Tourism Spatial Relation Extraction Based on Clustering

Jin Zhiyao, Du Junping
(School of Computer Science, Beijing University of Posts and Telecom-munications, Beijing 100876)

Abstract: This paper studied a relation extraction approach based on clustering to solve the problem of tourism spatial relation predefining. Firstly we extracted the spatial feature words from domain corpus basing on bootstrapping iteration. Secondly we calculated the similarity between words through semantic thesauruses. Finally, the feature words are clustered and every cluster represents a kind of entity relation discovered automatically. The experimental results showed that this approach has a good performance. The relation extraction based on HowNet ac-quires the best F value (0.618), which has almost the same effect as manual work, proving the efficiency and feasibility of the method.

Key words: computer applications; relation extraction; HowNet;CiLin;clustering

0 Introduction

With the rapid development of online travel, the Internet has become the largest collection of tourism data, which has moved into the age of big data. Most of the tourism big data are unstructured. These data consist of enormous information related with spatial positions. The starting point of our research is to extract the spatial information (especially spatial relations) from the tourism text. The information will be processed by methods like pattern analysis, data mining and knowledge discovery, and finally serve the development of intelligent travel.

The spatial relation extraction in text is divided into two classes: the approach based on knowledge engineering and the approach based on machine learning. The former artificially sets some patterns of morphology, syntax and semantics according to the features of spatial relation and matches new text with these patterns to realize the spatial relation extraction in sentences. The latter treats the relation extraction as a classification problem, acquires classifier parameters by learning, and completes the extraction target [1-2]. For now, most of the researches on machine learning based extraction focus on the supervised and semi-supervised approaches. However, these approaches can only be used if the entity relations are defined [3]. The tourism domain contains complex spatial relations and each relation can be described by various words, especially in Chinese text. In this case, it is very hard to predefine the relation types. Article [4] realizes tourism spatial relation extraction basing on the entropy. However, this approach can only estimate the existence of a relation, not the specific relation type. Therefore, an unsupervised approach needs to be proposed to realize the automatic extraction of tourism domain spatial relation.

On this basis, this paper proposes a clustering based approach for entity relation extraction. The rest of this paper is organized as follows. Section 1 and section 2 describe the algorithms to extract the spatial words and spatial relation. Section 3 presents the experiments and evaluations. Section 4 gives the conclusion.

Foundations: National Basic Research Program of China (973 Program) 2012CB821200 (2012CB821206); National Natural Science Foundation of China (No. 61320106006)
Brief author introduction:Jin Zhiyao(1990-), Female, Master candidate, Intelligent information processing
Correspondance author: Du Junping, Female, Professor,Artificial Intelligence. E-mail: junpingdu@126.com

1 Spatial words extraction

1.1 Spatial words extraction process

Aside from geographic named entity, the words used to describe spatial relation only contain preposition phrases, nouns of locality, spatial predicates and metaphorical spatial nouns [5]. As there are always lots of distractions like decorative noun or adjective sin real corpus, the sentence should be preprocessed before ex-traction and only keep those relevant words.

If we only use manual generalizations and common thesauruses, it is hard to ex-tract the spatial word efficiently due to its complexity. Therefore, we extracted the spatial words from domain corpus basing on bootstrapping iteration. The process is shown in Fig.1.

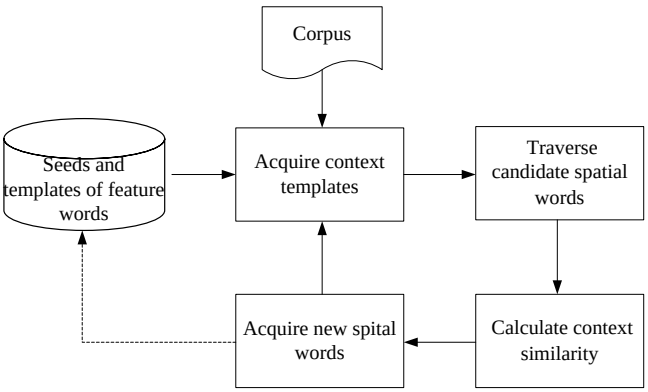

Fig.1 Spatial Words Extraction Process

1.2 Spatial words extraction based on bootstrapping iteration

1.2.1 Context similarity calculating

According to the selective constraint hypothesis, if two words have similar contexts, their semantics are also regarded as similar. Vector Space Model (VSM) is the most common model for context similarity calculating. The formula is:

$$sim(\vec{x},\vec{y}) = \cos(\vec{x},\vec{y}) = \frac{\langle \vec{x},\vec{y} \rangle}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad (1)$$

in which $\vec{x}$、$\vec{y}$ are the context feature vector of two words. The feature item is defined as follows:

$$x_i = 1 - \frac{1}{d_i + 2} \quad (2)$$

in which d_i is the text distance between language unit *i* and the word. The feature item is the smallest language unit in VSM, which can be character, word or GNE. The text distance can be calculated by the displacement between characters or chosen language units. The choice of feature item and distance calculating method can directly affect the context similarity, and then affect the precision and efficiency of the iteration.

Tab.1 Context Feature Vector Building

Word	China	Pingju	Opera	Theater	locates	in	Fengtai	Xiluo
POS	ns	n	n	n	v	p	ns	ns
NER	LOC						LOC	
POS&NER	LOC				v	p	LOC	

Take the case of Tab.1 to illustrate the building of context feature vector. The spatial word in this sentence is "locates". Choose the part of speech and the word displacement as feature item and text distance. The context window is set as 4. The context feature vector is built as: (n-1/2, a-2/4, n-3/4, ns-4/5) and (p-1/2, ns-2/3, ns-3/4). If we choose the Named Entity type as feature item and the character displacement as distance, the vector is: (LOC-1/2) and (LOC-2/3). The above comparison shows that POS is richer than named entity type as the feature item. Meanwhile, named entity type also has some constraint effect for the extraction. Therefore, the two kinds of annotation can be merged into a new feature item as POS&NER. New feature vector is built as: (LOC-1/2) and (p-1/2, LOC-2/3).

1.2.2 Iterating extraction algorithm

Basing on the iterating extraction algorithm raised in [6], we improved the context choosing process. The specific steps of improved algorithm are as follows:

Step 1: Put initial seeds and relevant context templates into the seed list and the seed template list respectively;

Step 2: Traverse the candidate spatial words and acquire their context. Calculate the similarity between the context and every seed template. If a similarity is higher than *ContextSimiThreshold*, put the candidate word into the seed list;

Step 3: Traverse new spatial words and acquire their context. Calculate the similarity between the context and every seed template. If a similarity is lower than *PattemDistictThreshold*, put the context into the seed context list. Repeat Step 2;

Step 4: Stop iterating when spatial words cease to increase.

As the spatial word has limited part of speeches and a relatively low word frequency, we filter a candidate word according to its POS and word frequency. Only the noun, verb and noun of locality with a frequency more than *minFreq* can be reserved.

2 Entity relation extraction based on clustering

After the spatial words are acquired, we used semantic thesauruses to calculate the similarity between words. Then we clustered the feature words basing on hierarchical clustering and realized the automatic extraction of entity relation type. Each cluster means a kind of spatial relation. We used two different semantic thesauruses for similarity calculating: HowNet and Cilin.

2.1 Similarity calculating based on HowNet

HowNet is a common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. There are two main concepts in HowNet: concept and sememe. The concept describes the semantic information of a word, composed of sememes. The similarity of two feature words w_i and w_j is described by their semantic similarity or semantic distance. The similarity $Sim(w_i, w_j)$ is calculated as:

$$Sim(w_i, w_j) = \frac{2*C_{i,j}}{C_i + C_j} \quad (3)$$

in which C_i and C_j are the number of sememes in the concept of w_i and w_j. $C_{i,j}$ is the number of the same sememes between w_i and w_j.

As a word has several concepts in HowNet, we should choose the concepts related with spatial relation to calculate. Define the upper sememes collection SR as {Space, Location, Direction, Part, Boundary, Distance, Situated}. Traverse all of the concepts and estimate whether the sememe of a concept is in SR to choose the spatial relevant concept.

