Toponym Disambiguation in Information Retrieval

Davide Buscaldi
Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

Under the supervision of

Dr. Paolo Rosso

2010 October

Abstract

In recent years, geography has acquired a great importance in the context of Information Retrieval (IR) and, in general, of the automated processing of information in text. Mobile devices that are able to surf the web and at the same time inform about their position are now a common reality, together with applications that can exploit these data to provide users with locally customised information, such as directions or advertisements. Therefore, it is important to deal properly with the geographic information that is included in electronic texts. The majority of such kind of information is contained as place names, or toponyms.

Toponym ambiguity represents an important issue in Geographical Information Retrieval (GIR), due to the fact that queries are geographically constrained. There has been a struggle to find specific geographical IR methods that actually outperform traditional IR techniques. Toponym ambiguity may constitute a relevant factor in the inability of current GIR systems to take advantage from geographical knowledge. Recently, some Ph.D. theses have dealt with Toponym Disambiguation (TD) from different perspectives, from the development of resources for the evaluation of Toponym Disambiguation (Leidner (2007)) to the use of TD to improve geographical scope resolution (Andogah (2010)). The Ph.D. thesis presented here introduces a TD method based on WordNet and carries out a detailed study of the relationship of Toponym Disambiguation to some IR applications, such as GIR, Question Answering (QA) and Web retrieval.

The work presented in this thesis starts with an introduction to the applications in which TD may result useful, together with an analysis of the ambiguity of toponyms in news collections. It could not be possible to study the ambiguity of toponyms without studying the resources that are

used as placename repositories; these resources are the equivalent to language dictionaries, which provide the different meanings of a given word. An important finding of this Ph.D. thesis is that the choice of a particular toponym repository is key and should be carried out depending on the task and the kind of application that it is going to be developed. We discovered, while attempting to adapt TD methods to work on a corpus of local Italian news, that a factor that is particularly important in this choice is represented by the "locality" of the text collection to be processed. The choice of a proper Toponym Disambiguation method is also key, since the set of features available to discriminate place references may change according to the granularity of the resource used or the available information for each toponym. In this work we developed two methods, a knowledge-based method and a map-based method, which compared over the same test set.

We studied the effects of the choice of a particular toponym resource and method in GIR, showing that TD may result useful if query length is short and a detailed resource is used. We carried out some experiments on the CLEF GIR collection, finding that retrieval accuracy is not affected significantly, even when the errors represent 60% of the toponyms in the collection, at least in the case in which the resource used has a little coverage and detail. Ranking methods that sort the results on the basis of geographical criteria were observed to be more sensitive to the use of TD or not, especially in the case of a detailed resource. We observed also that the disambiguation of toponyms does not represent an issue in the case of Question Answering, because errors in TD are usually less important than other kind of errors in QA.

In GIR, the geographical constraints contained in most queries are area constraints, such that the information need usually expressed by users can be resumed as "X in P", where P is a place name, and X represents the thematic part of the query. A common issue in GIR occurs when a place named by a user cannot be found in any resource because it is a fuzzy region or a vernacular name. In order to overcome this issue, we developed

Geooreka!, a prototype search engine with a map-based interface. A preliminary testing of this system is presented in this work. The work carried out on this search engine showed that Toponym Disambiguation can be particularly useful on web documents, especially for applications like Geooreka! that need to estimate the occurrence probabilities for places.

Abstract

En los últimos años, la geografía ha adquirido una importancia cada vez mayor en el contexto de la recuperación de la información (Information Retrieval, IR) y, en general, del procesamiento de la información en textos. Cada vez son más comunes dispositivos móviles que permiten a los usuarios de navegar en la web y al mismo tiempo informar sobre su posición, así como las aplicaciones que puedan explotar estos datos para proporcionar a los usuarios algún tipo de información localizada, por ejemplo instrucciones para orientarse o anuncios publicitarios. Por tanto, es importante que los sistemas informáticos sean capaces de extraer y procesar la información geográfica contenida en textos electrónicos. La mayor parte de este tipo de información está formado por nombres de lugares, llamados también *topónimos*.

La ambigüedad de los topónimos constituye un problema importante en la tarea de recuperación de información geográfica (Geographical Information Retrieval o GIR), dado que en esta tarea las peticiones de los usuarios están vinculadas geográficamente. Ha habido un gran esfuerzo por parte de la comunidad de investigadores para encontrar métodos de IR específicos para GIR que sean capaces de obtener resultados mejores que las técnicas tradicionales de IR. La ambigüedad de los topónimos es probablemente un factor muy importante en la incapacidad de los sistemas GIR actuales por conseguir una ventaja a través del procesamiento de las informaciones geográficas. Recientemente, algunas tesis han tratado el problema de resolución de ambigüedad de topónimos desde distintas perspectivas, como el desarrollo de recursos para la evaluación de los métodos de desambiguación de topónimos (Leidner) y el uso de estos métodos para mejorar la resolución de lo "scope" geográfico en documentos electrónicos (Andogah).

