

# Context Enhanced Keyword Extraction for Sparse Geo-Entity Relation from Web Texts

Li Yu<sup>1</sup>, Feng Lu<sup>1,2(✉)</sup>, Xueying Zhang<sup>3</sup>, and Xiliang Liu<sup>1,4</sup>

<sup>1</sup> State Key Laboratory of Resources and Environmental Information System,  
Institute of Geographic Sciences and Natural Resources Research,  
Chinese Academy of Sciences, Beijing 100101, China

{yu1, luf, liuxl}@lreis.ac.cn

<sup>2</sup> Jiangsu Center for Collaborative Innovation in Geographical Information  
Resource Development and Application, Nanjing 210023, China

<sup>3</sup> Key Laboratory of Virtual Geography Environment,  
Nanjing Normal University, Nanjing 210046, China

zhangsnowy@163.com

<sup>4</sup> Fujian Collaborative Innovation Center for Big Data Applications  
in Governments, Fuzhou 350003, China

**Abstract.** Geo-entity relation recognition from rich texts requires robust and effective solutions on keyword extraction. Compared with supervised learning methods, unsupervised learning methods attract more attention for their capability to capture the dynamic feature variation in text and to discover additional relation types. The frequency-based methods of keyword extraction have been widely studied. However, it is difficult to be applied into geo-entity keyword extraction directly because of the sparse distribution of geo-entity relations in texts. Besides, there are few studies on Chinese keyword extraction. This paper proposes a context enhanced keyword extraction method. Firstly the contexts for geo-entities are enhanced to reduce the sparseness of terms. Secondly two well-known frequency-based statistical methods (i.e., DF and Entropy) are used to build a large-scale corpus automatically from the enhanced contexts. Thirdly the lexical features and their weights are statistically determined based on the corpus to enhance the distinction of the terms. Finally, all terms in the enhanced contexts are measured with the lexical features, and the most important terms are selected as the keywords of geo-entity pairs. Experiments are conducted with mass real Chinese web texts. Compared with DF and Entropy, the presented method improves the precision by 41 % and 36 % respectively in discovering the keywords with sparse distribution and generates additional 60 % correct keywords for geo-entity relation recognition.

**Keywords:** Geographical information retrieval · Geo-entity relation · Keyword extraction · Text mining · Context enhancement

## 1 Introduction

The web provides important and even exclusive resources for geographic information retrieval and knowledge discovery [1]. At the same time, geo-entity relations are commonly used in describing the locations of entities and geographical phenomena which are crucial for building geographic knowledge systems [2]. To better understand the geographic semantics embedded in rich web texts, it's a pressing need for robust and effective solutions in geo-entity relation extraction.

The frequently used supervised learning methods which perform well with specified static texts behave poorly in extracting geo-entity relations from web texts [3]. Firstly, building massive patterns or corpora are expensive and training models is time-consuming, the massive web texts cannot be processed in real-time with supervised methods [4]. Secondly, web texts may cover various domains with strong heterogeneities, leading to a poor portability for model training [5]. Thirdly, the dynamic nature of web texts constantly generates additional relation types which cannot be captured by predefined patterns and pre-trained models [6]. The unsupervised learning methods have attracted more attentions in the field of web texts mining because they don't need large scale patterns and corpora. Additionally, they can be utilized for additional relation exploring, which are more suitable for dynamic text mining [7].

Keywords play an important role in relation recognition with unsupervised learning methods, which provide rich clues to describe the relations between entities [8]. Unsupervised methods regard keyword extraction as a ranking task and extract the top-ranked as keywords [9]. The existing keyword extraction methods for relation recognition are mainly based on frequency statistics. These methods are based on the hypothesis that there exist a large number of redundant terms which imply the relations for a specific entity pair. However, this hypothesis is not appropriate to extract keywords for geo-entity based on the following reasons: Firstly, the specific geo-entity pair rarely co-occurs in one sentence based on our experiments [10]. Besides, the number of terms in the context of the specific geo-entity pair is very limited, which makes the terms rather sparse. Secondly, the synonymy exacerbates the problem of sparseness [11]. Thirdly, there is a strong correlation between the types of geo-entity and the terms [12]. For example, "flow into" can only describe the relation between water bodies, not buildings. However, it is not applicable for semantic relations which are not restricted by the type of geo-entity pair. Therefore, only frequency statistic is hard to distinguish the keywords from others and will not work well in recognizing geo-entity relations with sparse distribution. Besides, different languages vary in word segmentation, part-of-speech (POS) tagging and syntactic analyzing, which have a great influence on keyword extraction. Compared with English, a character-based language like Chinese needs a different strategy of keyword extraction for geo-entity relation.

