Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval*

Davide Buscaldi
Universidad Politécnica de Valencia
Camino de Vera, s/n
Valencia, Spain
dbuscaldi@dsic.upv.es

Paolo Rosso
Universidad Politécnica de Valencia
Camino de Vera, s/n
Valencia, Spain
prosso@dsic.upv.es

Piedachu Peris García
Universidad Politécnica de Valencia
Camino de Vera, s/n
Valencia, Spain
pperis@dsic.upv.es

ABSTRACT

Many documents that can be found in the World Wide Web include some kind of geographical information, often in an implicit way. The use of resources like gazetteers and geographical ontologies can improve the results in Geographical Information Retrieval. Unfortunately, the construction of such ontologies is a long and laborious process; therefore, building an ontology in a semi-automatical way exploiting multiple sources is an interesting and useful task. Our work is focused on the integration of data from gazetteers (GNS and GNIS), the WordNet general domain ontology, and Wikipedia, the free encyclopedia. The result of our effort is an ontology implemented as a set of Prolog clauses, that can be easily expanded with both new data and relationships.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Geographical Information Retrieval, Ontologies, Wikipedia, WordNet, gazetteers

1. INTRODUCTION

Most of the information available in electronic format, such as in the World Wide Web or in digital libraries, involves some kind of spatial awareness. For instance, news usually describe an event and the place where this event occurred: "Earthquake in Turkey", "Visit of the Pope in Valencia". Currently, the Information Retrieval (IR) research community is increasing its efforts dedicated to the retrieval of geographical information, as testified by the creation of the GeoCLEF [1] [5] evaluation exercise at the CLEF 2005, recently repeated in 2006, and the advances of the SPIRIT[2] project [6]. These efforts are aimed to the solution of typical issues of the geographical IR task.

In many cases, explicit geographical information is missing from the documents, for instance the indication of a broader geographical entity is omitted when it is supposed to be well-known to the readers (e.g. usually *France* is not named in a news related to *Paris*). Another common problem is the synonymy, when there are many ways to indicate a geographical entity. This is particularly true for foreign names, where spelling variations are frequent. The solution to these problems has been generally individuated in the use of geographical-oriented ontologies [4, 6]. The manual construction of this kind of resources is usually a long, laborious process, and in many cases they are not freely available, such as the Getty Thesaurus of Geographical Names[3] (TGN). In order to overcome this issue, we made some attempts [2, 3] to use the geographical information included in WordNet, the well-known general domain ontology developed at the University of Princeton [7].

Unfortunately, the quantity of geographical information included in WordNet is quite small. Although it is quite difficult to calculate the number of geographical entities stored in WordNet, due to the lack of an explicit annotation of the synsets, we retrieved some figures by means of the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. On the other hand, gazetteers like the Geonet Names Server[4] (GNS) and the Geographic Names Information System[5] (GNIS) are freely available and provide plenty of geographical informations. The problems of these resources is that they do not organize the information in a structured way like ontologies, and that they contain *too many* names; therefore, increasing the ambiguity of geographical names (for instance, 16 places named "*Genoa*" can be found in various locations all over the world: one in Italy, another in Australia and the remaining ones in the United States).

*We would like to thank R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects for partially supporting this work.

[1] http://ir.shef.ac.uk/geoclef/
[2] http://www.geo-spirit.org
[3] http://www.getty.edu
[4] http://earth-info.nga.mil/gns/html/index.html
[5] http://geonames.usgs.gov/domestic/index.html

Encyclopedias also contain a quantity of geographical information. One of the most interesting and recent phenomena in the Web is the success of Wikipedia[6] as source of information. We already studied the possibility of using Wikipedia for Question Answering [1], a task related to the IR field, and we realized that it could be exploited also for the Geographical Information Retrieval, since the articles usually include useful information (such as boundaries) that can be used in order to extract relationships among geographical entities.

Each of the discussed resources presents advantages and disadvantages. For instance, gazetteers lack informations about the composition of geo-political entities such as Europe, England, Scotland. This information can be retrieved by means of WordNet and/or Wikipedia. In this paper we describe our work in order to integrate the information extracted from WordNet, the GNS and GNIS gazetteers and Wikipedia into a geographical ontology.