2.2 Similarity calculating based on Cilin(Extended)

Cilin is a common thesaurus of modern Chinese. The entries are sorted as large, middle and small with a tree-like hierarchical structure, which form a semantic classification system from wide concepts to specific meanings of words. Cilin(extended) extends the three-tier structure of Cilin to a five-tier one. In Cilin(extended), the similarity of two words w_i and w_j is reflected by their distance $Dist(w_i, w_j)$ in the tree. The similarity $Sim(w_i, w_j)$ is calculated as:

$$Sim(w_i, w_j) = \frac{1}{Dist(w_i, w_j)+1} \quad (4)$$

3 Experiment results and evaluations

The corpus in this paper is 5000 domain texts crawled from travel websites. Language Technology Platform (LTP) is used to preprocess the corpus.

3.1 Tourism domain spatial words extraction

First, we chose the most frequent spatial word “located” as the seed to determine the parameters. After the experiment, the *PattemDistictThreshold* and *ContextSimiThreshold* are set as 0.4 and 0.9. And we chose 4 as the context window. Then choose three typical spatial words: “west”, a noun of locality; “abut”, a spatial predicate; “foot”, a noun. Take these words as initial seed respectively and all together. The feature items are POS, NER and POS&NER. Besides, filter out the candidate words with a frequency lower than 3 and the words whose POS is not noun, verb or noun of locality. The result is shown in Tab.2.

Tab.2 Result of Spatial Words Iterative Extraction

Seeds	Feature Item	Iteration Times	Number of Spatial Words	Spatial Words Examples	Precision/%
West	POS	6	219	peak[n] join[v]	21.57
	NER	4	1295	scene[n] east[f]	10.38
	POS&NER	5	182	shed[n] pass[v]	29.91
Abut	POS	5	173	station[n] north[f]	16.75
	NER	4	3484	adjoin[v] south[f]	7.83
	POS&NER	5	244	site[n] north[f]	20.15
Foot	POS	4	398	center[n] east[f]	15.82
	NER	4	3013	detour[v] south[f]	10.15
	POS&NER	5	371	border[n] adjoin[v]	20.83
West Abut Foot	POS	4	238	peak[n] north[f]	27.18
	NER	4	1175	slope[n] join[v]	18.97
	POS&NER	5	177	scene[n] flow[v] east[f] around[f]	38.14

Result indicates that the new spatial words acquired always include all three kinds of POS with different seeds and feature items. In three kinds of POS, the noun of locality seed obtains the highest precision. When using all three kinds of words as seeds, the iteration has the highest precision due to the constraint of initial seed template. When choose NER as the feature item, the simple feature vector leads to and less iterating times and more templates and spatial words. Meanwhile, the precision is low. When we choose POS as feature item, the feature vector is more complex and more effective spatial words are acquired. And when POS&NER is used to build a richer feature vector, the precision reaches the highest point.

3.2 Tourism domain spatial relations extraction

Use the feature words acquired in 1.4.1 and other manual complement as the experiment data for the tourism domain spatial relations extraction based on clustering.

3.2.1 Feature words clustering based on HowNet

We clustered the feature words basing on the similarity calculating approach of HowNet. Let each word be a cluster, which means not clustering the feature words and taking this group as the baseline. In addition, let a person sort the feature words manually as the upper bound. The evaluations are two common indicators: purity and F value. Table 3 shows the performance.

Tab.3 Performance of cluster based on HowNet

Cluster Approach		F-measure	Purity
Singleton(baseline)		0.5555	1.0000
Manual		0.7158	0.7719
HowNet	HAC(single-linkage)	0.3841	0.4146
	HAC(complete-linkage)	0.6911	0.8049
	HAC(average-linkage)	0.6028	0.7073

Result shows that both indicators improve after clustering compared with the baseline, which proves the validity and necessity of clustering. The complete-linkage obtains the best F value (0.6911). The less difference with manual sort indicates that this automatic extraction almost acquires the same effect as manual work.

3.2.2 Feature words clustering based on Cilin (Extended)

We clustered the feature words basing on the similarity calculating approach of Cilin (extended). Table 4 shows the performance.

Tab.4 Performance of cluster based on Cilin (extended)

Cluster Approach		F-measure	Purity
Singleton(baseline)		0.5555	1.0000
Manual		0.7158	0.7719
Cilin	HAC(single-linkage)	0.4970	0.7317
	HAC(complete-linkage)	0.6099	0.7561
	HAC(average-linkage)	0.6168	0.7683

Result shows that the average-linkage obtains the best F value (0.6168). However, the performance of clustering based on Cilin is worse than clustering based on HowNet, which results from HowNet's better ability for the similarity calculating of unknown words.