En esta tesis se ha introducido un nuevo método de desambiguación basado en WordNet y por primera vez se ha estudiado atentamente la ambigüedad de los topónimos y los efectos de su resolución en aplicaciones como GIR, la búsqueda de respuestas (Question Answering o QA), y la recuperación de información en la web.

Esta tesis empieza con una introducción a las aplicaciones en las cuales la desambiguación de topónimos puede producir resultados útiles, y con una análisis de la ambigüedad de los topónimos en las colecciones de noticias. No sería posible estudiar la ambigüedad de los topónimos sin estudiar también los recursos que se usan como bases de datos de topónimos; estos recursos son el equivalente de los diccionarios de idiomas, que se usan para encontrar los significados diferentes de una palabra. Un resultado importante de esta tesis consiste en haber identificado la importancia de la elección de un particular recurso, que tiene que tener en cuenta la tarea que se tiene que llevar a cabo y las características específicas de la aplicación que se está desarrollando. Se ha identificado un factor especialmente importante constituido por la "localidad" de la colección de textos a procesar. La elección de un algoritmo apropiado de desambiguación de topónimos es igualmente importante, dado que el conjunto de "features" disponible para discriminar las referencias a los lugares puede cambiar en función del recurso elegido y de la información que este puede proporcionar para cada topónimo. En este trabajo se desarrollaron dos métodos para este fin: un método basado en la densidad conceptual y otro basado en la distancia media desde centroides en mapas. Ha sido presentado también un caso de estudio de aplicación de métodos de desambiguación a un corpus de noticias en italiano.

Se han estudiado los efectos derivados de la elección de un particular recurso como diccionario de topónimos sobre la tarea de GIR, encontrando que la desambiguación puede resultar útil si el tamaño de la query es pequeño y el recurso utilizado tiene un elevado nivel de detalle. Se ha descubierto que el nivel de error en la desambiguación no es relevante, al menos hasta el 60% de errores, si el recurso tiene una cobertura pequeña y un nivel de

detalle limitado. Se observó que los métodos de ordenación de los resultados que utilizan criterios geográficos son más sensibles a la utilización de la desambiguación, especialmente en el caso de recursos detallados. Finalmente, se detectó que la desambiguación de topónimos no tiene efectos relevantes sobre la tarea de QA, dado que los errores introducidos por este proceso constituyen una parte trascurable de los errores que se generan en el proceso de búsqueda de respuestas.

En la tarea de recuperación de información geográfica, la mayoría de las peticiones de los usuarios son del tipo "$XenP$", dónde P representa un nombre de lugar y X la parte temática de la query. Un problema frecuente derivado de este estilo de formulación de la petición ocurre cuando el nombre de lugar no se puede encontrar en ningún recurso, tratándose de una región delimitada de manera difusa o porqué se trata de nombres vernáculos. Para solucionar este problema, se ha desarrollado Geooreka!, un prototipo de motor de búsqueda web que usa una interfaz gráfica basada en mapas. Una evaluación preliminar se ha llevado a cabo en esta tesis, que ha permitido encontrar una aplicación particularmente útil de la desambiguación de topónimos, la desambiguación de los topónimos en los documentos web, una tarea necesaria para estimar correctamente las probabilidades de encontrar ciertos lugares en la web, una tarea necesaria para la minería de texto y encontrar información relevante.

Abstract

En els últims anys, la geografia ha adquirit una importància cada vegada major en el context de la recuperaci de la informació (Information Retrieval, IR) i, en general, del processament de la informaci en textos. Cada vegada són més comuns els dispositius mòbils que permeten als usuaris navegar en la web i al mateix temps informar sobre la seua posició, així com les aplicacions que poden explotar aquestes dades per a proporcionar als usuaris algun tipus d'informació localitzada, per exemple instruccions per a orientar-se o anuncis publicitaris. Per tant, és important que els sistemes informàtics siguen capaços d'extraure i processar la informació geogràfica continguda en textos electrònics. La major part d'aquest tipus d'informació est format per noms de llocs, anomenats també topònims.

L'ambigüitat dels topònims constitueix un problema important en la tasca de la recuperació d'informació geogràfica (Geographical Information Retrieval o GIR, atès que en aquesta tasca les peticions dels usuaris estan vinculades geogràficament. Hi ha hagut un gran esforç per part de la comunitat d'investigadors per a trobar mètodes de IR específics per a GIR que siguen capaos d'obtenir resultats millors que les tècniques tradicionals en IR. L'ambigüitat dels topònims és probablement un factor molt important en la incapacitat dels sistemes GIR actuals per a aconseguir un avantatge a través del processament de la informació geogràfica. Recentment, algunes tesis han tractat el problema de resolució d'ambigüitat de topònims des de diferents perspectives, com el desenvolupament de recursos per a l'avaluació dels mètodes de desambiguació de topònims (Leidner) i l'ús d'aquests mètodes per a millorar la resolució del "scope" geogràfic en documents electrònics (Andogah). L'objectiu d'aquesta tesi és estudiar l'ambigüitat dels topònims i els efectes de la seua resolució en aplicacions com en la tasca GIR, la cerca

de respostes (Question Answering o QA), i la recuperació d'informació en la web.

