Toponym Disambiguation by Ontology in Spanish: Geographical proximity between place names in the same context*

[1]Adriana López, [2]María J. Somodevilla, [3]Darnes Vilariño, [4]Ivo H. Pineda and [5]Concepción P. de Celis
[1,2,3,4,5]*Faculty of Computer Science, B. Autonomous University of Puebla, México,*
ady.lopez.cumplido@gmail.com, mariasg@cs.buap.mx, darnes@cs.buap.mx,
ipineda@cs.buap.mx, mcpcelish @gmail.com

Abstract

This article considers the issue of the Geographic Information Retrieval (GIR) suggesting the development of techniques for disambiguation of toponyms based on ontologies to address the ambiguity in GIR queries. In particular, mixed domain ontology was built to way of organizing the Spanish-speaking countries, in order to develop a disambiguator method of place names in Spanish. The proposed method is based on geographic proximity between toponyms in the same context, using hierarchical relationships provided by the ontology and an ontological hierarchical weighting, which is complemented by the Haversine distance.

Keywords*: Ontology, Disambiguation, Toponym, GIR*

1. Introduction

The ambiguity of place names is a major problem in the GIR, given that this task demands of the users are geographically linked. This paper introduces a new disambiguation method based on the construction of a mixed domain ontology, a way of organizing the territorial distribution of Spanish-speaking countries, in order to develop a new method for the disambiguation of toponyms which is based on geographical proximity between place names in the same context, and trying ambiguity in geographic information queries. The study was conducted on the place names included in a Mexican news collection in order to resolve the ambiguities from the geographical point of view. The information provided by the news these days is enriched with towns, cities, provinces and Spanish-speaking popular locations. Since many places in the world are named identically, not identify the correct referent could lead to a misinterpretation of the news. Section 2 begins with an introduction to the methods proposed by the research community, currently dealing with toponym disambiguation task and subsequently the geographical domain ontologies are analyzed. Section 3 describes the proposed disambiguation approach by using a mixed ontology. Section 4 discusses the experimental results followed by the conclusions of this work.

2. State of the Art

Toponym resolution aims to link place names with geographic representation, and is it is applied into areas such as Question Answering, mapping, GIR, among others. The Geo / information is abundant on the Weband digital libraries (e.g. collections of geo-referenced photographs like Flickr), where approximately 80% of Webpages contain references to other places [9]. In the task of GIR, most user requests are of the type "X" in "P", where P is a place name and X the subject of consultation [10].

This section presents an overview of the state of the art of GIR area, which involves different lines of research since it is multidisciplinary and recent birth. The following is a brief description of the most outstanding works. The work [1] presents a geographical similarity operator that calculates the relationship between two geographical locations and describes how to combine the textual

classification. The work [2] describes a system for toponym disambiguation based on the Perseus[1] digital library; this method calculates the geographic centroid candidates and then, removes those that are located at a distance more than twice the centroid standard deviation. In Jones et al. (2002), some methods to extract geographic information from Web pages, as part of the SPIRIT (Spatially-Aware Information Retrieval on the Internet) are reported. This project retrieves spatial information through geographical ontologies. Ontologies such as WordNet[2] have been widely exploited. In Leidner's work [4] , the main objective is to research how to reference spatial ambiguous entity names, which may be well founded or resolved, with respect to an extensional coordinated robust model of open news domain. Buscaldi and Rosso work [12] developed the conceptual density calculation, a measure of correlation between the meaning of a word and its context, and it is used for each candidate regarding the place's name. The reference that maximizes the conceptual density is then assigned to the ambiguous place name. The method uses hierarchical relationships such as hyponymy and hypernymy from WordNet. Among other works discussed previously, are those that establish lexical and semantic relations, the most common are: meronymy-holonymy, hyponymy-hypernymy and synonymy-antonymy. All relationships are represented by binary relations (correspondences between elements of the same set). In [5] a heuristic is presented for toponym disambiguation in texts, based on the quantification of the proximity of tree-names, figure 1 shows a part of this tree.