This paper focuses on how to extract keywords from mass Chinese web texts for recognizing geo-entity relations with extremely sparse distribution. Our contributions are as follows:

- (1) We propose the context enhanced method to reduce the term sparseness of keyword extraction. To the best of our knowledge, the sparse distribution of geo-entity relation is firstly presented in the field of geo-entity relation

recognition. We also prove sparseness reduction is essential for generating high-quality keywords and achieving an unsupervised recognizing method of sparse geo-entity relation.

- (2) In order to reveal the specific characteristics of the given web texts and deal with heterogeneous web texts, we use feature selection and weight statistics to increase the distinctions between the terms in context. Different with the frequency-based methods, we additionally explore multiple lexical features in real-time and dynamically adjust their weights.
- (3) Our method significantly outperforms other comparing algorithms (DF and Entropy), and has the ability of discovering additional keywords that is appropriate to dynamic text mining.

The remainder of this paper is organized as follows. A context enhanced methodology of keyword extraction for sparse geo-entity relation is presented in Sect. 2. The experiments and discussion are presented in Sect. 3. Conclusion is drawn in Sect. 4.

## 2 Methodology

### 2.1 Definitions

**Input:** Chinese texts crawled from assigned websites. One piece of texts is shown below.

中关村位于海淀区，邻近北京大学和清华大学。此外，中关村是中国的科技中心，被誉为“中国的硅谷”。

*Zhongguancun is located in Haidian District, along with the proximity to Peking University and Tsinghua University. Besides, Zhongguancun is a technology hub in China, and is referred to as “China’s Silicon Valley”.*

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

**Geo-entity pair** ( $e_1, e_2$ ): two geo-related entities co-occurring in one sentence. The first geo-entity appearing in one sentence is paired with other geo-entities in the same sentence. For example, (中关村Zhongguancun, 海淀区Haidian District), (中关村Zhongguancun, 北京大学Peking University) and (中关村Zhongguancun, 清华大学Tsinghua University) are geo-entity pairs in the first sentence.

**Geo-entity relation**  $r$ : a state of connectedness between geo-entities, divided into two types, spatial relations and semantic relations. Spatial relations consist of topological, directional and distance relations, such as “within”, “south” and “10 kilometres”. Semantic relations are “hypernym”, “hyponym”, “equal”, to name a few. Both of them can be represented as a set of facts with the form ( $e_1, r, e_2$ ). The examples of fact are

(中关村, 相邻, 北京大学) and (中关村, 别名, 中国的硅谷), which are (Zhongguancun, adjacent, Peking University) and (Zhongguancun, alias, China’s Silicon Valley) in English.

**Term  $t$ :** a phrase or a word with semantic information in a sentence, such as “位于”, “科技中心”(“be located in”, “technology hub” in English) and so on.

**Context  $c$ :** all 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. The stop words are function words, such as “被”, “并”, “都” (“by”, “and”, “both” in English) and so on. For example, the context of (中关村, 海淀区), (中关村, 北京大学) and (中关村, 清华大学) contains 2 terms, (位于, 邻近) which are (be located in, proximity) in English.

**Keyword  $k$ :** the terms picked out from context as indicators in relation expressions. For example, the term “proximity” picked out from the context (be located in, proximity) is a keyword revealing the topological relation “*adjacent*” for the geo-entity pair (Zhongguancun, Peking University) (Table 1).