Aquesta tesi comena amb una introducció a les aplicacions en les quals la desambiguació de topònims pot produir resultats útils, i amb un anàlisi de l'ambigüitat dels topònims en les col.leccions de notícies. No seria possible estudiar l'ambigüitat dels topònims sense estudiar també els recursos que s'usen com bases de dades de topònims; aquests recursos són l'equivalent dels diccionaris d'idiomes, que s'usen per a trobar els diferents significats d'una paraula. Un resultat important d'aquesta tesi consisteix a haver identificat la importància de l'elecció d'un particular recurs, que ha de tenir en compte la tasca que s'ha de portar a terme i les característiques específiques de l'aplicació que s'està desenvolupant. S'ha identificat un factor especialment important constitut per la "localitat" de la collecció de textos a processar. L'elecció d'un algorisme apropiat de desambiguació de topnims és igualment important, atès que el conjunt de "features" disponible per a discriminar les referències als llocs pot canviar en funció del recurs triat i de la informació que aquest pot proporcionar per a cada topnim. En aquest treball es van desenvolupar dos mètodes per a aquesta fi: un mètode basat en la densitat conceptual i altre basat en la distància mitja des de centroides en mapes. Ha estat presentat també un cas d'estudi d'aplicació de mètodes de desambiguació a un corpus de notícies en italià.

S'han estudiat els efectes derivats de l'elecció d'un particular recurs com diccionari de topònims sobre la tasca de GIR, trobant que la desambiguació pot resultar útil si la query és menuda i el recurs utilitzat té un elevat nivell de detall. S'ha descobert que el nivell d'error en la desambiguació no és rellevant, almenys fins al 60% d'errors, si el recurs té una cobertura menuda i un nivell de detall limitat. Es va observar que els mètodes d'ordenació dels resultats que utilitzen criteris geogràfics són més sensibles a la utilització de la desambiguació, especialment en el cas de recursos detallats. Finalment, es va detectar que la desambiguació de topnims no té efectes rellevants sobre la tasca de QA, atès que els errors introduïts per aquest procés constitueixen

una part trascurable dels errors que es generen en el procés de recerca de respostes.

En la tasca de recuperació d'informació geogràfica, la majoria de les peticions dels usuaris són del tipus "X en P", on P representa un nom de lloc i X la part temàtica de la query. Un problema freqüent derivat d'aquest estil de formulació de la petició ocorre quan el nom de lloc no es pot trobar en cap recurs, tractant-se d'una regió delimitada de manera difusa o perqu es tracta de noms vernacles. Per a solucionar aquest problema, s'ha desenvolupat "Geooreka!", un prototip de motor de recerca web que usa una interfície gràfica basada en mapes. Una avaluació preliminar s'ha portat a terme en aquesta tesi, que ha permès trobar una aplicació particularment útil de la desambiguació de topònims, la desambiguació dels topònims en els documents web, una tasca necessària per a estimar correctament les probabilitats de trobar certs llocs en la web, una tasca necessària per a la mineria de text i trobar informació rellevant.

The limits of my language mean the limits of my world

Ludwig Wittgenstein,

Tractatus Logico-Philosophicus 5.6

Supervisor:	Dr. Paolo Rosso
Panel:	Dr. Paul Clough
	Dr. Ross Purves
	Dr. Emilio Sanchis
	Dr. Mark Sanderson
	Dr. Diana Santos

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Glossary

ASR Automated Speech Recognition.

GAR Geographically Adjusted Ranking.

Gazetteer A list of names of places, usually with additional information such as geographical coordinates and population.

GCS Geographic Coordinate System: a coordinate system that allows to specify every location on Earth in three coordinates.

Geocoding The process of finding associated geographic coordinates, usually expressed as latitude and longitude, from other geographic data, such as street addresses, toponyms, or postal codes.

Geographic Footprint The geographic area that is considered relevant for a given query.

Geotagging The process of adding geographical identification metadata to various media such as photographs, video, websites, RSS feeds...

GIR Geographic (or Geographical) Information Retrieval : the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope (in Purves and Jones (2006)).

GIS Geographic Information System: any information system that integrates, stores, edits, analyzes, shares, and displays geographic information. In a more generic sense, GIS applications are tools that allow users to create interactive queries (user created searches), analyze spatial information, edit data, maps, and present the results of all these operations.