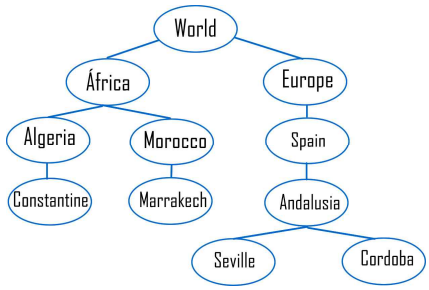

Figure 1. A part of the World hierarchical tree [5].

Ma A. Leite et al., works [6] propose mechanism for building ontologies, which explore a framework for encoding geographic knowledge base, consisting of multiple related ontologies, where relations are expressed as fuzzy as shown in figure 2.

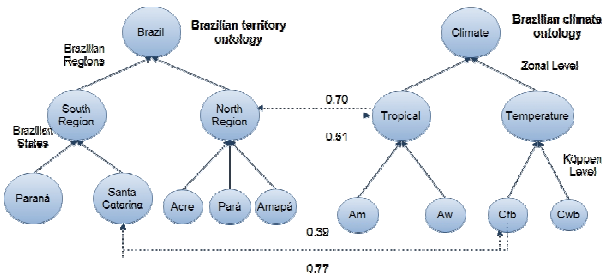

Figure 2. Brazilian territorial and Brazilian climate simple lightweight ontologies and their fuzzy associations [6].

A geographic space ontology which is based on an index structure is formalized in [7]. The main function of the ontology is to provide a vocabulary of classes and relationships to describe a specific area for the geographical space case. The ontology describes eight types of interest: SpatialThing, GeographicalThing, GeographicalRegion, GeopoliticalEntity, PopulatedPlace, Region, Country and

* This work has been partially supported by VIEP-BUAP project #SOGJ-ING11-I
[1] www.perseus.tufts.edu/hopper/
[2] http://wordnet.princeton.edu/

Continent. There exist hierarchical relationships between SpatialThing, GeographicalThing, and GeopoliticalEntity GeographicalRegion as shown in figure 2.

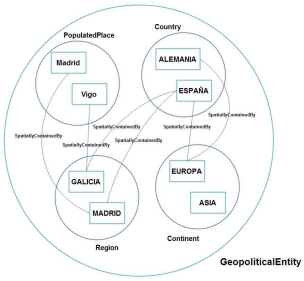

Figure 3. Ontology instances [7].

3. Ontology construction

As it was mentioned before, disambiguation of place names is the identification and categorization of their names. Our method for solving this task consist of using hierarchical and nonhierarchical relationship provided by a mixed Spanish ontology and at the same time using a hierarchical ontological weight, supplemented with the Haversine distance. Ontology provides a vocabulary of classes and relations to describe a specific area, in this case the geographical space. The goal is to build an ontology for Spanish-speaking countries, which include analysis of the distribution of the estimated 500 million Spanish speakers in the world.

When defining the ontology there are several ontological languages which provide different levels of formalism and reasoning facilities. The language OWL (WebOntology Language) was standardized by the W3C (World Wide WebConsortium) to define ontologies with various levels of detail. Such language can be categorized into three sublanguages: OWL-Lite, OWL-DL and OWL-Full. For the purpose of this research a geographic ontology was defined using OWL-DL, since it is designed for users who require maximum expressiveness while retaining computational completeness and solvability.

3.1. Definition of classes and subclasses

Classes are the basic units of an ontology which is intended to formalize. The ontology has built in seven main classes of interest: *Territorial_Division* from levels 1 to 4, *Continent*, *Country*, and *Abbreviation*. Figure 4, shows the radial extent of the classes of the toponyms ontology.

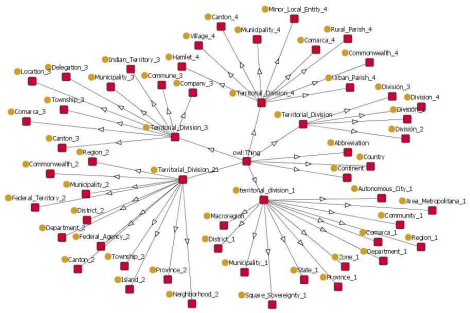

Figure 4. Radial extension of conceptual classes

3.2. Slots

Slots are related to search and sort, they are the fundamental objective of this study. Isolated classes such as *Continent*, *Country*, *State*, and so on do not provide enough information to answer questions like: *What countries belong to the American continent?, what continent each country belongs to?, what country belongs to certain group of states?*, *what is your continent and division level*?. Judging from this list of questions, the ontology includes information on several characteristics, called properties or slots. The main properties are:

- *is_part_of_the_continent*: to refer directly to the continent which the countries belong to.
- *is_part_of_the_country*: to refer directly to the country different levels of a territory belongs to.
- *is_part_of_the*: community, county, state, local, etc. to refer to different levels of territory.
- *has_abbreviation*: refers to the abbreviation of a particular territory.