**Table 1.** Examples of geo-entity pairs and corresponding keywords.

Geo-entity pairs	Keywords
(中关村, 海淀区) (Zhongguancun, Haidian District)	(位于) ( <i>be located in</i> )
(中关村, 北京大学) (Zhongguancun, Peking University)	(邻近) (proximity)
(中关村, 清华大学) (Zhongguancun, Tsinghua University)	(邻近) (proximity)
(中关村, 中国) (Zhongguancun, China)	(科技中心) (technology hub)
(中关村, 中国的硅谷) (Zhongguancun, China’s Silicon Valley)	(誉为) (as)

## 2.2 Sparseness Reduction

The terms in the context of a specific geo-entity pair are usually sparse. Merging the contexts of geo-entity pairs with the same type will reduce the sparseness of terms in one context. This requires a fine-grained mapping table connecting types to geo-entities. In this paper, an online Chinese encyclopaedia (Baidu Baike<sup>1</sup>) is used for obtaining the type labels of each geo-entity. Similar to Wikipedia, Baidu Baike attaches each piece of web texts with multiple type labels according to the ranked importance for each entry. For example, the entry “Beijing” has 4 type labels, “municipality”, “ancient capital”, “China” and “first-tier city”.

The process of sparseness reduction for terms is shown in Fig. 1. Firstly, we search the geo-entities in Baidu Baike one by one, and obtain the corresponding label types.

<sup>1</sup> <http://baike.baidu.com>.

Secondly, all type labels of the specified geo-entity are assessed by using their orders and frequencies, and the most important label is picked out as the geo-entity type. After all geo-entity is assigned its type, the type of geo-entity pair ( $e_x, e_y$ ) can be decided with the name  $T_{xy} = \langle \text{type}_{e_x}, \text{type}_{e_y} \rangle$ . Thirdly, we merge the contexts of geo-entity pairs with the same type, and the number of terms in context will be increased. This process enhanced the information used to extract keywords for geo-entity pairs. Moreover, the term's semantics are also fused with the help of the synonym dictionary *CiLin*<sup>2</sup> to reduce the sparseness of terms.

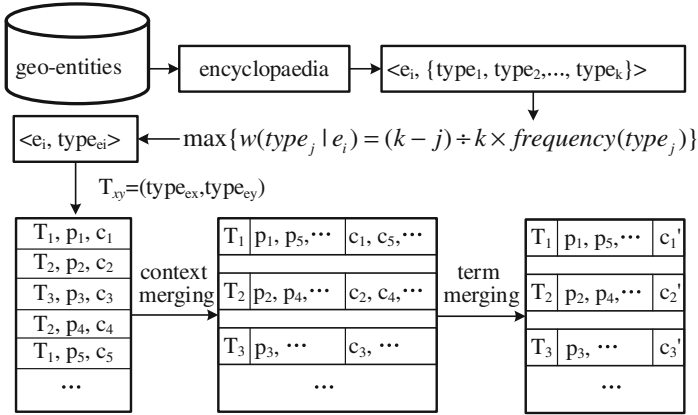


Fig. 1. Sparseness reduction for terms in contexts.

### 2.3 Corpus Generation

A large-scale corpus is needed to select the effective features for keyword extraction. It is generated automatically based on two well-known frequency-based statistical methods, namely the DF (Domain Frequency) and Entropy. DF and Entropy methods are used for extracting keywords from the entire web texts. The intersection of these two resulted keyword sets forms the corpus for feature selection. DF is shown in formula (1). Entropy is shown in formula (2)–(3).

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

$$S_{i,j} = \exp\left(\frac{\ln 0.5}{D} \times D_{i,j}\right) \quad (2)$$

<sup>2</sup> <http://www.datatang.com/data/42306/>.

$$Entropy_t = \begin{cases} -\sum_{i=1}^N \sum_{j=1}^N (S_{i,j} \log S_{i,j} + (1 - S_{i,j}) \log(1 - S_{i,j})), & 0 < S_{i,j} < 1 \\ 0, & \text{others} \end{cases} \quad (3)$$

In formula (1),  $f_{t,T_i}$  denotes the frequency of term  $t$  appearing in the contexts of geo-entity pairs with the type  $T_i \in TS$ .  $TS$  is the type set of geo-entity pairs with the size of  $N$ . In formula (2),  $S_{i,j}$  denotes the similarity between the context  $p_i$  and  $p_j$ , which is measured by the average distance of all contexts and the distance  $D_{i,j}$  between  $p_i$  and  $p_j$  after removing the term  $t$  from all contexts. Formula (3) denotes the entropy of term  $t$  measured by  $S_{i,j}$ .

