Geospatial Information and Documents

PAKDD Workshop

May 29 - June 1, 2012
Kuala Lumpur, Malaysia
EDITORIAL

Geographical or spatial information is now included in most of exchanged data. Sometimes, it is directly provided through metadata, but it is very often hidden and it becomes crucial to automatically discover it.

Natural Language Processing (NLP) and Data Mining communities have thus merged their efforts in order to extract geospatial information from textual documents, web pages, field data, and so forth. In this way, recent researches take into account the content of documents (e.g. terms) to identify geospatial data or to predict its geographic location. Nevertheless, spatial information has some specificities that make discovering spatial information and/or spatial correlations from large amount of data still challenging. In this context, some proposals have been focused on the formalization of geospatial concepts and relationships, on the extraction of geospatial relations (e.g. rivers / body of water, town / suburb) in free texts to offer to the database community a unified framework for geodata discovery.

This workshop aims at discussing and assessing some of these strategies, involving NLP or Data Mining techniques, covering all or part of the issues mentioned above.

Topics of interest but not limited to:
- Geospatial information retrieval in documents
- Geospatial knowledge acquisition from documents
- Classification of geospatial documents
- Geospatial analysis of textual data
- Integration of geospatial documents
- Extraction of geospatial information from documents
- Geospatial thesaurus/ontology building from documents
- Quality of extracted geospatial information
- Analysis and integration of geospatial data from web documents
- Visualization of geospatial information from documents

8 papers were submitted in response to the call for papers. The Program Committee who we thank for the quality of their reviewing work, selected 4 papers (2 full papers and 2 short papers). Moreover, we are grateful to Xueying Zhang (Key Laboratory of Virtual Geographic Environment, China) for having accepted our invitation to give a talk entitled Annotation Scheme for Geospatial Relations in Text.

Mathieu Roche and Maguelonne Teisseire
Co-chairs of GeoDoc'2012 Workshop
Program Committee

- Masanori Akiyoshi, Osaka University, Japan
- Torben Bach Pedersen, Aalborg University, Denmark
- Jason Baldridge, University of Texas, USA
- Mete Celik, Erciyes University, Turkey
- Robert Haining, University of Cambridge, UK
- Tahar Kechadi, UCD School of Computer Science and Informatics, Ireland
- Stan Matwin, University of Ottawa, Canada
- Donato Malerba, University of Bari, Italy
- Pascal Poncelet, University of Montpellier 2, France

List of accepted papers

- Full papers
  - The SpatialCIM methodology for spatial document coverage disambiguation and the entity recognition process aided by linguistic techniques,
    *Rosa Nathalie Portugal Vargas, Maria Fernanda Moura, Eduardo Antonio Speranza, Ercilia Rodriguez, and Solange Oliveira Rezende*

  - Evaluating different query reformulation techniques for the Geographical Information Retrieval task considering geospatial entities as textual terms,
    *José M. Perea-Ortega, Miguel A. García-Cumbreras, and L. Alfonso Urena-López*

- Short papers
  - The Relevance of Spatial Relation Terms and Geographical Feature Types,
    *Chunjue Zhang, Xueying Zhang*

  - Syntactic Patterns of Spatial Relations in Text,
    *Shaonan Zhu, Xueying Zhang*
Annotation Scheme for Geospatial Relations in Text

Xueying Zhang

Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, China

Abstract. Natural language describing the nature of peoples internal representation of space is the primary means for exchanging spatial information. The interpretation of geospatial expressions in natural language and the comparison of the semantics of spatial terms in different languages are special arguments for bridging semantic gap between natural language and computational models in geographical information systems (GIS). The understanding of spatial relations in natural language has been addressed as one of the most fundamental issues in GIS. Annotation schemes aim to identify specific spans of text which exhibit linguistic structures, and create metadata describing them to allow computer direct access to the structures. Corpus annotation is a task to provide both reference and training material for method development and benchmark data sets annotated with a given annotation scheme. After analysis of the linguistic characteristics, this paper proposes an annotation scheme for markup linguistic expressions for spatial relations in Chinese text. And then a natural language processing software GATE (General Architecture for Text Engineering) is introduced as the annotation tool. Based on the proposed annotation scheme, a corpus with "Encyclopedia of China Geography" as the source data is annotated by means of cross-validation to solve the problem of annotation inconsistency. Finally, the performance of the annotation scheme and the annotated corpus is evaluated and analyzed, and some related problems are discussed for further improvement and real-world applications. In conclusion, this study provides an annotation scheme and a large-scale benchmark dataset for interpretation of spatial relations in Chinese text. It can help people more in-depth understand the mechanism of geospatial cognition and spatial language, and then improve the intelligence of GIS spatial query, spatial reasoning, geographical information retrieval and geographical information services.
The SpatialCIM methodology for spatial document coverage disambiguation and the entity recognition process aided by linguistic techniques

Rosa Nathalie Portugal Vargas 1, Maria Fernanda Moura 2, Eduardo Antonio Speranza 2, Ercilia Rodriguez 2, Solange Oliveira Rezende 1

1University of São Paulo, Computer Science Department
{nathalie, solange}@icmc.usp.br
2Embrapa Agricultural Information
{fernanda, speranza, ercilia}@cnptia.embrapa.br

Abstract. Nowadays it is becoming more usual for users to take into account the geographical localization of the documents in the retrieval information process. However, the conventional retrieval information systems based on key-word matching do not consider which words can represent geographical entities that are spatially related to other entities in the document. This paper presents the SpatialCIM methodology, which is based on three steps: pre-processing, data expansion and disambiguation. In the pre-processing step, the entity recognition process is carried out with the support of the Rembrandt tool. Additionally, a comparison between the performances regarding the discovery of the location entities in the texts of the Rembrandt tool against the use of a controlled vocabulary corresponding to the Brazilian geographic locations are presented. For the comparison a set of geographic labeled news covering the sugar cane culture in the Portuguese language is used. The results showed a F-measure value increase for the Rembrandt tool from 45% in the non-disambiguated process to 0.50 after disambiguation and from 35% to 38% using the controlled vocabulary. Additionally, the results showed the Rembrandt tool has a minimal amplitude difference between precision and recall, although the controlled vocabulary has always the biggest recall values.

Keywords: Named Entity Recognition and Classification, Toponym Resolution, Ambiguity Problem.

1 Introduction

Internet has become a huge online repository of documents, news and several information. This repository is not easily interpretable, thus, requiring tools and techniques to organize, structure and extract interesting information from its documents. For instance, it is necessary to analyze the data structure, similar characteristics and geographic coverage in order to find new trends, problems and solutions for each geographic zone. It is becoming more usual for users to retrieve the information from the context and then consider the geographic location of the document [1]. For exam-
ple, a user searches the context “Main Artificial Intelligence conferences” and then the geographic localization can be specified as “in the United States”. This kind of search is also applied in the search of documents, for example, “Sugar cane production of the last year in Springfield”. However, conventional information retrieval systems based on key-word matching do not consider that words can represent geographic entities which are spatially related to other entities in the document [2]. For instance, for the previous search, the conventional retrieval system will extract all the documents that include in their content the same words used in the search without considering that the word “Springfield” refers to a geographic localization. This process might return a set of irrelevant documents for the user. It must also be considered the ambiguity in the search, for example, “Springfield” is related to three different cities in the United States. In order to determine and extract geographic features from the recognized entities in text it is necessary the development of systems. In this paper we consider all the names that represent a geographic location as geographic entities, such as, the word “Springfield” in the previous example. Some techniques of spatial data mining can be used to allow the geographic feature extraction. Spatial data mining is the process of discovering interesting patterns from spatial databases which were previously unknown but potentially useful [3]; [4].

Much evidence can be considered in the geographic context definition of documents, such as addresses, postal codes, telephone numbers and names of reference points. After the geographic evidence is identified, the possible semantic ambiguities must be checked. Some gazetteers or geographic information systems (GIS) are commonly used to identify the geographic references. A gazetteer is a catalog of locations or places (dictionary of toponyms) which provides a vocabulary of geographic terms along with their respective locations [5]; [6]. The ambiguity can be understood as an expression of the language (word or phrase) which has many different meanings and can be understood in different ways by a receiver. For example, if the city of Venice is mentioned in the text it is necessary to apply a disambiguation process since it can refer to Italian, Colombian or French cities. This ambiguity problem causes noises in the information retrieval process, therefore, the same term can be related to relevant and irrelevant information [7]. The disambiguation process must be applied to solve the ambiguity problems. The disambiguation processes of entities (toponym1) are responsible for finding out the spatial location in the text through their standardization in a structured representation as geographic coordinates, database entries, or locations with a geographic ontology [8].