In figure 5 some classes, instances and its relationships are shown.

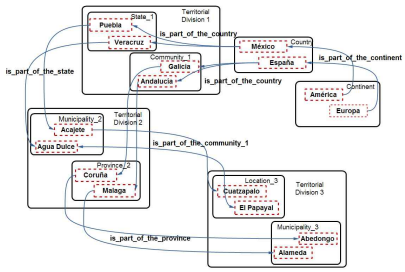

Figure 5. Some classes, instances and its relationships.

4. Disambiguation Algorithm

Although methods for toponym disambiguation are very different among them, they have factors in common and mostly influenced by two main phases. The next two steps are considered in the literature as the Generic Procedure for Toponym Disambiguation:

1. Extracting related candidates. All possible referents are extracted from a source of geographical knowledge (senses repositories, senses as Gazetter, ontology, etc.).
2. Choosing the correct referent. Apply a set of heuristics to determine, among all potential candidates, the most likely one, taking into account the context and resources as a source of evidence.

The proposed toponym disambiguation method for GIR queries in Spanish, considers seven important steps: 1) Contexts loading, 2) Preprocessing, 3) Searching for names in the ontological structure, 4) Toponyms ambiguity determination, 5) Searching for hierarchical and nonhierarchical relationships (holonymy, meronymy and sibling) for disambiguation case, 6) Application of hierarchical ontological weight and 7) Calculation of additional weight (Haversine distance). These seven steps as described in the following paragraphs.

Testing contexts all belong to online newspapers in Mexico, which are used for the following characteristics: rich in geo-spatial contents, reliable, frequently updated, have a kind of geographical ambiguity in its texts like (Valencia-Spain, Valencia-Venezuela), and also they contain spatial information (latitude and longitude). Context in Table 1 is presented in Spanish, since this work deals about a disambiguation method using ontologies and contexts for that language.

Table 1. Input context Sample

Context in Spanish
Muere mujer en fatal accidente ***** resumen ***** **Yanga**, **Veracruz**.- Tragica muerte fue la que tuvo una joven de 22 años de edad, luego de que el vehiculo que conducia, fuera impactado por una camioneta Ford Ranger debido a que su conductor al manejar en estado de ebriedad, ocasiono el accidente ya que la primera unidad termino impactandose contra un trailer tipo plataforma, sobre la carretera federal **Cordoba-Cuitlahuac** a la altura del libramiento a **Yanga**. De acuerdo al informe de la Policia Municipal, indica que la desgracia se registro ayer por la tarde cuando Elizabeth Vazquez Crisanto de 22 años de edad, al conducir un vehiculo Chevy con placas de circulacion YEZ-76-83 del estado en el cual lo acompañaba su hermano Santiago de 17 años, ambos con domicilio en el Municipio de **Cotaxtla**. Al viajar procedentes de la ciudad de **Cordoba** y con destino final a su domicilio, al llegar a la altura del kilometro 01 del mencionado libramiento, fueron impactados por la parte trasera del lado izquierdo por el conductor de una Ford Ranger, blanca con placas XH-98-213 del estado identificado como Jaime Marin con domicilio en la comunidad de **Mata Gallina** perteneciente a **Carrillo Puerto** ya que este ultimo conducia en estado de ebriedad.

The next step is to carry out a cleaning procedure. A cleaning procedure was performed, as the follows: each word in the context was converted to lowercase, written accents and exclamation marks were identified and eliminated, as well as repeated spaces, carriage returns, and stop words.

After the context was cleaned out, we look for toponyms inside ontology by using SPARQL[3] queries. In the query of Table 2, all the Spanish speaking continents are asked in the ontological structure of toponyms. What is relevant to this type of query is that you can get implicit knowledge from the explicit one.