## 2.4 Feature Selection

Feature selection is crucial for keyword extraction, which has been proved to have a positive effect on classification accuracy [13] as well as be able to reveal the nature of keywords more comprehensively from multiple perspectives instead of the single aspect “term frequency”. Taking the text piece example in Sect. 3.1, the selected features are defined as follows.

- (1) The POS of term (noun, verb, preposition or others). e.g., the POS of “邻近” is a verb in Chinese with a meaning of ‘be close to’.
- (2) The length of term, which is measured by the number of characters. e.g., the length of “邻近” is 2, which means “邻近” has 2 characters.
- (3) The location of term (left of  $e_1$ , between  $e_1$  and  $e_2$ , or right of  $e_2$ ). e.g., the location of “邻近” is between the geo-entity pair ( $e_1 = \text{中关村}$ ,  $e_2 = \text{北京大学}$ ).
- (4) The previous term just before  $e_1$ . e.g., the previous term just before  $e_1 = \text{中关村}$  is null.
- (5) The next term just after  $e_1$ . e.g., the next term just after  $e_1 = \text{中关村}$  is “位于”.
- (6) The previous term just before  $e_2$ . e.g., the previous term just before  $e_2 = \text{北京大学}$  is “邻近”.
- (7) The next term just after  $e_2$ . e.g., the next term after  $e_2 = \text{北京大学}$  is “和”.
- (8) The distance between the term and  $e_1$ . e.g., the distance between “邻近” and  $e_1 = \text{中关村}$  is 3. Note that the distances in features (8)–(11) are measured by the number of elements after word segmentation.
- (9) The distance between the term and  $e_2$ . e.g., the distance between the term “邻近” and  $e_2 = \text{北京大学}$  is 0.
- (10) The distance between the term and the head of sentence. e.g., the distance between the term “邻近” and the head of sentence is 4.
- (11) The distance between the term and the tail of sentence. e.g., the distance between the term “邻近” and the tail of sentence is 4.

## 2.5 Term Assessing

After selecting features, the process of term assessing is conducted, this considers the influence of the length, POS, location and distance of the terms, shown in formula (4)–(8). These lexical features are statistically determined according to the credible results of two frequency-based statistical methods and changed with the input texts in real time.

$$wgt^{(t)} = \theta_{LEN} \times (\theta_{POS} + \theta_{LOC} + \theta_{DIS}) \quad (4)$$

$$\theta_{LEN} = \begin{cases} 1, & \min < length(t_{pos}) < \max \\ 0, & others \end{cases} \quad (5)$$

$$\theta_{POS} = p(t_{POS}) \quad (6)$$

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

$$\theta_{DIS} = \begin{cases} p(dis(e_1)|t_{loc}) \\ p(dis(e_2)|t_{loc}) \\ p(dis(head)|t_{loc}) \\ p(dis(tail)|t_{loc}) \end{cases} \quad (8)$$

In formula (4),  $wgt^{(t)}$  denotes the weight of term  $t$  for the specified geo-entity pair, considering the importance of length  $\theta_{LEN}$ , part-of-speech  $\theta_{POS}$ , location  $\theta_{LOC}$  and distance  $\theta_{DIS}$ . Formula (5) denotes the weight of the length of term  $t$  affected by the POS of  $t$  ( $t_{pos}$ ). The length of each type of POS has its own valid range. The  $wgt^{(t)}$  will be equal zero if the length of  $t$  with  $t_{pos}$  is out of the range. Formula (6) denotes the weight of POS, which is the probability of the event that the POS of  $t$ , namely ( $t_{pos}$ ), is equal to the specific part-of-speech. Formula (7) denotes the weight of relative location affected by the previous and next terms of geo-entity.  $t_{loc}$  denotes relative location of term  $t$ , which can be left, between or right.  $tp(e_1)$  denotes the previous term of  $e_1$ ,  $tn(e_1)$  denotes the next term of  $e_1$ . For example,  $p(t_{loc} = between|tp(e_1))$  denotes the probability that the term  $t$  located between  $e_1$  and  $e_2$  is the keyword when the previous term of  $e_1$  is a specific term. Formula (8) denotes the weight of distance affected by the location of term.  $dis(e_1)$  denotes the distance between  $t$  and  $e_1$ .  $dis(e_2)$  denotes the distance between  $t$  and  $e_2$ .  $dis(head)$  denotes the distance between  $t$  and the head of the sentence.  $dis(tail)$  denotes the distance between  $t$  and the tail of the sentence. For example,  $p(dis(e_1)|t_{loc} = between)$  denotes the probability that the term  $t$  with a definite distance to  $e_1$  is the keyword when  $t$  is located between  $e_1$  and  $e_2$ .