In order to perform the disambiguation process, it is first necessary to identify the geographic entities in the text; this process of identification is known as named entity recognition process (NER). The NER in text is a complex task that aims at locating and classifying atomic elements in the text into predefined categories, such as persons, locations, organizations, expressions of time and money, among others. In fact there are some tools that perform the NER task in text. Such as, the Rembrandt linguistic tool [9] that performs the NER task in Portuguese documents. This tool uses

---

1 Toponym is a proper noun designating a place (country, city, continent, etc.)
Wikipedia as the base of a controlled vocabulary as well as set of grammatical rules in order to support the extraction of the entities. A controlled vocabulary is a list of words and phrases used to tag units of information. The Rembrandt tool was developed with the intention of recognizing the entities with a strong connection with geographic places such as countries, cities, rivers, universities, among others.

This paper presents the Spatial Coverage Identification Methodology (SpatialCIM) which allows the identification of the spatial coverage of the text, focusing on news in the Portuguese language. This methodology is based on three steps: pre-processing, data expansion and disambiguation. In the pre-processing step, the Named Entity Recognition and Classification (NERC) process is carried out with the support of the Rembrandt linguistic tool. The performance of the entity recognition process of the documents using the Rembrandt linguistic tool is evaluated against the use of a controlled vocabulary, representing the Brazilian geographic entities.

The set of news used in the experimental setup is related to the sugar cane culture in Brazil. This set of news was manually labeled by a geographic expert. Moreover, the Brazilian sugar mills named in the text and their respective geographic paths are considered as the geographic coverage of documents. The representation of the geographic paths follows the geopolitical hierarchical structure of “Region, State, Meso-Region, Micro-Region, City, Sugar Mill and Category”, according to the territorial division of Brazil, as recognized by the Brazilian Institute of Geography and Statistics (IBGE).

This paper is organized as follows: Section 2 covers related work; Section 3 introduces the proposed methodology to determine the spatial coverage of documents; Section 4 presents the preliminary experiments and obtained results, followed by the final considerations in Section 5.

2 Related Work

After recognizing the entities in the text, the following step is to obtain the geographic paths and solve the ambiguity problems in order to establish the correct geographic classification of the documents (or news, in the context of this work). In the literature, many approaches that solve the ambiguity problems are presented. In this paper, we are interested in exploring the (i) toponym resolution and (ii) ontology approach for the disambiguation process.

2.1 Toponym Resolution Approach

The toponym resolution can be defined as the task of setting a reference to a possible ambiguous place name related to its real localization, which is represented in a given context. The proposed approach geographic features are usually obtained with the help of gazetteers or geographic information systems (GIS). For the work present-

http://www.ibge.gov.br/home/
ed here, the interest is to explore the disambiguation method by means of the toponym resolution.

The work in [10] explores different toponym disambiguation methods based on the word frequency and the associated rankings of each entity. The recognized geographic paths represent a graph structure. This structure aims at helping the disambiguation process. When multiple geographical paths recognized for one entity exist, the longest path in the graph is considered and the others paths are discarded. If there is still ambiguity in the paths a democratic disambiguation is used. The method proposed for the authors is based on the co-occurrence of words in the documents and also considers the child nodes of each geographic path. The relevance is based on the frequency of the words from each node in the paths. Consequently, the most important geographic path is selected as the disambiguated path.

In [11], a conceptual density technique (CD) based in the Word Sense Disambiguation (WSD) for the geographical domain is used. The CD is given by the correlation between the word sense and the document context. The WordNet sub-hierarchies are used to determine the holonymy relationships (part-of relation) of the words. With the use of this hierarchy, it is easier to distinguish among the different types of localizations sharing the same name.

In [12], Leidner uses two minimalistic heuristics. The first one uses the linguistic technique called "one-referent-by-discourse". This technique assumes that a mentioned place name in the discourse refers to the same location in the whole discourse. The second minimalistic heuristic aims at solving the ambiguity problem. When more than one geographic path for a single entity exists, the lowest region formed among all the entities is the one that gives its interpretation.

2.2 The Ontology Approach

This approach is based on the use of geographic ontologies as the base for the geographic concepts and their relationships. Geospatial ontologies define classes and individuals to represent geographic regions, their features, and the relations among them [13]; [14].

The geographic features obtained from the geographic ontology are also explored in [15]. These features are used to discover the relationships among documents considering different geographic zones. The classification of the documents is done considering the geographic features to train different algorithms based on supervised and unsupervised machine learning techniques.

In [8], geographic ontology structures, as well as the structure of event ontology of the documents are considered for the disambiguation process. The authors affirm that the event ontology contributes for a better understanding of the spatial coverage of the documents. If two entities participate in the same event, it is a strong indicator of the geographic relationship between them. This indicator helps in the disambiguation process. The geographic ontology used by the authors consists of four levels (earth, country, state and city). The ontology of events considers the following classes: people, organization, and geopolitical localization. After the identification of the entities,
Wikipedia is used in order to establish the relationships between the event ontology and the geographic ontology.

3 The SpatialCIM Methodology

The SpatialCIM methodology allows the identification and localization of news considering the geopolitical hierarchical structure of the Brazilian territory with the support of linguistic techniques for the process of document entity recognition. The proposed methodology is based on three main steps: (1) pre-processing, (2) data expansion and (3) disambiguation. This paper explores the use of the Rembrandt linguistic tool as well as a Brazilian controlled vocabulary in the (1) pre-processing step in order to analyze which technique has a better performance in the news disambiguation process. After the entities are recognized, an entity selector filter is applied to extract only the geographic entities, as observed in the Fig. 1(a).

![Three stages of the SpatialCIM](image)

Fig. 1. Three stages of the SpatialCIM

After the geographic entities are extracted, step (2) data expansion, begins. It is necessary to use the spatial data base which contains information about the Brazilian geopolitical hierarchy to extract the geographic paths and the geographic coordinates of the entities. In the SpatialCIM, the spatial data base can be populated using a GIS or a geographic ontology. One example of the geographic paths obtained in this phase is presented in Fig. 1(b), in which the paths obey the hierarchy of Brazil, the Brazilian sugar mill, and the international entities. In Fig. 1(b) the entity column represents the geographic entities found in Fig. 1(a). For instance, for entity (a) “Alagoas” was recognized only as one geographic path marked as A1 was recognized. For entity (b) “Belém” four different geographic paths were recognized, whose paths are marked as B1, B2, B3 and B4 respectively. The same process is repeated for the other ambigu-
ous entities. As observed in Fig. 1(b), the extraction of the geographical paths can present ambiguous paths such as the B1, B2, B3 and B4. This ambiguity problem must be solved in order to allow a more efficient document localization. After all the possible geographic paths are extracted, step (3) disambiguation, begins. This process depends on the system used to populate the spatial data base. If a GIS is used, the applied disambiguation method would be the disambiguation by points. If ontology is selected, the applied disambiguation method would be the textual and structural disambiguation. For the disambiguation by points, the geographic coordinates must be extracted from the spatial data base as illustrated in Fig. 1(c).

![Fig. 2. Disambiguation by the Points Method](image)

In order to execute the process of disambiguation by points, all the non-ambiguous entities (Entity A1 and D1) are mapped in the space and a polygon is formed, as illustrated in Fig. 2(a). Fig. 2(b) presents the disambiguation process for entity “B”. This process considers the polygon formed by the non-ambiguous entities showed in Fig. 2(a). Then, the ambiguous entities (B1, B2, B3, and B4) are mapped in the space and the entity that belongs to the polygon or the entity is closest to any points of the polygon is selected. In Fig. 2(b), entity B4 is the disambiguated entity since it belongs to the formed polygon which is marked by a rectangle. This entity is now considered a non-ambiguous entity and a new polygon is formed considering the entities A1, D1 and B4 as demonstrated in Fig. 2(c). After a new polygon is formed the process is repeated until there are no more ambiguous entities as presented in Fig. 2(c) and 2(d).

In the disambiguation process, some problems might occur, such as, the non-existence of non-ambiguous entities at the beginning, or the tie between two ambiguous entities. Next there some cases as well as the solutions proposed are explained.

- If two ambiguous entities belong to the formed polygon, the closest entity to any point of the polygon is considered as correct.
- If the entities do not belong to the polygon and have the same distance to any point of the polygon, then an entity is selected randomly.
If all the recognized entities are ambiguous, some of the heuristics explained next are used in order to determine the most important entity among all the entities. When the first entity is selected, the disambiguation process continues as previously explained.

The heuristics used to determine the importance of the entities in this paper, considering the experimental set up, are listed next in order of importance:

- In this research the sugar mill entities are more important than other entities.
- As the geographic path considers the hierarchical structure of Brazil, the geographic paths of the entities that are more generic are considered. For example, the entity “São Paulo” is recognized as city and as state. In this case the “state of São Paulo” is selected since it represents a more general path.

At the end of the process, all the documents are marked with their respective associated geographic paths.

4 Experiments and Results

The collection used in the experimental set up is a set of news in the Portuguese language focused on the sugar cane culture. The set of news was marked by an Embrapa\textsuperscript{3} expert and considers all the geographic paths associated to each piece of news. The set of documents analyzed in this paper is formed by 237 documents, with 593 recognized entities by the expert, as well as approximately 350 words by document. In the set of news, it is considered that one document can have multiple geographic paths associated.