2. BUILDING THE ONTOLOGY

In this section we describe each of the steps taken in order to build the ontology. We have to remark that the construction of the ontology has been undertaken with the aim of improve and extend the portion of the WordNet ontology that was used for our previous works in the field of Geographical Information Retrieval [2, 3]. In detail, we were interested to extend geographical names in a document collection with the indication of the containing entities. Therefore, the largest part of the work was to extract information of containment between the entities. The ontology has been implemented as a Prolog database, therefore it can be easily expanded with new relationships and/or data.

2.1 Extraction of Data from Gazetteers

The GNS and GNIS gazetteers contain a lot of geographical information respectively about places outside and inside the United States (more than $5,500,000$ places for the GNS and almost $40,000$ for the GNIS - *concise* dataset). Both gazetteers use a similar format, where in each line there is at least the name of a place, its coordinates, the indication of the containing entity (state for the GNIS, a regional code + the ISO code of the country for GNS), and a class.The places are classified using *feature designation codes*; for instance, "populated places" are assigned the PPL code, "volcanoes" are identified with VLC and so on. In order to reduce *polysemy* among places of different type, we selected only places of the following classes: *bay, cape, gulf, hill, island, lake, mountain, ocean, populated places, port, sea* and *volcano*. These classes were chosen for their relative importance with respect to other classes (such as "garden" or "airfield"), and after an analysis of the GeoCLEF questions.

As an example, consider this portion of a line from the GNS:

```
123000  -695800   ISL   AA   00    Aruba
```

Since coordinates of geographical places are not given in WordNet, we considered as useful information only the class (ISL -island- in this case), the country code (AA) the region code (00), and obviously the name of the place ($Aruba$). This is how the previous data are translated into the Prolog database:

```
island("Aruba").
in("Aruba", "AA00").
```

The first clause states that $Aruba$ is an island, while the second one states that $Aruba$ is included in the region $AA00$. As it can be noticed, this is a combination of the ISO country code and the regional code. This uniquely identify a region in the world. We defined a predicate, $abbr/2$, which allows to individuate the full name of a code. Therefore, in the database the following clauses are also present:

```
abbr("AA00", "Aruba (general)").
in("AA00", "Aruba").
abbr("AA", "Aruba").
country("Aruba").
```

The information about the regional codes was retrieved from the "First-order Administrative Division Code to First-order Administrative Division Name Cross Reference".

In the case of the GNIS, US alphabetical state codes (FL, AZ, NJ, etc.) were used as regional codes, combined with the ISO code for the United States (US). For instance, the information about the containment relationship $Phoenix \rightarrow Arizona \rightarrow USA$ is expressed into the database by means of the following clauses:

```
city("Phoenix").
in("Phoenix", "AZUS").
abbr("AZUS", "Arizona").
in("Arizona", "US").
state("Arizona").
```

As said above, we used the *concise* dataset of the GNIS, that is, large features that should be labeled on maps with a scale of $1 : 250,000$. The reasons of our choice were, again, the reduction of polysemy, and the fact that queries over large databases are too slow to be used efficiently during the automated indexing of a large collection of documents (as we needed for our previous work [2, 3]). Unfortunately, as detailed above, the places listed in the GNS are far more than those listed in the GNIS, and it is not possible to download a reduced dataset such as for the GNIS. Therefore, we attempted to establish the importance of a named place by looking into an encyclopedia.

2.2 Filtering Names through Wikipedia

The consideration that stands behind the attempt is that large, important geographical features correspond to well-known names. Popular names can be found in an encyclopedia, and more if the encyclopedia can be edited by anyone, as in the case of Wikipedia. Therefore, we tried to add to the database only names that are included into the titles of the articles of Wikipedia. Unfortunately, there is an issue with this approach relative to popular names that are

[6]http://www.wikipedia.org

not names of geographical entities, or better, that are much more popular than the locations we would like to add to the database. For instance, *Leno* is a popular showman (Jay Leno) in the United States, but it is also a small town in Northern Italy and many other places that can be found in GNS but not in Wikipedia.