All terms in contexts are assessed by formula (4) and ranked in descending order. After ranking, a local ordered list of terms is generated for each geo-entity pair, which indicates the decreasing importance of the terms for geo-entity relation expression. The most important term is picked out as the keyword of the specified geo-entity pair.

### 3 Experiments

#### 3.1 Dataset

All the articles on Chinese national geography are crawled from Encyclopaedia of China<sup>3</sup>, with 2.3 million words in total. These articles describe the geographic, cultural and historical knowledge of toponyms, which provide rich information for geo-entity relation extraction. These articles are pre-processed using GATE<sup>4</sup> and 31,065 geo-entity pairs are generated. They are randomly divided into 3 groups to check the robustness of the proposed method.

#### 3.2 Baselines

The proposed method is compared with DF and Entropy. Specifically, DF method extends the classic TFIDF using the frequency of the terms in the context of the type-specific entity pairs, which would favor specific relational terms as opposed to generic ones. Entropy method converts the context to a vector of terms and assesses the discrimination of each term based on the informatics theory, which would provide useful heuristic information for keyword extraction.

#### 3.3 Metrics

Because the number of the keywords in the experiment is unknown, we can only define the precision as shown in formula (9).  $Cnt(right\ set)$  denotes how many the extracted keywords are correct.  $Cnt(result\ set)$  denotes the total number of keywords in the results.

$$Precision = \frac{Cnt(right\ set)}{Cnt(result\ set)} \quad (9)$$

We randomly sample part of data from the results, and manually evaluate them by two people, and evaluate the coherence of their annotation by  $kappa$  coefficient ( $\kappa$ ) as formula (10).  $P_o$  denotes the relative annotation agreement between the two people,  $P_e$  denotes the hypothetical probability of chance agreement. If  $\kappa > 0.8$ , the annotations

<sup>3</sup> [http://www.360doc.com/content/11/01/10/01/694750\\_85358960.shtml](http://www.360doc.com/content/11/01/10/01/694750_85358960.shtml).

<sup>4</sup> <https://gate.ac.uk/>.



are accepted and the mean precision of the two evaluations is calculated. Otherwise, evaluation is conducted again.

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (10)$$

### 3.4 Results

#### 3.4.1 Keyword Extraction

We utilize the proposed method with the first group as an example. The results are shown in Fig. 2. The terms in context are ordered by their descending importance ranks, the one with the maximal weight is picked out as the keyword for each pair of geo-entities. Note that some geo-entity pairs own multiple keywords because multiple terms in one context have the equal weight. For example, the geo-entity pair (*Zhejiang Province, Qiandao Lake*) has keywords “artificial-lake” and “reservoir”.

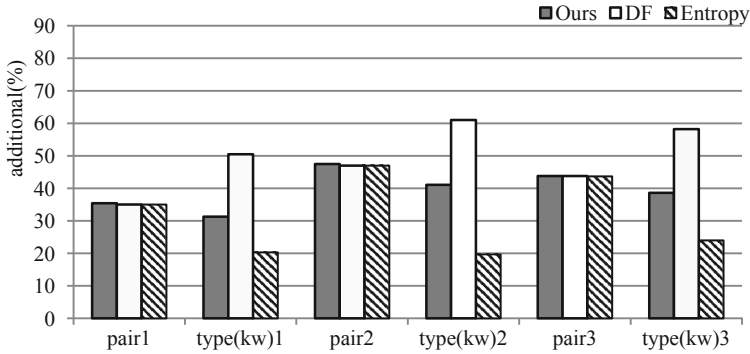
1	$e_1$	$e_2$	terms in context	keyword
2	Pukou	Yangtze River	north, locate	north
3	Pulan	Menshi	west, coal-mining, have	west
4	China	Mount Putuo	buddhism, four, mount	buddhism
5	Longsha Park	Wangjiang Tower	scenic, have	scenic
6	Tsitsihar	Songnen plain	locate	locate
7	Qimen County	Mount Qishan	near, have	near
8	JiangSu Province	Qidong City	govern, granary, cotton-area	govern
9	Zhejiang Province	Qiandao Lake	artificial-lake, reservoir	artificial-lake, reservoir
10	Qiantang River	Xin'an River	upstream, origin	upstream
11	Mount Qingcheng	Mount Tiancang	contain	contain
12	Qingtian County	Shimen Cave	relic, have	relic
13	Wuding River	Luhe River	tributary	tributary
14	Mount Tianshan	Turpan	basin, locate	basin
15	Mount Data	Qianfo Cave	main-pick, north, south, east	main-pick
16	Qinling	Canglongpu		name