The main objective of these experiments is to compare the performance of the disambiguation by points using the recognized entities obtained with the help of the Rembrandt linguistic tool and a Brazilian controlled vocabulary. The Rembrandt tool uses linguistics processes in order to determine all the entities in the text. The Rembrandt can automatically recognize entities of the types person, location, time and organization. The Brazilian controlled vocabulary is formed by the intersection of a list with the Brazilian geographic entities and the geographic entities founds in the documents. The process performed by the Rembrandt tool can be computationally expensive since it needs to analyze each phrase and recognize the entities, as well as applying the sense disambiguation. This kind of disambiguation is used when trying to determine the type of entity. For example, the “Paris” entity can be understood as a city, organization, or person entity. If the document is too large, the Rembrandt tool can take considerable time to determine the entities. Considering that linguistic processes are computationally expensive we experimented the use of the controlled vocabulary as a non-linguistic alternative.

To obtain the controlled vocabulary, a list of all the Brazilian geographic entities was considered. One problem with this technique is that the context of the words can-

\textsuperscript{3} Embrapa: The Brazilian Agricultural Research Corporation (http://www.embrapa.br/english)
not be recognized; consequently, it can take several types of ambiguities. For example, if the term “São Paulo” is in the text, the controlled vocabulary will recognize this term as a geographic entity without considering the document context. In the document, the term “São Paulo” could be a reference to the São Paulo football team or the state of São Paulo or the São Paulo city.

To achieve the experimental objectives, the geographic paths automatically obtained were compared to the geographic paths marked by an expert. The paths obtained with the entity recognition carried out by the Rembrandt tool and the controlled vocabulary were also compared, in order to determine which technique achieves better geographic paths compared with those obtained by the expert or if the process has no significant differences. To evaluate how much the disambiguation process contributes to better news localization, the resulting geographic paths for each of the news are compared with the geographic paths marked by the expert. Table 1 illustrates the geographic paths of the news marked by the expert. In this case, labels given by the expert are considered as the correct ones for the geospatial classification of the documents. In Table 2, the geographic paths automatically obtained with the entity recognition provided by the Rembrandt linguistic tool can be observed. Observe that two ambiguous geographic paths for the Micro-region of “Catanduva” (rows L1 and L2) and “Jau” (rows L3 and L4) were recognized.

Table 1. Geographic Paths marked by an expert at news 41755

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>41755</td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>L1</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São José</td>
<td>São Paulo</td>
<td>Rio Preto</td>
<td>Catanduva</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>Ribeirão Preto</td>
<td>São Paulo</td>
<td>Preto</td>
<td>-</td>
<td>-</td>
<td>L3</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>Piracicaba</td>
<td>São Paulo</td>
<td>Piracicaba</td>
<td>-</td>
<td>-</td>
<td>L4</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>Bauru</td>
<td>São Paulo</td>
<td>Jau</td>
<td>-</td>
<td>-</td>
<td>L5</td>
</tr>
</tbody>
</table>

Comparing these results with the results marked by the expert, it can be observed that the “São Paulo”, “Ribeirão Preto” and “Piracicaba” entities (rows L1, L3 and L4 in Table 1) were not recognized. Table 3 illustrates the geographic paths automatically obtained with the entity recognition provided by the controlled vocabulary presented in Table 1.
Table 2. Geographic Paths Generated by SpatialCIM as well as with the Entity Recognition Provided by the Rembrandt tool at news 41755

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>41755</td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>São Jose</td>
<td>Catanduva</td>
<td>Catanduva</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>São Jose</td>
<td>Catanduva</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Bauru</td>
<td>Jau</td>
<td>Jau</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Bauru</td>
<td>Jau</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3 presents ambiguous geographic paths for the “Catanduva” (rows L1 and L2), “Jau” (rows L3 and L4), “Piracicaba” (rows L5, L6 and L7), “Ribeirão Preto” (rows L9, L10 and L11) and “São Paulo” (rows L12, L13 and L14) entities. Observe that all the entities marked by the expert, in Table 1, were recognized in Table 3. Due to the evidence of ambiguity presented in Table 2 and 3, the disambiguation by points process is applied as detailed in Section 3. Table 2 keeps the L2 and L4 rows and the Table 3 keeps the L2, L4, L5, L8, L9 and L14 rows.

Table 3. Geographic Paths Generated by SpatialCIM as well as with the Entity Recognition Provided by the Controlled Vocabulary at news 41755.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>41755</td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>São Jose</td>
<td>Catanduva</td>
<td>Catanduva</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>São Jose</td>
<td>Catanduva</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Bauru</td>
<td>Jau</td>
<td>Jau</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Bauru</td>
<td>Jau</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Piracicaba</td>
<td>Piracicaba</td>
<td>Piracicaba</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Piracicaba</td>
<td>Piracicaba</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Piracicaba</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Ribeirão Preto</td>
<td>Ribeirão Preto</td>
<td>Rib. Preto</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Ribeirão Preto</td>
<td>Ribeirão Preto</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Ribeirão Preto</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Met. São Paulo</td>
<td>São Paulo</td>
<td>São Paulo</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>Met. São Paulo</td>
<td>São Paulo</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sudeste</td>
<td>São Paulo</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

As observed, the disambiguation by points and the recognition of entities made with the controlled vocabulary (presented in Table 3) presents a significant number of correct entities (presented in Table 1). The disambiguated results in Table 3 are close to the entities marked by the expert and presented in Table 1.
Table 4. Correct recognized entities with the Rembrandt tool and the controlled vocabulary considering the disambiguated and the non-disambiguated geographic paths

<table>
<thead>
<tr>
<th></th>
<th>Disambiguated Geographic paths</th>
<th>Non-disambiguated Geographic paths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Rembrandt tool</td>
<td>0.62</td>
<td>0.42</td>
</tr>
<tr>
<td>Brazilian controlled vocabulary</td>
<td>0.73</td>
<td>0.26</td>
</tr>
</tbody>
</table>

On the other hand, in Table 4, the general result of the precision, recall and F-measure of the methods for the complete news collection is showed. It can be observed that the Rembrandt tool gives a larger F-measure value of 0.50 over 0.38 obtained by the controlled vocabulary in the disambiguated process. Although, the controlled vocabulary always presents bigger recall values it fails in the precision values. Additionally, the disambiguation process improves the results for the Rembrandt tool as well as controlled vocabulary for both the recall and the precision measures.

Fig. 3. Recall measure of the Rembrandt tool and the Controlled Vocabulary for the disambiguated and non-disambiguated process.

The presented experiments show the results of the comparison between the Rembrandt tool and the controlled vocabulary for the achievement and disambiguation of the geographic paths.
The results presented in Fig.3 and Fig.4 show the distribution of the recall and precision according to tolerance thresholds. These results also confirm that Rembrandt tool has minimal amplitude difference between recall and precision measure, while recall of the controlled vocabulary is always higher and precision falls. Consequently, the Rembrandt tool is able to obtain a higher F-measure value than the controlled vocabulary.

5 Conclusions

In this paper, the SpatialCIM methodology was detailed which can determine the spatial coverage of documents. Additionally, a comparison between the Rembrandt tool and the use of a controlled vocabulary for the process of recognition of the geographic entities is presented. This comparison was performed in order to evaluate if there was a significant advantage in using the linguistic tool when we also have a controlled vocabulary. For the experimental set up, a set of news in Portuguese about the sugar cane culture, and previously marked by an Embrapa expert was used. Finally, the disambiguation method by points with the support of a GIS and the table of the Brazilian geopolitical division obtained from IBGE was also explored.

The experiments showed a larger F-measure value for the geographic entity recognition process with the use of the Rembrandt tool. It also showed that the disambiguation process actually improves the correct recognition of the entities and it works with a lower number of geographic paths. In future work we intend to analyze the textual and structural disambiguation method with the support of geographic ontologies.

Acknowledgements We’d like to thank the University of São Paulo, Computer Science Department; Embrapa Agricultural Information; CNPQ and FAPESP by the support.
References

Evaluating different query reformulation techniques for the Geographical Information Retrieval task considering geospatial entities as textual terms

José M. Perea-Ortega¹, Miguel A. García-Cumberras², and L. Alfonso Ureña-López²

¹ Languages and Information Systems Department, University of Sevilla
E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, 41012, Sevilla (Spain)
jmperea@us.es

² Computer Science Department, University of Jaén
Escuela Politécnica Superior, Campus Las Lagunillas s/n, 23071, Jaén (Spain)
{magc,laurena}@ujaen.es

Abstract. Geographic Information Retrieval (GIR) is an active and growing research area that focuses on the retrieval of textual documents according to a geographical criteria of relevance. However, since a GIR system can be treated as a traditional Information Retrieval (IR) system, it is important to pay attention to finding effective methods for query reformulation. In this way, the search results will improve their quality and recall. In this paper, we propose different Natural Language Processing (NLP) techniques of query reformulation related to the modification and/or expansion of both parts thematic and geospatial that are usually recognized in a geographical query. We have evaluated each of the reformulations proposed using GeoCLEF as an evaluation framework for GIR systems. The results obtained show that all proposed query reformulations retrieved relevant documents that were not retrieved using the original query.