Table 2. SPARQL query sample

Query
PREFIX toponyms: <http:// www.owl-ontologies.com/Toponimos.owl#> SELECT DISTINC Continent FROM http://www.owl-ontologies.com/Toponimos.owl> WHERE{ country toponyms:is_a_continent ?continent }ORDER BY ?name

Once the toponyms are found, a search is performed for each of them in the ontological structure, if it is found more than once, it is hypothesized that the toponym is ambiguous in the context, even for those compounds names. In Table 1 context it was found that *Carrillo Puerto Cordoba* toponyms appear more than once in the ontological structure, this indicates that the toponym is ambiguous in this context.

Toponyms can be ambiguous, as it was shown in the previous steps. It is then necessary, to look for each ambiguous toponym, its hierarchical and nonhierarchical relationships, such as holonymy (has-parts), meronymy (is-part -of) and sibling relationship (they share the same level of territorial division), by using SPARQL queries.

This step consists on to compare the hierarchical and non-hierarchical relationships with toponyms relationships with geographical proximity in the same context. Equation 1 shows how this weight is obtained; where for each hierarchical level that has agreement with the ambiguous toponym is weighted as follows:

$$f_i = \left\{ \frac{(1/n) * \sum_{i=1}^{n} f(i)}{m} \right\} \quad (1)$$

Where:

[3] hppt://www.w3.org/TR/rdf-sparql-query

- n= number of analyzed relationships
- m= maximum number of siblings found for ambiguous toponyms
- h= number of siblings for the toponym under analysis
-

5.2 Making additional weight (Haversine distance)

Haversine's formula is an important equation in celestial navigation. It helps to calculate the great circle distance between two points of the globe from its longitude and latitude.

$$\Delta lat = lat_2 - lat_1$$

$$\Delta long = long_2 - long_1$$

$$a = sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat_1.\cos(lat_2(\Delta long/2)))$$

$$c = 2.atan_2(\sqrt{a}.\sqrt{1-a})$$

$$d = R.c \quad (2)$$

Where:

- d= distance between points (along a great circle of the sphere)
- Δlat= the difference in latitude
- Δlong= the difference in longitude
- lat=latitude
- long=longitude
- R= radius of the sphere

To calculate the Haversine distance, two points are taken, for example in table 1 context, to disambiguate the toponym *Puerto Carrillo* (*Veracruz*) from the toponym *Felipe Carrillo Puerto* (*Quintana Roo*), both toponyms will take as a reference, the non-hierarchical relationship of "siblings" (toponym located *Yanga*), to calculate this distance, see table 3.

Table 3. Latitude and longitude data

Data	Felipe Carrillo Puerto (Quintana Roo)	Puerto Carrillo (Veracruz)	Yanga (Veracruz)
Latitude	19.172717	19.181739	18.830226
Longitude	-96.133269	-88.479138	-96.799306

The weighting process is performed by comparing the Haversine distance between each ambiguous toponym and their brothers who were located for the toponym which obtained the maximum value in the ontological hierarchical weighting. The figure 8 shows the Haversine's distance calculation applied to the ambiguous toponyms *Carrillo Puerto* and *Felipe Carrillo Puerto*. The performed calculation showed that the minimum distance was found between the points *Puerto Carrillo* and *Yanga* **(**brother located in the context of ambiguous toponym) and it was 79.71km. This result confirms the research hypothesis, existence of geographic proximity between toponyms in the same context.

Figure 6. Haversine's distance for ambiguous toponyms.

6. Experimental Results

The ontology has the advantage of having a structure that can be queried with semantics (holonymy, meronymy, etc.) and not just by simple keywords. Experiments will be made later, being labeled the collection of articles relating to the right of each ambiguous toponym. Table 4 shows statistics for the test collection.