GKB Geographical Knowledge Base: a database of geographic names which includes some relationship among the place names.

IR Information Retrieval: the science that deals with the representation, storage, organization of, and access to information items (in Baeza-Yates and Ribeiro-Neto (1999)).

LBS Location Based Service: a service that exploits positional data from a mobile device in order to provide certain information to the user.

MAP Mean Average Precision.

MRR Mean Reciprocal Rank.

NE Named Entity: textual tokens that identify a specific "entity, usually a person, organization, location, time or date, quantity, monetary value, percentage.

NER Named Entity Recognition: NLP techniques used for identifying Named Entities in text.

NERC Named Entity Recognition and Classification: NLP techniques used for the identifiying Named Entities in text and assigning them a specific class (usually person, location or organization).

LIST OF TABLES

NLP Natural Language Processing: a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

QA Question Answering: a field of IR where the information need of a user is expressed by mean of a natural language question and the result is a concise and precise answer in natural language.

Reverse geocoding The process of back (reverse) coding of a point location (latitude, longitude) to a readable address or place name.

TD Toponym Disambiguation: the process of assigning the correct geographic referent to a place name.

TR Toponym Resolution: see *TD*.

1

Introduction

Human beings are familiar with the concepts of *space* and *place* in their everyday life. These two concepts are similar but at the same time different: a space is a three-dimensional environment in which objects and events occur, where they have relative position and direction. A place is itself a space, but with some added meaning, usually depending on culture, convention and the use made of that space. For instance, a city is a place determined by boundaries that have been established by their inhabitants, but it is also a space since it contains buildings and other kind of places, such as parks and roads. Usually, people move to one place to another to work, to study, to get in contact with other people, to spend free time during holidays, and to carry out many other activities. Even without moving, we receive everyday information about some event that occurred in some place. It would be impossible to carry out such activities without knowing the names of the places. Paraphrasing Wittgenstein, "We can not go to any place we can not talk about"[1]. This information need may be considered as one of the roots of the science of geography. The etymology of the word *geography* itself, "to describe or write about the Earth", reminds of this basic problem. It was the Greek philosopher Eratosthenes who coined the term "geography". He and others ancient philosophers regarded Homer as the founder of the science of geography, as accounted by Strabo (1917) in his "Geography" $(i, 1, 2)$, because he gave, in the "Iliad" and the "Odyssey", descriptions of many places around the Mediterranean Sea. The

[1]The original proposition as formulated by Wittgenstein was "What we cannot speak about we must pass over in silence" Wittgenstein (1961)

geography of Homer had an intrinsic problem: he named places but the description of where they were located was, in many cases, confuse or missing.

A long time has passed since the age of Homer, but little has changed in the way of representing places in text: we still use *toponyms*. A toponym is, literally, a place name, as its etymology says: *τόπος* (place) and *όνυμα* (name). Toponyms are contained in almost every piece of information in the Web and in digital libraries: almost every news story contains some reference, in an explicit or implicit way, to some place on Earth. If we consider places to be objects, the semantics of toponyms is pretty simple, if compared to words that represent concepts, such as "happiness" or "truth". Sometimes, toponyms meanings are more complex because there is no agreement on their boundaries, or because they may have a particular meaning that is perceived subjectively (for instance, people that inhabits some place will give it also a "home" meaning). However, in most cases, for practical reasons, we can approximate the meaning of a toponym with a set of coordinates in a map, which represent the location of the place in the world. If the place can be approximated to a point, then its representation is just a $2-uple$ (*latitude, longitude*). Just as for the meanings of other words, the "meaning" of a toponym is listed in a dictionary[1]. The problems of using toponyms to identify a geographical entity are related mostly to ambiguity, synonymy and the fact that names change over time.

The ambiguity of human language is one of the most challenging problems in the field of Natural Language Processing (NLP). With respect to toponyms, ambiguity can be of various types: a proper name may identify different class of named entities (for instance, 'London' may identify the writer 'Jack London' or a city in the UK), or may be used as a name for different instances of a same class; e.g. 'London' is also a city in Canada. In this case we talk about *geo-geo* ambiguity and this is the kind of ambiguity addressed in this thesis. The task of resolving *geo-geo* ambiguities is called *Toponym Disambiguation* (TD) or *Toponym Resolution* (TR). Many studies show that the number of ambiguous toponyms is greater than one would expect: Smith and Crane (2001) found that 57.1% of toponyms used in North America are ambiguous. Garbin and Mani (2005) studied a news collection from Agence France Press, finding that 40.1% of toponyms used in the collection were ambiguous, and in 67.8% of the cases they could not resolve ambiguity. Two toponyms are synonyms where they are different names referring to the same place. For instance, "Saint Petersburg" and "Leningrad" are two toponyms that indicates the same city. In this example, we also see that toponyms are not fixed but change over time.

