Elia, A., Guglielmo, D., Maisto, A., & Pelosi, S. (2013). A linguistic-based method for automatically extracting spatial relations from large non-structured data. In *Proceedings of the 13th international conference on algorithms and architectures for parallel processing* (pp. 193–200). Vietri sul Mare, Italy.
Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd annual meeting on association for computational linguistics* (p. 415). Barcelona, Spain.
Jin, Y., Zhao, S., & Wu, Y. (2019). Geographic entity relationship extraction model based on piecewise convolution of residual network. In the 3rd international conference on machine learning and soft computing (ICMLSC 2019) (pp. 160–165). Da Lat, Viet Nam.

Khan, A., Vasardani, M., & Winter, S. (2013). Extracting spatial information from place descriptions. In Proceedings of the first ACM SIGSPATIAL international workshop on computational models of place (p. 62). Orlando, Florida.

Li, W., Goodchild, M. F., & Raskin, R. (2014). Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. International Journal of Digital Earth, 7(1), 17–37.

Liu, K., Gao, S., Qiu, P., Liu, X., Yan, B., & Lu, F. (2017). Road2Vec: Measuring traffic interactions in urban road system from massive travel routes. ISPRS International Journal of Geo-Information, 6(11), 321.

Loglisci, C., Lenzo, D., Roche, M., Teisseire, M., & Malerba, D. (2012). Toward geographic information harvesting: Extraction of spatial relational facts from web documents. In 2012 IEEE 12th international conference on data mining workshops (pp. 789–796). Brussels, Belgium.

Mai, G., Yan, B., Janowicz, K., & Zhu, R. (2019). Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In the 22nd AGILE Conference on Geographic Information Science (AGILE 2019) (pp. 21–39). Limassol, Cyprus.

Mesquita, F. (2012). Clustering techniques for open relation extraction. In Proceedings of the SIGMOD/PODS 2012 PhD symposium (pp. 27–32). New York, USA.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In the Workshop at International Conference on Learning Representations 2013 (pp. 1–12). Scottsdale, Arizona, USA.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP (ACL ‘09) (pp. 1003–1011). Suntec, Singapore.

Mirrazei, S. I., Martins, B., & Cruz, I. F. (2016). A distantly supervised method for extracting spatio-temporal information from text. In the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL ‘16) (pp. 74:1–74:4). Burlingame, California.

Naughton, M., Stokes, N., & Carthy, J. (2010). Sentence-level event classification in unstructured texts. Information Retrieval, 13(2), 132–156.

Pershina, M., Min, B., Xu, W., & Grishman, R. (2014). Infusion of labelled data into distant supervision for relation extraction. In Proceedings of the 52nd annual meeting of the association for computational linguistics (pp. 732–738). Baltimore, MD.

Quan, C., Wang, M., & Ren, F. (2014). An unsupervised text mining method for relation extraction from biomedical literature. Plos ONE, 9(7), e102039.

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). Open language learning for information extraction. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 523–534). Jeju Island, Korea.

Shen, M., Liu, D. R., & Huang, Y. S. (2012). Extracting semantic relations to enrich domain ontologies. Journal of Intelligent Information Systems, 39(3), 749–761.

Smole, D., Čeh, M., & Podobnikar, T. (2011). Evaluation of inductive logic programming for information extraction from natural language texts to support spatial data recommendation services. International Journal of Geographical Information Science, 25(11), 1809–1827.

Stefanakis, E. (2003). Modelling the history of semi-structured geographical entities. International Journal of Geographical Information Science, 17(6), 517–546.

Xiong, S., Mao, J., Duan, P., & Miao, S. (2017). Chinese geographical knowledge entity relation extraction via deep neural networks. In the 9th international conference on agents and artificial intelligence (ICAART 2017) (pp. 24–33). Porto, Portugal.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., & Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the web. In Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 2-Volume 2 (pp. 1021–1029). Suntec, Singapore.
Zhang, P., Li, W., Hou, Y., & Song, D. (2011). Developing position structure-based framework for Chinese entity relation extraction. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(3), 14.

Zhang, W., Sun, L., & Zhang, X. (2012). An entity relation extraction method based on wikipedia and pattern clustering. *Journal of Chinese Information Processing*, 26(2), 75–81.

Zhu, Y., Zhu, A.-X., Feng, M., Song, J., Zhao, H., Yang, J., … Yao, L. (2017a). A similarity-based automatic data recommendation approach for geographic models. *International Journal of Geographical Information Science*, 31(7), 1403–1424.

Zhu, Y., Zhu, A.-X., Song, J., Yang, J., Feng, M., Sun, K., … Zhao, H. (2017b). Multidimensional and quantitative interlinking approach for linked geospatial data. *International Journal of Digital Earth*, 10(9), 923–943.