Fig. 2. Examples of extracted keywords in the first group of data.

#### 3.4.2 Additional Keywords

Compared with the corpus, some new geo-entity pairs and keywords in each group are extracted, as shown in Fig. 3. In the horizontal axis, pair denotes geo-entity pair, type (kw) denotes the number of keyword’s types, and the index numbers correspond to each group. The vertical axis denotes how many new objects are extracted. For example, in the first group, 35.4 % additional geo-entities pairs and 31.3 % additional keyword’s types are generated with our method.

The extraction percentage of additional geo-entity pairs is almost the same in the three methods. Additionally, the DF explores the largest number of new types of



**Fig. 3.** Additional geo-entity pairs and keywords

keywords (average 56.6 % in three groups), while the Entropy misses the most of the keywords.

### 3.4.3 Precision

The extracted keywords are evaluated manually and the *kappa* coefficient  $\kappa$  is calculated. The additional objects are evaluated to assess the ability of a keyword extraction method adapting to the unknown data in the corpus. 100 additional geo-entity pairs with additional types of keywords are sampled randomly from the results, and added into the evaluation set. Then two people simultaneously check if the extracted keyword in the evaluation set is the relational term of one specific geo-entity pair. The *kappa* coefficient  $\kappa$  with a value 0.83 declares a high coherence and proves the validity of the evaluation.

**Table 2.** The precision of three methods for all additional extracted keywords (%)

Precision(%)	Ours	DF	Entropy
new(kw)	60.3	31.7	26.7
AVG	85.5	44.1	49.4

Table 2 shows that how many geo-entity pairs with the additional types of keyword are extracted correctly (new(kw), in short), and the mean of all results which contains the existed keywords and the new discovered ones extracted correctly (AVG, in short). The proposed method gets an average precision of 85.5 %, which is about 41 % and 36 % higher than DF and Entropy. More importantly, the precision of new types of keywords extracted with the presented method is 60.3 %, surpassing by 28 % and 33 % with DF and Entropy respectively. Although DF method obtains the largest number of new types of keywords (shown in Fig. 3), it has the low precision of new types of keywords (31.7 %). Moreover, Entropy method misses the most keywords and has the lowest precision.

### 3.4.4 Discussion

As mentioned in Sect. 1, the frequency-based methods for keyword extraction are derived from TF-IDF and Entropy. TF-IDF is under the premise that entity relations would appear frequently in massive texts. And Entropy is dependent on the hypothesis that the relational terms used to describe the specific relation appear more often than others. Both TFIDF and Entropy assess the importance of terms by frequency statistic. Unfortunately, there is usually no significant frequency difference between keywords and other terms because the keywords are sparsely distributed. Thus, it is difficult to distinguish the keywords from contexts using the frequency-based methods. Therefore, TFIDF (including DF) and Entropy do not perform well in keyword extraction for sparse geo-entities, especially on the additional types of keywords.

On the contrary, we extract keywords not only with the term frequency, but also the lexical features to reveal the specific characters of the given texts. Besides, the reliability is kept with combining the types of geo-entities with the lexical features, which produces massive keywords with a higher quality dealing with the sparse geo-entity relations. Moreover, our method can discover additional keywords from the original web texts, which is a step forward comparing with supervised learning methods.