Keywords: Geographic query reformulation, Geographic Information Retrieval, GeoCLEF

1 Introduction

In the Information Retrieval (IR) field [2], the approach based on the modification of the user query to improve the quality of the IR results is known as query reformulation. The aim of such process is to satisfy the user information need, usually improving the quality and recall of the results obtained using the original user query. This feature is explicitly supported by some search engines suggesting related queries or providing different completions of the initial user query. Moreover, other search engines also support query reformulation in an implicit manner, by expanding the original query with terms related to their keywords, for example.
Geographic Information Retrieval (GIR) is an active and growing research area that focuses on the retrieval of textual documents according to a geographical criteria of relevance. For this reason, GIR is considered as an extension of the field of IR. Specifically, GIR is concerned with improving the quality of geographically-specific information retrieval, focusing on access to unstructured documents [10, 13]. The IR community has primarily been responsible for research in the GIR field, rather than the Geographic Information Systems (GIS) community. The type of query in a IR engine is based usually on natural language, in contrast to the more formal approach common in GIS, where specific geo-referenced objects are retrieved from a structured database. In a GIR system, a geographic query can be structured as a triplet of \(<theme><spatial relationship><location>\), where \(<theme>\) is the main subject of the query, \(<location>\) represents the geographical scope of the query and \(<spatial relationship>\) determines the relationship between the subject and the geographical scope. For example, the triplet for the geographical query “airplane crashes close to Russian cities” would be \(<airplane crashes><close to><Russian cities>\). Thus, a search for “castles in Spain” should return not only documents that contain the word “castle”, but also those documents which have some geographical entity related to Spain.

Since a GIR system can be treated as a traditional search engine (the results for a query are displayed as a ranked list), it is important to pay attention to finding effective methods for query reformulation. These methods can take into account both lexical-syntactic features and geographical aspects. In this way, the search results will improve their quality and recall. The objective of this paper is to evaluate several geographic query reformulations for the GIR task, considering that a GIR system can perform as a IR system. To carry out this evaluation, we have used the most important evaluation framework in this context: GeoCLEF\(^3\) [7, 14].

The remainder of this paper is structured as follows: in Section 2, the most important works related to the geographic query reformulation in GIR are expounded; in Section 3, we describe the GIR system used for the experiments; in Section 4, we describe briefly the evaluation framework; in Section 5, the experiments carried out and an analysis of the results are presented; finally, in Section 6, we draw some conclusions and future work is expounded.

2 Related work

Jansen et al. [9] define the concept of query reformulation as the process of altering a given query in order to improve search or retrieval performance. Sometimes, query reformulation is applied automatically by search engines as with relevance feedback technique. It is a method that allows users to judge whether a document is relevant or not, so that automatic rewritings can be generated depending on it. At other times, query reformulation is carried out analysing

\(^3\) http://ir.shef.ac.uk/geoclef/
the top retrieved documents without the user’s intervention, taking into account term statistics. However, it has been found that users rarely utilize the relevance feedback options [19] and usually reformulate their needs manually [1].

The focus of this paper is geographic queries. According to Gravano [8], search engines are criticized because of their ignorance to the geographical constraints on users’ queries and, therefore, retrieve less relevant results. This could be attributed to the way search engines handle queries in general as they adopt a keywords matching approach without spatially inferring the scope of the geographic terms. However, it shall be noted that a number of services to deal with this issue have recently been proposed in major search engines, but not in the general purpose tools.

Several authors have studied what users are looking for when submitting geographic queries [18, 6, 11]. One of the main conclusions of these studies is that the structure of geographic queries consists of thematic and geographical parts, with the geo-part occasionally containing spatial or directional terms. From a geographical point of view, Kohler [12] provides a research about geo-reformulation of queries. She concludes that the addition of more geo terms in the query is commonly used to differentiate between places that share the same name. This is also known as query expansion using geographic entities.

In the literature, we can find various works that have addressed the spatial query expansion. Cardoso et al. [4] present an approach for geographical query expansion based on the use of feature types, readjusting the expansion strategy according to the semantics of the query. Fu et al. [5] propose an ontology-based spatial query expansion method that supports retrieval of documents that are considered to be spatially relevant. They improve search results when a query involves a fuzzy spatial relationship, showing that proposed method works efficiently using realistic ontologies in a distributed spatial search environment. Buscaldi et al. [3] use WordNet during the indexing phase by adding the synonyms and the holonyms of the encountered geographical entities to each documents index terms, proving that such method is effective. Finally, Stokes et al. [20] conclude that significant gains in GIR will only be made if all query concepts (not just geospatial ones) are expanded.

Our work could be positioned within those works that treat geographical part as textual terms, i.e., from a Natural Language Processing (NLP) point of view, exclusively. For this reason, the proposed query reformulations are based on expansions of the thematic and geographical parts detected in a geographical query, using synonyms and geospatial terms related with the keywords and geographical entities found in the query.

3 The SINAI-GIR architecture

In this Section we describe an example of a GIR system. Specifically, we have used our own GIR system called SINAI-GIR [17]. GIR systems are usually composed of three main stages: preprocessing of the document collection and queries,
textual-geographical indexing and searching and, finally, reranking of the retrieved results using a particular relevance formula that combines textual and geographical similarity between the query and the document retrieved. The GIR system used in this work follows a similar approach, as can be seen in Figure 1.

With respect to the document collection processing, it is based on detecting all the geographical entities in each document and generating a geo-index with them. In this phase, the stop words are removed and the stem of each word is taken into account. We have used our own Named Entity Recognition (NER) tool to detect geographical entities. It is called GeoNER [16] and it is based on external knowledge resources such as GeoNames\(^5\) and Wikipedia.

Regarding query processing, each query is preprocessed and analyzed, identifying the geographical scope and the spatial relationship that may contain. It also involves specifying the triplet explained in Section 1, which will be used later during the filtering and reranking process. To detect such triplet, we have used a Part Of Speech tagger (POS tagger) like TreeTagger\(^6\), taking into account some lexical syntactic rules such as preposition + proper noun, for example. Moreover, the stop words are removed and the Snowball stemmer\(^7\) is applied to each word of the query, except for the geographical entities. During the text retrieval process, we obtain 1,000 documents for each query. We have used Terrier\(^8\) as a search engine. According to a previous work [15], it was shown that Terrier is

\(^5\) GeoNames is a geographical database covers all countries and contains over eight million placenames that are available for download free of charge. [http://www.geonames.org](http://www.geonames.org)

\(^6\) TreeTagger v.3.2 for Linux. Available in [http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html](http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html)

\(^7\) Available in [http://snowball.tartarus.org](http://snowball.tartarus.org)

\(^8\) Version 2.2.1, available in [http://terrier.org](http://terrier.org)
one of the most used IR tools in IR systems in general and GIR systems in particular, obtaining promising results. The weighting scheme used has been \textit{inL2}, which is implemented by default in Terrier. This scheme is the Inverse Document Frequency (IDF) model with Laplace after-effect and normalization two. As a final step, each preprocessed query (including their geographical entities) is run against the search engine. The retrieved documents are filtered and reranked, setting in the last positions those documents that do not match with the geographical scope detected in the query. By contrast, those documents that fit the geographical scope detected, are set in the first positions.

Although GIR systems usually apply a geo-reranking process after the IR module, it is important to note that it is not necessary in this work particularly, because we are interested in evaluating the precision and recall of each query reformulation proposed from an IR point of view.

4 GeoCLEF: the evaluation framework

In order to evaluate the proposed query reformulations, we have used the GeoCLEF framework [7, 14], an evaluation forum for GIR systems held between 2005 and 2008 under the CLEF$^9$ conferences. GeoCLEF provides a document collection that consists of 169,477 documents, composed of stories and newswires from the British newspaper \textit{Glasgow Herald} (1995) and the American newspaper \textit{Los Angeles Times} (1994), representing a wide variety of geographical regions and places. On the other hand, there are a total of 100 textual queries or topics provided by GeoCLEF organizers (25 per year). They are composed of three main fields: \textit{title} (T), \textit{description} (D) and \textit{narrative} (N). For the experiments carried out in this work, we have only taken into account the \textit{title} field. Some examples of GeoCLEF topics are: “vegetable exporters of Europe”, “forest fires in north of Portugal”, “airplane crashes close to Russian cities” or “natural disasters in the Western USA”.