Table 4 Information about the contexts used in experiments

Data	Measures
Total number of toponyms	2352
Toponyms ambiguous	672
Total number of contexts analyzed	200
Average toponyms per context	11.76
Unduplicated number of toponyms in the same context	6
Average unduplicated toponyms per context	4.14
Number of duplications with different senses in the same context	2.2

7. Conclusions

This article presented a novel method for toponym disambiguation, which was designed to be used in texts written in Spanish, since there is little research on the issue of disambiguation in comparison with other languages, specifically English. Having this toponym disambiguation method for the Spanish language, opens up new topics of research in different areas. On the other hand, it has been achieved in understanding the role of semantic relationships, hierarchical and nonhierarchical, between toponyms in the same context.

8. Future Work

The future research aims toward improving the ontology of toponym ontology and the disambiguation method. Regarding the ontology, this considers the incorporation of spatial relationships. That is, allows searching for geographic information according to the semantic space and through the Web. For example: "Epidemics in Mexico" where the *inside* relationship is processed as a topological relationship. The approach proposed can be implemented as an extension of existing location-based services (eg, Yahoo Local, Google Local, etc.), where in addition to displaying a geographical location, a description could be included in the conceptual hierarchy or geographical area (Continent-Country-State) of a geographic object. Finally, it suggests that the disambiguation method could include the concepts of context distance and size where the influence of the weight of words according to the distance from the toponym under analysis could also be considered, so that closer

words have more influence and more distant words have less influence, and thus not just consider the geographic proximity between toponyms in the same context.

9. References

[1] Leonardo Andrade and Mário J. Silva. "Relevance ranking for geographic IR", Proceedings of sigir, pp. 64~67, Agosto 2006.
[2] David A. Smith and Gregory Crane. "Disambiguation geographic names in a historical digital library", Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, Vol. 2163, pp. 127~136, 2001.
[3] Ruas A. Sanderson M. Sester M. Van Kreveld M. Jones C., Purves R. And Weibel R. "Spatial information retrieval and geographical ontologies an overview of the spirit project", Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 387~388, 2002.
[4] Jochen L. Leidner. "Toponym resolution in text (annotation, evaluation and applications of spatial grounding)", Newsletter ACM SIGIR Forum., Vol. 41, No. 2, pp. 124~126, 2007.
[5] Imene Bensalem and Mohamed-Khireddine Kholladi. "Toponym disambiguation by arborescent relationships", Journal of Computer Science, Vol. 6, No. 6, pp. 631~637, 2010.
[6] Maria Angelica A. Leite and Ivan L. M. Ricarte. "Document Retrieval using fuzzy related geographic ontologies", In Proceedings of the 2nd International Workshop on Geographic Information Retrieval, GIR'08, pp 47~54, New York, NY, USA, 2008.AC.
[7] Diego Seco Naveiras. "Técnicas de indexación y Recuperación de Documentos utilizando referencias Geográficas y Textuales". PhD thesis, Universidad de Coruña, Departamento de Computación. 2008.
[8] Mark Sanderson and Janet Kohler. "Analyzing geographic queries", Proceedings of the Workshop on Geographic Information Retrieval, Sheffield, UK. SIGIR, 2004.
[9] Spink, A., Wolfram, D., Jansen, B.J. Saracevic, T. "Searching the Web: The Public and Their Queries", Journal of the American Society for Information Science and Technology, Vol. 52, No. 3, pp. 226~234.
[10] Davide Buscaldi, Toponym Disambiguation in Information Retrieval, Procesamiento del Lenguaje Natural. Vol. 46, No. 02, pp. 125~126, 2011.
[11] Miguel Félix Mata Rivera. "Recuperación y ponderación de información geográfica desde repositorios no estructurados conducidas por ontologías". Instituto Politécnico Nacional, pp. 138~140, 2009.
[12] Buscaldi, D. and P. Rosso. "A conceptual density-based approach for the disambiguation of toponyms." Int. Journal of Geographic Information Science, 22:301~313, 2008.
[13] Che-Yu Yang, Hua-Yi Lin, "Semantic Annotation for the Webof Data: An Ontology and RDF based Automated Approach", JCIT: Journal of Convergence Information Technology, Vol. 6, No. 4, pp. 318 ~ 327, 2011.
[14] Chunchen Liu, Dayou Liu, Shengsheng Wang, "A Fuzzy Geospatial Ontology Model and Its Application in Geospatial Semantic Retrieval", JCIT: Journal of Convergence Information Technology, Vol. 6, No. 6, pp. 85 ~ 97, 2011.