Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 125-126 recibido 04-01-11 revisado 21-02-11 aceptado 24-02-11

Toponym Disambiguation in Information Retrieval*

Desambiguación de Topónimos en la Recuperación de Información

Davide Buscaldi
DSIC - Universidad Politécnica de Valencia

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans, Orléans (France)
davide.buscaldi@univ-orleans.fr

Resumen: Tesis doctoral (con mención de doctorado europeo) en Informática realizada por Davide Buscaldi y dirigida por el doctor Paolo Rosso (Univ. Politécnica de Valencia). El acto de defensa de la tesis tuvo lugar en Valencia en Octubre de 2010 ante el tribunal formado por los doctores: Paul David Clough (University of Sheffield), Ross Purves (Universität Zürich), Emilio Sanchis Arnal (Univ. Politécnica de Valencia), Mark Sanderson (Royal Melbourne Institute of Technology), Diana Santos (Sintef-ICT, Oslo). La mención europea se obtuvo a través de una estancia en el FBK-IRST (Italia) bajo la dirección de Bernardo Magnini. La calificación obtenida fue de *Sobresaliente Cum Laude*.
Palabras clave: Recuperación de Información Geográfica, Desambiguación de Topónimos, Geoinformática

Abstract: Ph.D. thesis (European doctorate mention) in Computer Science written by Davide Buscaldi under the supervision of Dr. Paolo Rosso (Univ. Politécnica de Valencia). The author was examined in Valencia in October 2010 by a panel composed by the following doctors: Paul David Clough (University of Sheffield), Ross Purves (Universität Zürich), Emilio Sanchis Arnal (Univ. Politécnica de Valencia), Mark Sanderson (Royal Melbourne Institute of Technology), Diana Santos (Sintef-ICT, Oslo). The European mentions was received after a 3 months stage at the FBK-IRST (Italy) under the guidance of Bernardo Magnini. The obtained grade was *Cum Laude*.
Keywords: Geographical Information Retrieval, Toponym Disambiguation, Geoinformatics

1 Introduction

In this Ph.D. thesis we have investigated the effects of resolving ambiguous place names (a task commonly referred to as *Toponym Disambiguation* (TD)) in some Information Retrieval (IR) tasks, more specifically the Geographical Information Retrieval (GIR) and Question Answering (QA) tasks.

In order to achieve this goal, we have studied different toponym repositories, mostly WordNet[1] and Geonames[2]. We developed toponym disambiguation methods, applying them to the GeoCLEF[3] collection, evaluating results in IR with the queries of past GeoCLEF campaigns and in QA with questions from past CLEF-QA campaigns. We have studied the following problems:

1. The difference in performance of different TD methods on the same collection;
2. The difference in performance of the same TD method, using different toponym repositories, on the same collection;
3. How these differences may affect the overall performance of the GIR and QA tasks;
4. How the overall performance in GIR and QA scales with respect to the number of errors introduced in TD;
5. How task granularity may affect the performance in TD (disambiguation at

* Thesis supported by a FPI Grant of the Valencian government (ref. BFPI06/97).

[1] http://wordnet.princeton.edu

[2] http://www.geonames.org

[3] http://ir.shef.ac.uk/geoclef/

street level implies more difficulties that disambiguation at city level).

In order to carry out part of the work, we produced two resources that have been freely released: Geo-WordNet[4], a mapping of WordNet synsets to geographical coordinates extracted from Geonames; and GeoSemCor, a version of SemCor where the geographical entities have been automatically tagged. Part of the work resulted in the development of a prototype of a geographically-aware search engine, *Geooreka*[5].

2 Thesis Overview

In this thesis, we have addressed the challenges raised by the ambiguity of toponyms in Information Retrieval. The work is structured as follows:

In Chapter 1, we introduce the objectives of the thesis and detail the motivations of this work and how it relates with similar works. We also explain the possible applications of TD in various tasks and which problems may arise from the ambiguity of place names in such applications. Chapter 2 provides an overview of Information Retrieval and its evaluation.

Chapter 3 is dedicated to the most important resources used as toponym repositories: gazetteers and geographic ontologies. Moreover, the chapter provides an overview of the currently existing text corpora in which toponyms have been labelled with geographical coordinates: GeoSemCor, CLIR-WSD, TR-CoNLL and SpatialML.

In Chapter 4, the focus is on the ambiguity of toponyms and the methods for its resolution; two different methods, one based on WordNet and another based on map distances, were presented and compared over the GeoSemCor corpus. A case study related to the disambiguation of toponyms in an Italian local news collection is also presented in this chapter.

In Chapter 5, we detail the experiments that we carried out in order to study the relation between GIR and toponym disambiguation, especially to understand under which conditions toponym disambiguation may help, and how disambiguation errors affects the retrieval results.

In Chapter 6 we studied the effects of TD on Question Answering, using the SemQUASAR QA engine as a base system. In Chapter 7, we present the geographical web search engine Geooreka! and discuss the importance of the disambiguation of toponyms in the web. Finally, in Chapter 8 we summarise the contributions of the Ph.D. thesis and present some ideas for further work.

3 Thesis Contributions

The most important findings of the research work are the following ones: *off-the-shelf gazetteers are not enough*, by themselves, to cover the needs of toponym disambiguation above a certain detail, especially when the toponyms to be disambiguated are road names or vernacular names. In the “L’Adige” news collection, about 10% of toponyms were road or street names. The probability of a street name to be ambiguous was calculated to be 0.83, with respect to other classes of toponyms which presented an ambiguity of 0.58. On average, the resource built for this task presented an average ambiguity of 4.68 referents per toponym, compared to an average of 1.42 of a resource such as Geonames.

The second finding is that *the ambiguity level that is found in resources like WordNet does not represent a problem*: all referents can be used in the indexing phase to expand the index without affecting the overall performance. In Table 1 we show the results calculated on the GeoCLEF test collection, with toponym disambiguation or not, using WordNet or Geonames as toponym repository.

	Disamb.	Not Dis.
WordNet	0,217	0,221
Geonames	0,226	0,220

Table 1: Average MAP on GeoCLEF.

Finally, we found that *disambiguation is useful only in the case of short queries with a detailed toponym repository*, reflecting the working configuration of web search engines. The results in Table 1 were calculated using only the topic title, where difference in average MAP between Geonames and WordNet-based runs was statistically relevant (according to Student’s t-test at 95%). The preliminary results obtained with Geooreka! confirms that TD is a crucial feature in geographically constrained web searches.

[4] Listed in the official WordNet “related projects” page: *http://wordnet.princeton.edu/wordnet/related-projects/*

[5] http://www.geooreka.eu