[1] dictionaries mapping toponyms to coordinates are called *gazetteers* - cfr. Chapter 3

The growth of the world wide web implies a growth of the geographical data contained in it, including toponyms, with the consequence that the coverage of the places named in the web is continuously growing over time. Moreover, since the introduction of map-based search engines (Google Maps[1] was launched in 2004) and their diffusion, displaying, browsing and searching information on maps have become common activities. Some recent studies show that many users submit queries to search engines in search for geographically constrained information (such as "Hotels in New York"). Gan et al. (2008) estimated that 12.94% of queries submitted to the AOL search engine were of this type; Sanderson and Kohler (2004) found that 18.6% of the queries submitted to the Excite search engine contained at least a geographic term. More recently, the spreading of portable GPS-based devices and, consequently, of location-based services (Yahoo! FireEagle[2] or Google Latitude[3]) that can be used with such devices is expected to boost the quantity of geographic information available on the web and introduce more challenges for the automatic processing and analysis of such information.

In this scenario, toponyms are particularly important because they represent the bridge between the world of Natural Language Processing and Geographic Information Systems (GIS). Since the information on the web is intended to be read by human users, usually the geographical information is not presented by means of geographical data, but using text. For instance, is quite uncommon in text to say "$41.9^oN12.5^oE$" to refer to "Rome, Italy". Therefore, automated systems must be able to disambiguate toponyms correctly in order to improve in certain tasks such as searching or mining information.

Toponym Disambiguation is a relatively new field. Recently, some Ph.D. theses have dealt with TD from different perspectives. Leidner (2007) focused on the development of resources for the evaluation of Toponym Disambiguation, carrying out some experiments in order to compare a previous disambiguation method to a simple heuristic. His main contribution is represented by the TR-CoNLL corpus, which is described in Section 3.4.3. Andogah (2010) focused on the problem of geographical scope resolution: he assumed that every document and search query have a geographical scope, indicating where the events described are situated. Therefore, he aimed his efforts to exploit the notion of geographical scope. In his work, TD was considered in order to enhance the scope determination process. Overell (2009) used Wikipedia[4]

[1] http://maps.google.com

[2] `http://fireeagle.yahoo.net/`

[3] `http://www.google.com/latitude`

[4] `http://www.wikipedia.org`

to generate a tagged training corpus that was applied to supervised disambiguation of toponyms based on co-occurrences model. Subsequently, he carried out a comparative evaluation of the supervised disambiguation method with respect to simple heuristics, and finally he developed a Geographical Information Retrieval (GIR) system, Forostar, which was used to evaluate the performance of GIR using TD or not. He did not find any improvements in the use of TD, although he was not able to explain this behaviour.

The main objective of this Ph.D. thesis consists in giving an answer to the question "*under which conditions may toponym disambiguation result useful in Information Retrieval (IR) applications?*".

In order to reply to this question, it is necessary to study TD in detail, and understand what is the contribution of resources, methods, collections, and the granularity of the task over the performance of TD in IR. Using less detailed resources greatly simplifies the problem of TD (for instance, if Paris is listed only as the French one), but on the other side, it can produce a loss of information that deteriorates the performance in IR. Another important research question is "*Can results obtained on a specific collection be generalised to other collections, too?*". The previously listed theses did not discuss these problems, while this thesis is focused on them.

Speculations that the application of TD can produce an improvement of the searches both in the web or in large news collections have been made by Leidner (2007), who also attempted to identify some applications that could benefit from the correct disambiguation of toponyms in text:

- *Geographical Information Retrieval*: it is expected that toponym disambiguation may increase precision in the IR field, especially in GIR, where the information needs expressed by users are spatially constrained. This expectation is based on the fact that, by being able to distinguish documents referring to one place from another with the same name, the accuracy of the retrieval process would increase.

- *Geographical Diversity Search*: Sanderson et al. (2009) noted that current IR techniques fail to retrieve documents that may be relevant to distinct interpretations of their search terms, or, in other words, they do not support "diversity search". In the Geographical domain, "spatial diversity" is a specific case, where a user can be interested in the same topic over a different set of places (for instance: "brewing industry in Europe") and a set of document for each place can be more useful than a list of documents covering the entire relevance area.

- *Geographical document browsing*: this aspect embraces GIR from another point of view, that of the interface that connects the user to the results. Documents

containing geographical information can be accessed by means of a map in an intuitive way.

- *Question Answering*: toponym resolution provides a basis for geographical reasoning. Firstly, questions of a spatial nature (Where is X? What is the distance between X and Y?) can be answered more systematically (rather than having to rely on accidental explicit text spans mentioning the answer).