However, there are still two kinds of keywords we can't effectively deal with: (1) Keywords with semantic constraints. Sometimes relations depend on time, spatial or semantic constraints, which no longer meet the format of the triplet. For example, the sentence “艾比湖蒙古语称为艾比淖尔(Aibi Lake is called Ebi Bur in Mongol).” expresses the facts (Aibi Lake, alias in Mongol, Ebi Bur). Our method can extract the keywords “be called” which is the meaning of “alias”, but miss the semantic constraint. More features should be considered when dealing with keywords with semantic constraints, such as grammatical structure, semantic coherence and so on. Besides, dependency parsing is also an effective solution for completing relation expression. (2) Implicit keywords. One sentence implies a kind of relation between two geo-entities, whereas the keywords describing this relation do not appear in the sentence. For example, the sentence “The water resources of Min River are 13.32 million kilowatt, accounting for 18.85 % of water resources of Sichuan Province” describes a topological relation (Min River; Sichuan Province; within), but there are no terms meaning “within” in the sentence. Geometric information from geographical knowledge bases (such as Geonames and OpenStreetMap) would be beneficial to extract implicit spatial keywords.

Note that the main contribution of this study is to alleviate the influence of context sparseness. The proposed method solves this problem with the help of a fine-grained mapping table and an open synonym dictionary. Because the languages only influence the feature selection and the weights of features, specific features should be selected in the context enhanced method for different languages.

## 4 Conclusion

This paper proposed a context enhanced method to extract the keywords from mass web texts to recognize geo-entity relations with sparse distributions. We adopt two strategies to reduce the sparseness of terms in contexts. The first is a fine-grained type

table used to merge the contexts for increasing the number of terms, and the second is semantic fusion conducted to reduce the sparseness of terms in all contexts. Moreover, we consider the global and local features by introducing the characteristics of length, part-of-speech, position and distance of terms to improve the performance. It is demonstrated that the proposed method can efficiently enhance the ability of discovering geo-entity relation keywords with sparse distributions. This method also generates massive additional keywords which is helpful to realize the unsupervised learning methods of geo-entity relation recognition.

**Acknowledgments.** This work was partially supported by the National High-Tech Research and Development Program of China (2013AA120305) and the National Natural Science Foundation of China (41271408).

## References

1. Jones, C.B., Purves, R.S.: Geographical information retrieval. *Int. J. Geogr. Inf. Sci.* **22**(3), 219–228 (2008)
2. Kordjamshidi, P., Otterlo, M.V., Moens, M.F.: Spatial role labeling: towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.* **8**(3), 1–39 (2011)
3. Purves, R.S., Clough, P., Jones, C.B.: The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *Int. J. Geogr. Inf. Sci.* **21**(7), 717–745 (2007)
4. Zhu, S.N., Zhang, X.Y., Zhang, C.J.: Syntactic pattern recognition of geospatial relations described in natural language. In: *Proceedings of 2010 International Conference on Broadcast Technology and Multimedia Communication*, 13 December, pp. 354–357. CNKI, Chongqing (2010)
5. Li, W.W., Goodchild, M.F., Raskin, R.: Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *Int. J. Digit. Earth* **7**(1), 17–37 (2014)
6. Loglisci, C., Ienco, D., Roche, M., et al.: Towards geographic information harvesting: extraction of spatial relational facts from web documents. In: *2012 IEEE 12th International Conference on Data Mining Workshops*, 10 December, pp. 789–796. IEEE, Brussels (2012)
7. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* **37**, 141–188 (2014)
8. Zhang, W.R., Sun, L., Han, X.P.: A entity relation extraction method based on Wikipedia and pattern clustering. *J. Chin. Inf. Process.* **26**(2), 75–127 (2012)
9. Liu, Z.Y., Sun, M.S.: Can prior knowledge help graph-based methods for keyword extraction? *Front. Electr. Electron. Eng.* **7**(2), 242–253 (2012)
10. Vasardani, M., Winter, S., Richter, K.F.: Locating place names from place descriptions. *Int. J. Geogr. Inf. Sci.* **27**(12), 2509–2532 (2013)
11. Shen, M.M., Liu, D.R., Huang, Y.S.: Extracting semantic relations to enrich domain ontologies. *J. Intell. Inf. Syst.* **39**(3), 749–761 (2012)
12. Zhang, X.Y., et al.: SVM based extraction of spatial relations in text. In: *2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, 29 June–01 July, pp. 529–533. IEEE, Fuzhou (2011)
13. Naughton, M., Stokes, N., Carthy, J.: Sentence-level event classification in unstructured texts. *Inf. Retrieval* **13**(2), 132–156 (2010)