Regarding the evaluation measures used, results are evaluated using the relevance judgements provided by the GeoCLEF organizers and the TREC evaluation method. The evaluation has been accomplished by using the Mean Average Precision (MAP), Recall (R) and Precision at \(n\) (P@$n$). The MAP measure computes the average precision over all queries. The average precision is defined as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved. Recall is a measure of the extent to which relevant documents are found or retrieved. Recall is 1.0 when every relevant document is retrieved. Finally, Precision at \(n\) is the precision at the number of \(n\) relevant documents in the collection for the query. Precision is the fraction of the relevant documents divided by the total number of documents retrieved. Therefore, if P@$n$ is 1.0, it means a perfect relevance ranking and a perfect recall at \(n\) documents retrieved.

\textsuperscript{9}http://www.clef-initiative.eu/
5 Experiments and results

Several query reformulations for the GIR task are analyzed in this work. They try to use both the thematic part and the geographical scope detected in the query. The objective of these query reformulations is to improve the retrieval process trying to find relevant documents that are not retrieved using the original query. Starting from the preprocessed original query, we have generated the following query reformulations:

- **QR1**: the geographical scope is removed, leaving only the thematic part of the original query.
- **QR2**: the thematic part is expanded, repeating its terms. In this way, we try to give more importance to the thematic part than the geo-part.
- **QR3**: the thematic part is expanded using only synonyms of the keywords detected in the thematic part of the query. We have considered as keywords the nouns recognized in such part.
- **QR4**: the geographical part is expanded using only synonyms of the geographical scope detected in the query.
- **QR5**: the geographical part is expanded using locations or places that match with the geographical scope and the spatial relationship detected in the query.
- **QR6**: the thematic and geographical parts are expanded, combining the QR3 and QR5 reformulations.

<table>
<thead>
<tr>
<th>Reformulation</th>
<th>Text of the query</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>visit American presid Germany</td>
</tr>
<tr>
<td>QR1</td>
<td>visit American presid</td>
</tr>
<tr>
<td>QR2</td>
<td>visit American presid visit American presid Germany</td>
</tr>
<tr>
<td>QR3</td>
<td>#and(#or(visit meet stay ) American presid Germany)</td>
</tr>
<tr>
<td>QR4</td>
<td>#and(visit American presid #or(Germany #3(Federal Republic of Germany) Deutschland FRG ))</td>
</tr>
<tr>
<td>QR5</td>
<td>#and(visit American presid #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen ))</td>
</tr>
<tr>
<td>QR6</td>
<td>#and(#or(visit meet stay ) of the American presid ) #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen )</td>
</tr>
</tbody>
</table>

Table 1. Example of query reformulations generated for the query “Visits of the American president to Germany”

Table 1 shows an example of the different query reformulations generated for the query “Visits of the American president to Germany”. As can be seen, QR2 and QR3 are query reformulations that expand only the thematic part of the queries and, on the other hand, QR4 and QR5 expand only the geographical
part of them. Finally, QR6 can be considered a combination of expansions using both parts.

The different results obtained using each query reformulation (QR) along with the result obtained using the original query are shown in Table 2. In such table, we show the average score of precision at the 5, 10 and 100 first documents retrieved, recall (R) and MAP for each query reformulation proposed. Although none of the proposed QRs improve the MAP score obtained using the original query, it is interesting to note that QR2 (the thematic part is expanded, repeating its terms) achieves the best P@10 score in three of the four topic sets.

<table>
<thead>
<tr>
<th>Topic Set</th>
<th>QR</th>
<th>P@5</th>
<th>P@10</th>
<th>P@100</th>
<th>R</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>original</td>
<td>0.5520</td>
<td>0.4560</td>
<td>0.1904</td>
<td>0.8364</td>
<td>0.3514</td>
</tr>
<tr>
<td></td>
<td>QR1</td>
<td>0.2640</td>
<td>0.2560</td>
<td>0.1260</td>
<td>0.6748</td>
<td>0.1638</td>
</tr>
<tr>
<td></td>
<td>QR2</td>
<td>0.5200</td>
<td>0.4920</td>
<td>0.1840</td>
<td>0.8276</td>
<td>0.3353</td>
</tr>
<tr>
<td></td>
<td>QR3</td>
<td>0.3680</td>
<td>0.3160</td>
<td>0.1400</td>
<td>0.7596</td>
<td>0.2035</td>
</tr>
<tr>
<td></td>
<td>QR4</td>
<td>0.3120</td>
<td>0.2800</td>
<td>0.1212</td>
<td>0.6552</td>
<td>0.2242</td>
</tr>
<tr>
<td></td>
<td>QR5</td>
<td>0.1440</td>
<td>0.1240</td>
<td>0.0772</td>
<td>0.5624</td>
<td>0.0952</td>
</tr>
<tr>
<td></td>
<td>QR6</td>
<td>0.1600</td>
<td>0.1480</td>
<td>0.0780</td>
<td>0.5692</td>
<td>0.0942</td>
</tr>
<tr>
<td>2006</td>
<td>original</td>
<td>0.2400</td>
<td>0.1920</td>
<td>0.0716</td>
<td>0.7288</td>
<td>0.2396</td>
</tr>
<tr>
<td></td>
<td>QR1</td>
<td>0.0560</td>
<td>0.0640</td>
<td>0.0252</td>
<td>0.4604</td>
<td>0.0615</td>
</tr>
<tr>
<td></td>
<td>QR2</td>
<td>0.2320</td>
<td>0.2040</td>
<td>0.0664</td>
<td>0.6796</td>
<td>0.2314</td>
</tr>
<tr>
<td></td>
<td>QR3</td>
<td>0.1440</td>
<td>0.1400</td>
<td>0.0604</td>
<td>0.7356</td>
<td>0.1419</td>
</tr>
<tr>
<td></td>
<td>QR4</td>
<td>0.1920</td>
<td>0.1720</td>
<td>0.0636</td>
<td>0.6984</td>
<td>0.2064</td>
</tr>
<tr>
<td></td>
<td>QR5</td>
<td>0.2240</td>
<td>0.1840</td>
<td>0.0612</td>
<td>0.6524</td>
<td>0.1811</td>
</tr>
<tr>
<td></td>
<td>QR6</td>
<td>0.1840</td>
<td>0.1760</td>
<td>0.0580</td>
<td>0.6772</td>
<td>0.1486</td>
</tr>
<tr>
<td>2007</td>
<td>original</td>
<td>0.3040</td>
<td>0.2560</td>
<td>0.1188</td>
<td>0.7156</td>
<td>0.2311</td>
</tr>
<tr>
<td></td>
<td>QR1</td>
<td>0.1600</td>
<td>0.1320</td>
<td>0.0796</td>
<td>0.4452</td>
<td>0.1255</td>
</tr>
<tr>
<td></td>
<td>QR2</td>
<td>0.2640</td>
<td>0.2120</td>
<td>0.1072</td>
<td>0.6656</td>
<td>0.1871</td>
</tr>
<tr>
<td></td>
<td>QR3</td>
<td>0.2000</td>
<td>0.1800</td>
<td>0.0884</td>
<td>0.6284</td>
<td>0.1774</td>
</tr>
<tr>
<td></td>
<td>QR4</td>
<td>0.2160</td>
<td>0.2000</td>
<td>0.1020</td>
<td>0.6608</td>
<td>0.1687</td>
</tr>
<tr>
<td></td>
<td>QR5</td>
<td>0.2240</td>
<td>0.2000</td>
<td>0.0928</td>
<td>0.6720</td>
<td>0.1874</td>
</tr>
<tr>
<td></td>
<td>QR6</td>
<td>0.2240</td>
<td>0.2040</td>
<td>0.0836</td>
<td>0.6344</td>
<td>0.1763</td>
</tr>
<tr>
<td>2008</td>
<td>original</td>
<td>0.3760</td>
<td>0.2680</td>
<td>0.1104</td>
<td>0.7368</td>
<td>0.2484</td>
</tr>
<tr>
<td></td>
<td>QR1</td>
<td>0.1760</td>
<td>0.1400</td>
<td>0.0928</td>
<td>0.5996</td>
<td>0.1301</td>
</tr>
<tr>
<td></td>
<td>QR2</td>
<td>0.3440</td>
<td>0.2680</td>
<td>0.1124</td>
<td>0.7196</td>
<td>0.2381</td>
</tr>
<tr>
<td></td>
<td>QR3</td>
<td>0.2960</td>
<td>0.2320</td>
<td>0.1024</td>
<td>0.6884</td>
<td>0.1972</td>
</tr>
<tr>
<td></td>
<td>QR4</td>
<td>0.2640</td>
<td>0.1960</td>
<td>0.0924</td>
<td>0.6404</td>
<td>0.1619</td>
</tr>
<tr>
<td></td>
<td>QR5</td>
<td>0.2720</td>
<td>0.2040</td>
<td>0.0964</td>
<td>0.6984</td>
<td>0.1906</td>
</tr>
<tr>
<td></td>
<td>QR6</td>
<td>0.2720</td>
<td>0.2280</td>
<td>0.0948</td>
<td>0.7028</td>
<td>0.2028</td>
</tr>
</tbody>
</table>

Table 2. Evaluation results obtained for each query reformulation proposed.