For the three kinds of hierarchical cluster, we made a contrast experiment to find the

influence of different thresholds. The result is shown in Fig.2.

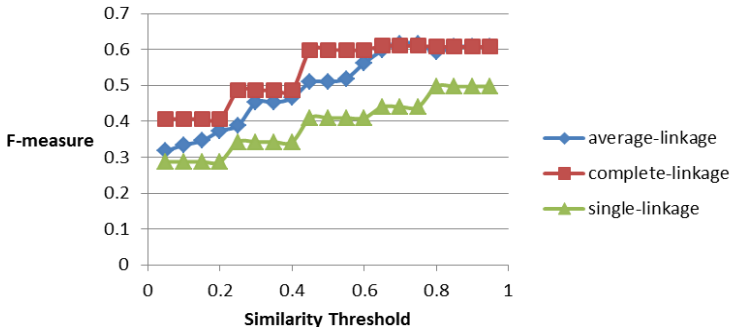

Fig.2 Performance of HAC based on Cilin(extended)

Result indicates that complete-linkage and average-linkage have more stable performance than single-linkage. Both algorithms achieve their best performance when the threshold is 0.75 and average-linkage performs better.

4 Conclusion

To solve the complexity and predefining difficulty of tourism spatial relation, this paper raises an approach based on clustering. We extracted the spatial feature words from domain corpus basing on bootstrapping iteration, clustered these words by semantic similarity and the clustering results are just the spatial relations automatically discovered. Results showed that this approach can effectively realize the automatic extraction of tourism spatial relations. However, we only made simple improvement in semantic thesauruses according to the trait of spatial relation. In the next step of our work, we will make further research on sematic parameters of spatial words to make the approach more suitable for tourism spatial relation extraction.

Acknowledgements (Optional)

This work was supported by the National Basic Research Program of China (973 Program) 2012CB821200 (2012CB821206) and the National Natural Science Foundation of China (No. 61320106006).

References

[1] Zhang C J, Zhang X Y, Jiang W M, et al. Rule-Based Extraction of Spatial Relations in Natu-ral Language Text[A]. In: Proceedings of International Conference on Computational Intelligence and Software Engineering (CiSE)[C]. 2009. 1-4.
[2] Zhang Xueying, Zhang Chunju, Du Chaoli, et al. SVM based Extraction of Spatial Relations in Text[A]. In: Proceedings of the First IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services[C]. 2011. 179-254.
[3] Wang Lifeng. Research on Domain Adaptive Chinese Entity Relation Extraction[D]. Harbin: Harbin Institute of Technology, 2011.
[4] Guo Jianyi, Lei Chunya, Yu Zhengtao, et al. A semi-supervised learning method based on information entropy to extract the domain entity relation[J]. Journal of Shandong University(Engineering Science), 2011, 41(4): 7-12.
[5] Yuan Yecheng, Liu Haijiang, Pei Tao, et al. Spatial Relation Extraction from Chinese Characterized Documents Based on Semantic Knowledge[J]. Journal of Geo-Information Science, 2014,16(5): 681-690.
[6] Jiang Wenming. Research on Spatial Relation Extraction from Chinese Texts[D]. Nanjing: Nanjing Normal University, 2010.

基于聚类的旅游领域实体空间关系抽取研究

靳知瑶，杜军平
（北京邮电大学计算机学院，北京 100876）

摘要：针对旅游领域实体空间关系预先定义困难的问题，本文研究了一种基于聚类的实体空间关系获取方法。首先使用 BootStrapping 方法从领域文本中迭代获取描述实体间空间关系的特征词集合，借助语义词典《知网》和《同义词词林扩展版》计算特征词间的相似度，利用层次聚类等方法对特征词聚类，所得的每一类即为自动发现的一个实体关系类别，从而实现旅游领域实体关系类型的自动获取。实验结果表明，本方法取得了较好的实验性能，其中基于知网的空间关系获取得到了最好的 F 值，为 0.6911，与人工分类的效果相差较小，验证了方法的有效性和可行性。

关键词：计算机应用；关系抽取；知网；同义词词林；聚类
中图分类号：TP391