- *Location-Based Services*: as GPS-enabled mobile computing devices with wireless networking are becoming pervasive, it is possible for the user to use its current location to interact with services on the web that are relevant to his or her position (including location-specific searches, such as "where's the next hotel/restaurant/post office round here?").

- *Spatial Information Mining*: frequency of co-occurrences of events and places may be used to extract useful information from texts (for instance, if we can search "forest fires" on a map and we find that some places co-occur more frequently than others for this topic, then these places should retain some characteristics that make them more sensible to forest fires).

Most of these areas were already identified by Leidner (2007), who considered also applications such as the possibility to track events as suggested by Allan (2002), and improving information fusion techniques.

The work carried out in this Ph.D. thesis in order to investigate the relationship of TD to IR applications was complex and involved the development of resources that did not exist at the time in which the research work started. Since toponym disambiguation is seen as a specific form of Word Sense Disambiguation (WSD), the first steps were taken adapting the resources used in the evaluation of WSD. These steps involved the production of GeoSemCor, a geographic labelled version of SemCor, which consists in texts of the Brown Corpus which have been tagged using WordNet senses. Therefore, it was necessary also to create a TD method based on WordNet. GeoSemCor was used by Overell (2009) and Bensalem and Kholladi (2010) to evaluate their own TD systems. In order to compare WordNet to other resources and to compare our method to map-based existing methods, such as the one introduced by Smith and Crane (2001), which used geographical coordinates, we had to develop Geo-WordNet, a version of WordNet where all placenames have been mapped to their coordinates. Geo-WordNet has been downloaded until now by 237 universities, institutions and private companies, indicating the level of interest in this resource. This resource allows the creation of

a "bridge" between GIS and GIR research communities. The work carried out to determine whether TD is useful in GIR and QA or not was inspired by the work of Sanderson (1996) on the effects of WSD in IR. He experimented with pseudo-words, demonstrating that when the introduced ambiguity is disambiguated with an accuracy of 75%, the effectiveness is actually worse than if the collection is left undisambiguated. Similarly, in our experiments we introduced artificial levels of ambiguity on toponyms, discovering that, using WordNet, there are small differences in accuracy results, even if the number of errors is 60% of the total toponyms in the collection. However, we were able to determine that disambiguation is useful only in the case of short queries (as observed by Sanderson (1996) in the case of general WSD) and if a detailed toponym repository (e.g. Geonames instead of WordNet) is used.

We carried out also a study on an Italian local news collection, which underlined the problems that could be met in attempting to carry out TD on a collection of documents that is specific, both thematically and geographically, to a certain region. At a local scale, users are also interested in toponyms like road names, which we detected to be more ambiguous than other types of toponyms, and thus their resolution represents a more difficult task. Finally, another contribution of this Ph.D. thesis is represented by the Geooreka! prototype, a web search engine that has been developed taking into account the lessons learnt from the experiments carried out in GIR. Geooreka! can return toponyms that are particularly relevant to some event or item, carrying out a spatial mining in the web. The experiments showed that probability estimation for the co-occurrences of place and events is difficult, since place names in the web are not disambiguated. This indicates that Toponym Disambiguation plays a key role in the development of the geospatial-semantic web.

The rest of this Ph.D. thesis is structured as follows: in Chapter 2, an overview of Information Retrieval and its evaluation is given, together with an introduction on the specific IR tasks of Geographical Information Retrieval and Question Answering. Chapter 3 is dedicated to the most important resources used as toponym repositories: gazetteers and geographic ontologies, including Geo-WordNet, which represents a connection point between these two categories of repositories. 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 is discussed the ambiguity of toponyms and the methods for the resolution of such kind of ambiguity; 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. Chapter 5 is dedicated to the experiments that explored the relation between GIR and toponym disambiguation, especially to understand in which conditions toponym disambiguation may help, and how disambiguation errors affects the retrieval results. The GIR system used in these experiments, GeoWorSE, is also introduced in this chapter. In Chapter 6 the effects of TD on Question Answering have been studied using the SemQUASAR QA engine as a base system. In Chapter 7, the geographical web search engine Geooreka! is presented, and the importance of the disambiguation of toponyms in the web is discussed. Finally, in Chapter 8 are summarised the contributions of the work carried out in this thesis and some ideas for further work on the Toponym Disambiguation issue and its relation to IR are presented. Appendix A presents some data fusion experiments that we carried out in the framework of the last edition of GeoCLEF in order to combine the output of different GIR systems. Appendix B and Appendix C contain the complete topic and question sets used in the experiments detailed in Chapter 5 and Chapter 6, respectively. In Appendix D are reported some works that are based on or strictly related to the work carried out in this Ph.D. thesis.

Chapter 2

Applications for Toponym Disambiguation

Most of the applications introduced in Chapter 1 can be considered as applications related to the process of retrieving information from a text collection, or, in other words, to the research field that is commonly referred to as *Information Retrieval* (IR). A generic overview of the modules and phases that constitute the IR process has been given by Baeza-Yates and Ribeiro-Neto (1999) and is shown in Figure 2.1.