At this point, we wonder if the QRs proposed were really retrieving relevant documents that the original query was not retrieving. Using the relevance judgements provided by the GeoCLEF organizers, we get the relevant documents retrieved by each QR that were not retrieved by the original query, as shown.
in Figure 2. The total number of documents retrieved was always 1,000. On the other hand, according to the relevance judgements, the total number of relevant documents for each topic set (2005, 2006, 2007 and 2008) is 1,028, 378, 650 and 747, respectively. Moreover, the number of relevant documents retrieved by the original query was 908, 284, 543 and 588 for the topic sets 2005, 2006, 2007 and 2008, respectively.

Analyzing these results, we can observe that all proposed query reformulations always retrieve relevant documents that are not retrieved using the original query. It is interesting to note the behaviour in general of the reformulations related to the geographical expansion (QR4 and QR5). Specifically, QR5 achieves a remarkable difference using the 2008 topic set with a total of 68 relevant documents not retrieved by the original query. In fact, this means that of 159 relevant documents not retrieved by the original query using the 2008 topic set (747-588), 42.77% of them are retrieved using the QR5 reformulation. Another example occurs with QR4 that obtains the highest value for the 2007 topic set, retrieving 32.71% of the relevant documents not retrieved using the original query. On the other hand, the reformulations related to the thematic expansion (QR2 and QR3) also achieve good results in general, as can be seen for the 2005 and 2006
topic sets. All this makes that the reformulation that combines the QR3 and QR5 reformulations (QR6) also obtain good results, as shown for all topic sets. Finally, QR1 achieves the best score for the 2005 topic set, so the idea of removing the geographical part in the original query can sometimes be a good strategy.

6 Conclusions and further work

In this paper we propose different NLP techniques of query reformulation related to the modification and/or expansion of both parts thematic and geospatial that are usually recognized in a geographical query. We have evaluated each of the reformulations proposed using GeoCLEF as an evaluation framework for GIR systems. This evaluation has been carried out from an IR point of view, that is, without taking into account any geo-reranking procedure after the retrieval process. The results obtained show that all proposed query reformulations retrieved relevant documents that were not retrieved using the original query. Therefore, for future work, we will analyze the different types of geographical queries and then we will study in depth when is more suitable to apply these techniques in a GIR system depending on the type of the query.

Acknowledgments

This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. It has been also partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) through the TEXT-COOL 2.0 project (TIN2009-13391-C04-02) from the Spanish Government, and a grant from the Instituto de Estudios Giennenses through the Geocaching Urbano project (RFC/IEG2010).

References

The Relevance of Spatial Relation Terms and Geographical Feature Types

Chunju Zhang     Xueying Zhang

Key Laboratory of Virtual Geography Environment (Nanjing Normal University), MOE, Nanjing, China
zcjtwz@sina.com

Abstract: Spatial relation terms can generally indicate spatial relations described in natural language context. Their semantic representation is closely related to geographical entities and their characteristics e.g. geometry, scale and geographical feature types. This paper proposes a quantitative approach to explore the semantic relevance of spatial relation terms and geographical feature types of geographical entities in text. Firstly, a classification of spatial relation terms is performed. Secondly, the “Overlap” similarity measure is introduced to define the relevance of spatial relation terms and geographical feature types based on a large scale annotation corpus. Finally, a knowledge base based on protégé is developed to formally represent and visualize geographical feature types, spatial relation classifications, and the relevance of spatial relation terms and geographical feature types. This study indicates that spatial relation terms are strongly relevant to geographical feature types. The semantic representation of topological relation terms is diverse and their relevance with geographical feature types is much stronger than directional relation and distance relation terms, but the annotation quality and the classification granularity of geographical entities in the corpus have a great effect on the performance.

Keywords: spatial relation; geographical feature type; spatial relation term; relevance

1 Introduction

Natural language describes the nature of people’s internal representation of space and is the primary means for representation and exchange of geographical information, such as geographical entities, spatial relations, etc. Spatial relations are the associations or connections between different real world features, and play an important role in spatial data modeling, spatial query, spatial analysis, spatial reasoning,
and map comprehension [1]. The semantic research of spatial relations is the premise and basis for the description and expression of spatial relations. Spatial relations have been in a high priority in many research fields, such as linguistics, cognitive science, GIS and spatial reasoning. The linguistics field focus on the words, lexical, syntactic and semantic structure of spatial relation expressions, and the relationship with human’s spatial cognition [2][3]. In recent years, spatial relations in natural language have become a hot topic of geographical information science. Mark [4] and Lautenschütz [5] investigated the influence of geometry and scale characteristics, spatial relation types and geographical feature types on human’s chosen of spatial relation terms by questionnaire method, and then Mark[6] made a further research on the mapping between spatial relation terms and GIS computational model. Shariff [7] and Xu [8] summarized the knowledge rules of different geographical feature types and spatial relation vocabularies to construct a semantic mapping model of spatial relation terms. Du and Wang [9] explored the formal expression of GIS querying sentences described in a restricted syntactic pattern of spatial relation descriptions in natural language.

Spatial relation terms can indicate spatial relations described in natural language context. Different from the early models of spatial relations which focused on the geometry, it is now widely recognized that the semantic meaning of spatial relation terms is also dependent on functional and pragmatic features in situated context [5]. Their semantic descriptions in natural language are closely related to geographical entities and their characteristics of geometry, scale and geographical feature types. Especially, some spatial relation terms can be used for several different geographical feature types, while some are just for a certain geographical feature type. For example, the spatial relation term of watershed is used to indicate the junction between mountains and waters, and can’t describe geographical entities of other geographical feature types. This paper proposes a quantitative approach to explore the relevance of spatial relation terms and geographical feature types from text corpuses. Properly understanding the semantic meaning of spatial relation terms in text will improve geographical information retrieval, GIS natural language query, extraction of spatial relations from text, and qualitative spatial reasoning.

The reminder of this paper is structured as follows: Section 2 investigates the basic categories of spatial relations in natural language, and the classification of spatial relation terms. Section 3 discusses the calculation of relevance of spatial relation terms and geographical feature types, and the semantic knowledge expression of spatial relation terms based on Ontology. The conclusion and future work are given in Section 4.

2 Classification of Spatial Relation Terms

Spatial relations are considered to be one of the most distinctive aspects of spatial information. According to Egenhofer and Franzosa’s argument, spatial relations can be grouped into three different categories of topological relations, direction relations and distance relations. Natural-language spatial relations are spatial relations de-
scribed in natural language among people’s daily communication, it is much closer to people's habit than GIS spatial relations [10]. The description and expression forms of spatial relations in natural language and GIS are very different. Spatial relations in natural language are richer, but with a qualitative, fuzzy, uncertainty and unstructured characters, while spatial relations in GIS are quantitative, structured, and accurate. Topological relations have long been considered as the most important spatial relations in GIS while direction and distance relations are with the highest using frequency in people’s daily life. Spatial relations in natural language are expressed through a series of spatial relation terms. In different language, those terms are with different diversity and complexity. Taking the spatial relation term of “cross” in English for example, in Chinese it can be expressed as “穿越(chuanyue)”, “交叉(jiaocha)”, “横贯(hengguan)”, etc. Meanwhile, some spatial relation terms in Chinese indicate more than one spatial relation type. For example, the spatial relation term “北靠(beikao)” not only expresses the north direction but also implies a topological relation of extended connection. In addition, there are some spatial relation terms in text descriptions whose semantic meanings can’t be expressed with existing calculation models of GIS spatial relations. Taking the spatial relation term of “支流(zhiliu, anabranch)” for example, it may describe an including relation of the main vein and tributaries of a river, however, this semantic relation can’t be expressed in GIS spatial relation models.