In order to overcome this issue, we had to select from the whole collection only the articles referring actually to some geographical name. We extracted from WordNet a set of geographical *trigger words* using the *holonymy* (part-of) relationship and its reverse, *meronymy*. We retrieved iteratively all the meronyms from two root synsets: *northern_hemisphere* and *southern_hemisphere*. The result is the list of all the geographical synset included in WordNet. The words contained in these synsets and in the definition of each one (the *gloss*) were added to the set of trigger words, with the exception of stop-words. For instance, consider the following synset (between braces) and its gloss:

```
{Paris, City of Light, French capital, capital of
France} - the capital and largest city of France;
international center of culture and commerce.
```

The terms added to the set of trigger words in this case are: Paris, City, Light, French, capital, France, largest, city, international, center, culture, commerce. The 10 most frequent words extracted in this way are: city, state, population, area, world, km, country, new, north, river.

Therefore, the obtained words can be considered as quite representative of the geographical domain. In order to determine whether a Wikipedia article is in the geographical domain or not, we need to measure its similarity to the set of trigger words. Let us name W_a the set of words in an article a of Wikipedia, T the set of trigger words extracted from WordNet, then the similarity score $S(a,T)$ between a and T is computed by means of the Dice formula:

$$S(a,T) = \frac{2|W_a \cap T|}{|W_a| + |T|} \quad (1)$$

We indexed only the documents with $S(a,T) > 0,04$. We used the Xapian[7] search engine to index the Wikipedia snapshot.

2.3 Integration with WordNet

At this point, we realized that some useful informations were missing from the ontology. For instance, we had the information that *Cambridge* is a place in the United Kingdom, Cambridgeshire, but the GNS does not contain any reference that Cambridge is in *England*. Neither the GNS tells whether France is in Europe or in America. Considering that many topics of the GeoCLEF make reference to locations in this way, (for instance, topic 26: *Wine regions around rivers in Europe*), this was a major issue to deal with. Fortunately, the missing information can be found in WordNet. For instance, in WordNet *England* is a meronym of *United Kingdom*.

The integration with WordNet is done in two steps. In the first one, for each name n from the ontology obtained from the GNS and GNIS are extracted the names of the two containing entities (respectively region and country) c_1 and c_2, the two holonyms h_1 (direct) and h_2 (direct holonym of h_1) of n in WordNet and the class of n from WordNet through the relationship *instance_of* (this can be city, river, country, etc.). Let us define D as the set of clauses in the database. The actions that can be taken at this point are:

- if $h_1 = c_1 \wedge h_2 = c_2$: add to the database the synonyms (S_n) of n in WordNet; $\forall s \in S_n$ add the clause $in(\text{"}s\text{"},\text{"}c_1\text{"})$. to the database, if it does not contain already s.
- if $h_2 = c_2 \wedge h_1 \neq c_1 \wedge in(h_1, c_2) \notin D$: add to the database the clauses $in(\text{"}h_1\text{"},\text{"}c_2\text{"})$. and $in(\text{"}n\text{"},\text{"}h_1\text{"})$.

A manual mapping has been done between regional codes and WordNet synsets in order to write the clauses following the same format used when extracting data from the gazetteers.

The second step was done semi-automatically, adding to the database all the clauses of containment between countries and continents entities, selected from WordNet starting from the root geographical concepts (northern and southern hemisphere).

2.4 Boundaries

Another interesting relationship that it is not present in the gazetteers but can be retrieved from Wikipedia is if two entities are bounding each with another one. We analyzed some articles in order to identify textual ("shallow" as opposed to syntactical, "deep" ones) patterns that are often used. The recent efforts of the Wikipedia community in order to improve the quality of the geographical section (WikiProject Geography[8]) give an important contribution to the standardization of the patterns.

Common expression are "X borders Y", "X shares borders with Y", "X is a ... bordering Y", "X is bordered by Y", where X is the country described in the article and Y are one or more internal links to other countries. We collected manually a set of regular expressions matching these patterns. Internal links can be easily identified in Wikipedia since they are written between double square brackets (in the xml dumps).

For simplicity, we included boundaries only for location of the class *country*. For every country in the database, the respective entry in Wikipedia was examined in order to individuate one of the defined patterns. If the countries in the links were already present in the database, a relationship `bound(X, Y).` was added for each country.