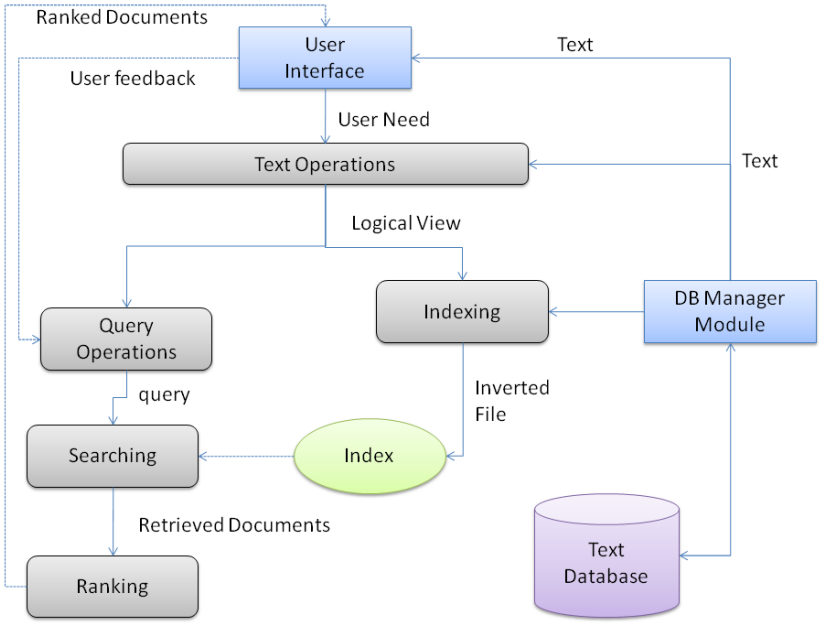


Figure 2.1: An overview of the information retrieval process.

The basic step in the IR process consists in having a document collection available (*text database*). The document are analyzed and transformed, by means of text operations. A typical transformation carried out in IR is the *stemming* process (Witten et al. (1992)), which consists in transforming inflected word forms to their root or base form. For instance, "geographical", "geographer", "geographic" would all be reduced to the same stem, "geograph". Another common text operation is the elimination of *stopwords*, with the objective of filtering out words that are usually considered not informative (e.g. personal pronouns, articles, etc.). Along with these basic operations, text can be transformed in almost every way that is considered useful by the developer of an IR system or method. For instance, documents can be divided in passages, or information that is not included in the documents can be attached to the text (for instance, if a place is contained in some region). The result of text operations constitutes the logical view of the text database, which is used to create the *index* as a result of a *indexing* process. The index is the structure that allows fast searching over large volumes of data.

At this point, it is possible to initiate the IR process by a user who specifies a *user need* which is then transformed using the same text operations used in indexing the text database. The result is a *query* that is the system representation of the user need, although the term is often used to indicate the user need themselves. The query is processed to obtain the retrieved documents, that are ranked according a likelihood or *relevance*.

In order to calculate relevance, IR systems first assign weights to the terms contained in documents. The term weight represents how important is the term in a document. Many weighting schemes have been proposed in the past, but the best known and probably one of the most used is the $tf \cdot idf$ *scheme*. The principle at the basis of this weighting scheme is that a term that is "frequent" in a given document but "rare" in the collection should be particularly informative for the document. More formally, the weight of a term t_i in a document d_j is calculated, according to the $tf \cdot idf$ weighting scheme, in the following way (Baeza-Yates and Ribeiro-Neto (1999)):

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \tag{2.1}$$

where N is the total number of documents in the database, n_i is the number of documents in which term t_i appears, and $f_{i,j}$ is the normalised frequency of term t_i in the document d_j:

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \tag{2.2}$$

where $freq_{i,j}$ is the raw frequency of t_i in d_j (i.e., the number of times the term t_i is mentioned in d_j). The $\log \frac{N}{n_i}$ part in Formula 2.1 is the *inverse document frequency* for t_i.

The term weights are used to determine the importance of a document with respect to a given query. Many models have been proposed in this sense, the most common being the *vector space* model introduced by Salton and Lesk (1968). In this model, both the query and the document are represented with a T-dimensional vector (T being the number of terms in the indexed text collection) containing their term weights: let us define $w_{i,j}$ the weight of term t_i in document d_j and $w_{i,q}$ the weight of term t_i in query q, then d_j can be represented as $\vec{d_j} = (w_{1,j}, \ldots, w_{T,j})$ and q as $\vec{q} = (w_{1,q}, \ldots, w_{T,q})$. In the vector space model, relevance is calculated as a cosine similarity measure between the document vector and the query vector:

$$\begin{aligned} sim(d_j, q) &= \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^{T} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{T} w_{i,j}} \times \sqrt{\sum_{i=1}^{T} w_{i,q}}} \end{aligned}$$

The ranked documents are presented to the user (usually as a list of *snippets*, which are composed by the title and a summary of the document) who can use them to give feedback to improve the results in the case of not being satisfied with them.