Region connection calculus (RCC) model takes geographical entities in the real world as a region and describes spatial relations with the region connectedness [11]. Therefore, it is in accordance with human’s cognition habit and more suitable for qualitative representation and reasoning of spatial relations. The ternary point configuration calculus (TPCC) describes directions such as front, back, left and right [12]. Distance relations specify that how far the object is away from the reference object. Based on RCC8, TPCC, and the frequency of spatial relation terms in natural language context, basic categories of spatial relations and classifications of spatial relation terms are described in Table 1. From table 1, we can see that one spatial relation category may include multi-spatial relation terms, and one spatial relation term may correspond to more than one spatial relation categories. Also, there are some commonly used spatial relation terms which can’t be clustered into these categories, such as between, round, etc. Here it should be noted that this paper only discusses a binary instance of spatial relations between two geographical entities, not consider the composite spatial relations. For some compound spatial relation terms, the classification will be determined by the last direction word, such as “中南部(zhongnanbu, south-centre)” with a direction of south. Also, there are some connected words which can’t reflect topological or direction relations, but provide the connection between the source and target objects, so they play a role in auxiliary judgments of spatial relations, such as “located”, “is”, “as”, “with”, “by”, etc.
### Table 1. Samples of classifications of spatial relation terms

<table>
<thead>
<tr>
<th>Spatial Relations</th>
<th>Spatial Relation Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topological relation</strong></td>
<td></td>
</tr>
<tr>
<td>......IN(tangential and non-tangential proper parts)</td>
<td>包含(包括, including), 流入(流入, flow into), 属于(属于, belong to)</td>
</tr>
<tr>
<td>......EC(extended connection)</td>
<td>相接(相接, touch), 起点(起点, starting point), 流入(流入, flow into)</td>
</tr>
<tr>
<td>......DC(discrete connection)</td>
<td>相离(相离, discrete connection), 相隔(相隔, compartment), 相距(相距, apart)</td>
</tr>
<tr>
<td>......PO(Partially overlap)</td>
<td>贯穿(贯穿, run through), 交叠(交叠, overlap), 穿越(穿越, run through)</td>
</tr>
<tr>
<td>......EQ(equality)</td>
<td>相等(相等, equal), 别名(别名, alias), 简称(简称, named)</td>
</tr>
<tr>
<td><strong>Directional relation</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Relative direction</strong></td>
<td></td>
</tr>
<tr>
<td>......F(front)</td>
<td>前头(前头, front), 前面(前面, front), 前部(前部, foreshore)</td>
</tr>
<tr>
<td>......BE(behind)</td>
<td>后端(后端, back-end), 后面(后面, behind), 后部(后部, behind)</td>
</tr>
<tr>
<td>......L(left)</td>
<td>左边(左边, left side), 左面(左面, left), 左(左, left)</td>
</tr>
<tr>
<td>......R(right)</td>
<td>右边(右边, right), 右端(右端, right), 右(右, right)</td>
</tr>
<tr>
<td>......A(above)</td>
<td>上端(上端, above), 上(上, above), 上面(上面, above)</td>
</tr>
<tr>
<td>......BW(below)</td>
<td>下端(下端, below), 下(下, below), 下面(下面, below)</td>
</tr>
<tr>
<td>......INT(inner)</td>
<td>内(内, inner), 内部(内部, inside), 内面(内面, inside)</td>
</tr>
<tr>
<td>......EXT(exterior)</td>
<td>外(外, outer), 外部(外部, exterior), 外面(外面, outside)</td>
</tr>
<tr>
<td><strong>Absolute direction</strong></td>
<td></td>
</tr>
<tr>
<td>......E(east)</td>
<td>东面(东面, east), 东方(东方, east), 东端(东端, east)</td>
</tr>
<tr>
<td>......W(west)</td>
<td>西面(西面, west), 西部(西部, west), 西(西, west)</td>
</tr>
<tr>
<td>......S(south)</td>
<td>南面(南面, south), 南(南, south), 南方(南方, south)</td>
</tr>
<tr>
<td>......N(north)</td>
<td>北面(北面, north), 北方(北方, north), 北(北, north)</td>
</tr>
<tr>
<td>......C(centre)</td>
<td>中心(中心, centre), 中心(中心, centre), 中间(中间, middle)</td>
</tr>
<tr>
<td>......NE(northeast)</td>
<td>东北面(东北面, northeast), 东北方(东北方, northeast)</td>
</tr>
<tr>
<td>......SE(southeast)</td>
<td>东南面(东南面, southeast), 东南方(东南方, southeast)</td>
</tr>
<tr>
<td>......NW(northwest)</td>
<td>西北面(西北面, northwest), 西北方(西北方, northwest), 西北部(西北部, northwest)</td>
</tr>
<tr>
<td>......SW(southwest)</td>
<td>西南部(西南部, southwest), 西南部(西南部, southwest), 西南部(西南部, southwest)</td>
</tr>
<tr>
<td><strong>Distance relation</strong></td>
<td>距离(距离, distance), 相离(相离, distance), 相距(相距, distance)</td>
</tr>
</tbody>
</table>

### 3 Calculation of Relevance

In linguistics, a text corpus is a large and structured set of texts which are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. This paper takes the large scale annotation corpus (Geocorpus) of spatial relations of “Chinese Encyclopedia (geography)” in paper [13] as an experimental data, and summarizes 600 commonly used spatial relation terms...
among the annotation process. Based on those experimental data, a quantitative approach to evaluate the relevance of spatial relation terms and geographical feature types from text corpuses is proposed.

A binary spatial relation could be formalized as \(<\text{geographical entity } A, \text{ spatial relation terms, geographical entity } B >\) in natural language context. Obviously, one spatial relation term should associate with a pair of geographical entities. For the concept characteristics of geographical entities could be defined by the type of geographic features, therefore, a single spatial relation can be further abstracted as \(<\text{feature type of geographical entity } A’, \text{ spatial relation term, feature type of geographical entity } B’ >\). In spatial relation descriptions, there is an order for target and reference objects. To simplify the calculation, the order between geographical entity A and B is not distinguished in this paper. Overlap is a classic calculation method for semantic relations, and it is based on the co-occurrence frequency of two events in a data set [14]. Therefore, the relevance of spatial relation terms and the type of geographical entities based on Geocorpus can be defined as in formula 1.

\[ R(T, A’B’) = \frac{|T \cap A’B’|}{\min(|T|, |A’B’|)} \]

In formula 1, T represents the occurrences of a spatial relation term in the Geocorpus, A’ and B’ denote the occurrence of two geographical feature types, R indicates the relevance degree between spatial relation term (T) and a pair of geographical feature types (A’ and B’). Taking the spatial relation term of “流入 (liuru, flow into)” as an example, the results of the relevance are just as shown in figure 1.

![Fig. 1. The relevance of “流入 (flow into)” and geographical feature types](image-url)
The result shows that spatial relation terms are strongly relevant to geographical feature types. Some of the relevance have a higher R-value, such as the relevance of “流入 (flow into)” with rivers, lakes and oceans, but some of the relevance is lower than 0.05. One reason for this phenomenon is the imbalance and coarse granularity of geographical concepts in the Geocorpus, and another reason is that there is an imbalance for geographical phenomena in the real world. The R-value is higher when the phenomenon is more common. As we all know, there are a lot of spatial relation terms to describe spatial relations of entities of river, ocean and lake, while the term “流入 (flow into)” can describe spatial relations of geographical entities of multi-feature types. However, people are used to choose “流入 (flow into)” to describe spatial relations for geographical entities of river, ocean and lake in daily life. Therefore, this result comparatively conforms to people’s language and cognitive habit. In order to keep the balance of the relevance we can set and adjust a threshold to filter the uncommon relevance in a text corpus, such as 0.05 for “流入 (flow into)”. In addition, the relevance of topological relation terms and geographic feature types are significantly stronger than directional relation and distance relation terms.

Ontology formally represents rich knowledge as a set of concepts and the relationships between those concepts within a domain. It can improve the consistency, accuracy, reusability and sharing features of knowledge to understand and use. In this paper, a knowledge base is developed based on protégé to formally represent and visualize geographical feature types, spatial relation classifications, spatial relation terms and their relevance with geographical feature types (see Figure 2). Geographical feature types and spatial relation classifications are expressed with a class in OWL language, and the hierarchy relationship is established by the subClassOf. The semantic relations between spatial relation terms and geographic feature types are expressed by the ObjectProperty, and the quantitative constraint value are organized in DatatypeProperty. Then the relevance is defined a property with “rdfs: domain” and “rdfs: range” respectively, which can restrict the application fields and scope. Finally, instances of spatial relation terms can be made according to the semantic relevance of ObjectProperty and DatatypeProperty. This knowledge base can improve GIS natural language query, extraction of spatial relations from text, geographical information retrieval, qualitative spatial reasoning, etc.
4 Conclusion

Based on a large scale text corpus and the geographical feature classification scheme this paper proposed a method to explore the relevance of spatial relation terms and geographical feature types. The experiment indicates that our proposed approach can effectively obtain meaningful results. However, the annotation quality of the corpus and the classification granularity of geographical entities have a great effect on the performance, especially for a general dataset. In our future work, we will start the classification on different kinds of texts describing the same kind of data (e.g. documents addressing only water, sea) in order to better extract relations specified for a particular domain. Moreover, the relevance of spatial relation terms and the type of geographical entities have big limitations based on manual annotation and judgment completely. There are rich semantic and inherited hierarchy relationships among the classification system of geographic feature types, especially in the Ontology of geographic feature types. Semantic relations can be expanded through a calculation of the semantic distance. Therefore, we will also make a further research to expand the relevance of spatial relation terms and the type of geographical entities based on this semantic hierarchy relationship in the future work.
5 References

Syntactic Patterns of Spatial Relations in Text

Shaonan Zhu, Xueying Zhang

Key Laboratory of Virtual Geography Environment, Ministry of Education, Nanjing Normal University, Nanjing, China

Abstract: Natural language is considered as the important resource of geospatial information including geospatial relations. Syntactic patterns represent linguistic characteristics of spatial relations described in natural language, which can be effectively used in spatial information retrieval, spatial query in GIS, cartographic language and spatial information retrieval. The previous research focused on induction syntactic patterns of spatial relations manually. In this paper, a machine learning approach is proposed, which is based on an annotation corpus marked by spatial relations. Firstly, the instances of the spatial relations in the corpus are extracted; secondly, sequence alignment algorithm is used to calculate the similarity between s of spatial relations; finally, the instances with higher similarity are generalized to generate the syntactic patterns of geospatial relations.