For instance, consider the page of *France* in Wikipedia; the passage describing the boundaries of France with neighboring countries is: "France is bordered by Belgium, Luxembourg, Germany, Switzerland, Italy, Monaco, Andorra, and Spain". These countries are also present in the database, there fore the following clauses are added to it:

[7] http://xapian.sourceforge.net

[8] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geography

```
bound(France, Belgium).
bound(France, Luxembourg).
bound(France, Germany).
...
bound(France, Spain).
```

In order to reduce the number of clauses, we observed that the boundary relationship is reflexive. Therefore we checked also the presence of `bound(Y,X).` in the database before adding `bound(X,Y).` to it.

3. USING THE ONTOLOGY

Thanks to the flexibility of Prolog, all the information in the ontology can be easily retrieved. For instance, suppose that we need to find all the cities in a given region Y; therefore, we have to add to the database only the following rule:

```
city_of(X,Y):- city(X), abbr(Z,Y), in(X,Z).
```

The most useful information, we planned to use for GeoCLEF, is the containment. In this case the rules are:

```
cont(X,Y):- in(X,Y).
cont(X,Y):- in(Z, Y), cont(X,Y).
```

Our aim was to use the ontology for the expansion of geographical terms in a collection of documents. The indexing process was performed by means of the Lucene search engine, generating two index for each text: a *geo* index, containing all the geographical terms included in the text and those obtained through the ontology, and a *text* index, containing the stems of text words that are not related to geographical entities. Thanks to the separation of the indices, a document containing "John Houston" will not be retrieved if the query contains "Houston", the city in Texas. The adopted weighting scheme is the usual *tf-idf*. The geographical terms in the text are identified by means of a Named Entity (NE) recognizer based on maximum entropy[9], and put into the *geo* index.

For instance, consider the following text:

"`In 2001, Genova was the seat of the G8.`"

The NE recognizer identifies *Genoa* as a geographical entity. A search for `cont("Genova", X).` returns {*IT08 (Liguria), Italy, Europe, northern hemisphere*}. Therefore, the following index terms are put into the *geo* index: {Genova, Liguria, Italy, northern hemisphere}. The result of the expansion of index terms is that the above text will be indexed also by words like *Liguria* and *Italy* that were not explicitly mentioned in it.

There are two problems that we still need to solve: one is the synonymy: we are not adding to the index the English name of Genova, Genoa. The second one is the polysemy: suppose that *Genoa* was found in the text instead of *Genova*, then we would have also the names related to the Genoas in the United States and in Australia.

Whereas in the first case we can simply add to the database a *synonymy* relationship, based on hierarchy similarity and on footprints (but coordinates can be obtained only from the gazetteers), in the second one we need to *disambiguate* the term inside the text, a task recognized as one of the most difficult in the field of Natural Language Processing.

[9]Freely available from the OpenNLP project: http://opennlp.sourceforge.net

4. CONCLUSIONS

This paper describes our efforts in order to create in a semi-automatical way an ontology that can be used effectively as a resource for the Geographical Information Retrieval task. We made use of three resources: gazetteers, Wikipedia and WordNet. The information contained in the three resources is very heterogeneous, and we succeeded to integrate them only at a small extent, mostly because the construction of the ontology was subordinated to the needs arisen from our previous experience in the use of WordNet for the GeoCLEF task. We still need to verify if the use of the ontology give benefits over the use of WordNet alone. Moreover, the ontology itself does not solve the ambiguity problems. Further investigations will be aimed at developing disambiguation methods that use the ontology (possibly an improved version of it), and a performance comparison with an implementation of the same ontology with a relational database instead of Prolog.

5. REFERENCES

[1] D. Buscaldi and P. Rosso. Mining knowledge from wikipedia for the question answering task. In *Proceedings of the LREC 2006*, 2006.

[2] D. Buscaldi, P. Rosso, and E. Sanchis. Using the wordnet ontology in the geoclef geographical information retrieval task. In *Proceedings of the CLEF 2005*, 2005.

[3] D. Buscaldi, P. Rosso, and E. Sanchis. Wordnet as a geographical information resource. In *Proceedings of the 3rd Global WordNet Association (GWA06)*, 2006.

[4] G. Fu, C. Jones, and A. Abdelmoty. Bulding a geographical ontology for intelligent spatial search on the web. In *Proceedings of the IASTED International Conference on Databases and Applications*, 2005.

[5] F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough. Geoclef: the clef 2005 cross-language geographic information retrieval track. In *CLEF 2005 Working Notes, C.Peters Ed.*, 2005.

[6] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the spirit project. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 2002.

[7] G. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.