Keywords: spatial relation; sequence alignment; syntactic pattern; text

1 Introduction

Spatial relation described in natural language is much closer to human spatial cognition [1]. To understand and analyze how people express spatial relations is a meaningful issue in geographical information science [5]. National Geographic Information and Analysis Center (NCGIA) has implemented a study plan on spatial relations, since 1988. From a linguistics point of view, the description of spatial relations in natural language is mainly dependent on the spatial relation terms and syntactic structures. Spatial relations terms reflect the state of geographical entities and spatial relations. Syntactic structures delimit the semantic of spatial relations described in the specific pragmatic. Although the spatial relation terms, syntax and
semantics have strong fuzziness and uncertainty, there still exist some typical syntactic structures, i.e. syntactic patterns. Therefore, the identification of syntactic patterns of spatial relations is helpful for the understanding of spatial relations in natural language and their usage in GIS query, spatial reasoning and geographic information retrieval.

The studies aiming at syntactic patterns mainly focus on the field of scene reconstruction from natural language. The Wordseye system using part of speech tagging and syntactic analysis, on the basis of the initial annotation and analysis, transforms syntactic parsing tree into a dependency structure, and expresses spatial interdependent relationship between the spatial objects. Reinbergerr extract the spatial relations with the syntactic structure similar to “subject-verb-object” and special field knowledge. In general, the syntactic patterns about spatial relations are manually summarized [4-7].

Corpus is the resource of linguistics research and information extraction. Therefore, a large-scale annotated corpus will provide a new way to automatically identify of syntactic patterns of spatial relations in text. This paper is based on a annotation corpus marked by spatial relations, analyses the instances from the corpus, uses the sequence alignment algorithm to calculate the similarity between instances of spatial relations, builds the similarity matrix, group instances of high similarity, generalized to generate the syntactic patterns of spatial relations.

2 The similarity matrix for instances of spatial relations

A spatial relation can be abstracted as a binary relation between the two geographical entities, has apparent target and reference. In structure, a instance must contain two geographical entities with the order. This paper defines <left, GNE1, middle, GNE2, right> as the structure of the instance. GNE1 and GNE2 mean the target and reference in the instance of spatial relations (GNE: geographic named entity). Left, middle, right mean the context. In general, a description about the spatial relation is in a sentence, so the context of the description is usually in a sentence. The part of speech is the most basic language feature in the instance of spatial relations. On the other hand, the vocabulary of spatial relations has a strong indication about spatial relations. Geographic named entity, POS (part of speech), and the word which is in the vocabulary of spatial relations can express natural language characteristics of the instance of spatial relations. Figure 1 expresses the structural result of
the instances of spatial relations. For example, “阿里山位于嘉义市以东(ALISHAN is east of JIAYI City)” is a instance. After the annotation features are extracted and structured, the final feature sequence is “<GNE><V><GNE><SIGNAL>”.

<table>
<thead>
<tr>
<th>Left</th>
<th>GNE1</th>
<th>Middle</th>
<th>GNE2</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;&gt;</td>
<td>&lt;GNE&gt;</td>
<td>&lt;V&gt;</td>
<td>&lt;GNE&gt;</td>
<td>&lt;SIGNAL&gt;</td>
</tr>
</tbody>
</table>

Fig. 1. The instance of spatial relations

In order to find the rules from spatial relations, the structure of the sentence which describes spatial relations must be researched. The words in the instances can be regarded as symbols. This paper uses sequence alignment algorithm which is widely applied in the field of biology to compute the similarity. A major theme of genomics is to compare DNA sequences and to find out the public part of the two sequences. Two ways are frequently used. The first one is local sequence alignment, which gets the similarity from local information. Smith-Waterman Algorithm is the classical algorithm. The other one is global sequence alignment, and Needleman-Wunsch Algorithm is the typical algorithm.

This paper extends Needleman-Wunsch Algorithm to handle language unit in the sequence. For example, there are two sequences: 【GNE1】【V】【GNE】【GNE2】 and 【GNE1】【V】【GNE2】. Firstly, two-dimensional table must be generated. Secondly, two sequences with the same length can be built.

Thirdly, the similarity between the sequences is quantified (see 1).
\[ \text{SIM} = \frac{\text{SUM}}{\text{LENGTH}} \] (1)

SUM means the score of the whole sequence. If at the same position the language unit in the target sequence is the same with the reference sequence, the score is added one point, and called one match. LENGTH means the sequence length. SIM is the result. The score of the example which is expressed in Figure 3 is 0.75.

Because the similarity should be computed between two instances, the matrix can be built. The similarity matrix for instances of spatial relations show the similarity of the whole sequence list. It is prepare for the generalization of syntactic patterns.

3 Generalization of syntactic patterns

The sequence can be considered as a syntactic pattern, but it has not a representative. In fact, the instances of spatial relations can be considered as a sample set, then syntactic patterns are generalized according to the value of the similarity matrix. The procedure is defined as follows:

1. traverse the instances, and pick up the instance and the list which corresponds the instance in the similarity matrix;
2. get the most similar instances, the number of the instances is variable (1, 2, …, n);
3. generalize the instances from step two. If the result contains language feature, go back to step two. Otherwise, go back to step one.
4. loop until all the instances are traversed in step one.

The key point is how to define the generalization templates. When the two sequences are compared, the same kind of feature will be retained. For example, the target sequence is “<GNE1><a/n/n/w/t/v/v/t/p><GNE2><n>”, and table 1 shows the list which corresponds the instance in the similarity matrix.

<table>
<thead>
<tr>
<th>Reference Template</th>
<th>Similarity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;GNE1&gt;&lt;m/n/c/w&gt;&lt;GNE2&gt;&lt;n&gt;</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>&lt;GNE1&gt;&lt;v/a/n/f/w/c/SIGNAL/w/v/s/w/f/v/w/GNE/w/w&gt;&lt;GNE2&gt;&lt;n&gt;</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>
The result of the above-mentioned is “<GNE><n/w><GNE><n>”, “<GNE><GNE><n>”.

4 Experimental evaluation

The syntactic patterns serve for the extraction of spatial relations in Chinese text with the rules. Traditionally, the rules are generalized manually. This paper gives a helpful method to rich the rules. How the text is handled with the rules is as follows. Firstly, the Chinese word segmentations in a sentence are got. At the same time the POS can be got. At last, the spatial signal word can be found by the spatial terms. So the syntactic patterns can be used to match the text which has been handled. That is why the patterns are important.

This paper uses the annotated corpus called GeoCorpus. GeoCorpus gets data form encyclopedia of china which contains a full description of the geographic elements of the administrative unit, mountains, rivers, hills, plateaus, plains, basins, etc. This experiment selected 2355 instances from GeoCorpus. After the instances were structured and generalized, 5295 syntactic patterns were obtained, and someone repeated several times. Through statistical analysis, there were 996 syntactic patterns except the patterns were repeated. In 5295 syntactic patterns, there were 920 syntactic patterns which the number of repetitions is less than 10. 50 syntactic patterns which had the highest frequency accounted for 70% of the total. Tab.2 shows the syntactic patterns which are the top ten.

<table>
<thead>
<tr>
<th>Syntactic patterns</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>【GNE1】【SIGNAL】【GNE2】</td>
<td>622</td>
</tr>
<tr>
<td>【GNE1】【GNE2】【SIGNAL】</td>
<td>403</td>
</tr>
<tr>
<td>【GNE1】【W】【GNE2】</td>
<td>274</td>
</tr>
<tr>
<td>【GNE1】【GNE】【GNE2】</td>
<td>197</td>
</tr>
<tr>
<td>【GNE1】【V】【GNE2】</td>
<td>156</td>
</tr>
<tr>
<td>【GNE1】【GNE2】【N】</td>
<td>135</td>
</tr>
<tr>
<td>【GNE1】【GNE2】【GNE】</td>
<td>120</td>
</tr>
</tbody>
</table>
The experimental results show that in Chinese text, spatial relations usually appear in a sentence; there is almost no qualifier before the former GNE in the syntactic patterns, some syntactic patterns is the part of other syntactic patterns. These Linguistic phenomena and rules help to further understand human cognition on spatial relations.

5 Conclusion

Based on an annotation corpus, this paper proposes an approach to get the syntactic patterns with the help of the similarity matrix by sequence alignment algorithm. This method overcomes the shortcomings of manual induction, and finds out hidden rules about spatial relations. The future research should focus on how to extract the spatial relations in Chinese text by syntactic patterns, and extend their usage in GIS natural language query.

6 References