The evaluation of IR systems is carried out by comparing the result list to a list of relevant and non-relevant documents compiled by human evaluators.

2.1 Geographical Information Retrieval

Geographical Information Retrieval is a recent IR development which has been object of great attention IR researchers in the last few years. As a demonstration of this interest, GIR workshops[1] have been taking place every year since 2004 and some comparative evaluation campaigns have been organised: *GeoCLEF*[2], which took place between 2005 and 2008, and NTCIR *GeoTime*[3]. It is important to distinguish GIR from Geographic Information Systems (GIS). In fact, while in GIS users are interested in the extraction of information from a precise, structured, map-based representation, in GIR users are interested to extract information from unstructured textual information, by exploiting

[1] `http://www.geo.unizh.ch/~rsp/other.html`

[2] `http:/ir.shef.ac.uk/geoclef/`

[3] `http://research.nii.ac.jp/ntcir/ntcir-ws8/`

geographic references in queries and document collection to improve retrieval effectiveness. A definition of Geographical Information Retrieval has been given by Purves and Jones (2006), who may be considered as the "founders" of this discipline, as "the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope". It is noteworthy that despite many efforts in the last few years to organise and arrange information, the majority of the information in the world wide web is still constituted by unstructured text. Geographical information is spread over a lot of information resources, such as news and reports. Users frequently search for geographically-constrained information: Sanderson and Kohler (2004) found that almost the 20% of web searches include toponyms or other kinds of geographical terms. Sanderson and Han (2007) found also that the 37.7% of the most repeated query words are related to geography, especially names of provinces, countries and cities. Another study by Henrich and Luedecke (2007) over the logs of the former AOL search engine (now Ask.com[1]) showed that most queries are related to housing and travel (a total of about 65% of the queries suggested that the user wanted to actually get to the target location physically). Moreover, the growth of the available information is deteriorating the performance of search engines: every time the searches are becoming more demanding for the users, especially if their searches are very specific or their knowledge of the domain is poor, as noted by Johnson et al. (2006). The need for an improved generation of search engines is testified by the SPIRIT (Spatially-Aware Information Retrieval on the Internet) project (Jones et al. (2002)), which run from 2002 to 2007. This research project, funded through the EC Fifth Framework programme, that has been engaged in the design and implementation of a search engine to find documents and datasets on the web relating to places or regions referred to in a query. The project has created software tools and a prototype spatially-aware search engine has been built and has contributed to the development of the Semantic Web and to the exploitation of geographically referenced information.

In generic IR, the relevant information to be retrieved is determined only by the topic of the query (for instance, "whisky producers"); in GIR the search is based both on the topic and the geographical scope (or geographical *footprint*) for instance, "whisky producers *in Scotland*". It is therefore of vital importance to assign correctly a geographical scope to documents, and to correctly identify the reference to places in text. Purves and Jones (2006) listed some key requirements by GIR systems:

1. the extraction of geographic terms from structured and unstructured data;

[1] http://www.ask.com

2. the identification and removal of ambiguities in such extraction procedures;
3. methodologies for efficiently storing information about locations and their relationships;
4. development of search engines and algorithms to take advantage of such geographic information;
5. the combination of geographic and contextual relevance to give a meaningful combined relevance to documents;
6. techniques to allow the user to interact with and explore the results of queries to a geographically-aware IR system; and
7. methodologies for evaluating GIR systems.

The extraction of geographic terms in current GIR systems relies mostly on existing Named Entity Recognition (NER) methods. The basic objective of NER is to find names of "objects" in text, where the "object" type, or class, is usually selected from person, organization, location, quantity, date. Most NER systems also carry out the task of classifying the detected NE into one of the classes. For this reason, they may be also be referred to as NERC (Named Entity Recognition and Classification) systems. NER approaches can exploit machine learning or handcrafted rules, such as in Nadeau and Sekine (2007). Among the machine learning approaches, Maximum Entropy is one of the most used methods; see Leidner (2005) and Ferrández et al. (2005). Off-the-shelf implementations of NER methods are also available, such as GATE[1], LingPipe[2] and the Stanford NER by Finkel et al. (2005) based on Conditional Random Fields (CRF). These systems have been used for GIR in the works of Martínez et al. (2005), Buscaldi and Rosso (2007) and Buscaldi and Rosso (2009a). However, these packages are usually aimed at general usage; for instance, one could be interested not only in knowing that a name is the name of a particular location, but also in knowing the class (e.g. "city", "river", etc.) of the location. Moreover, off-the-shelf taggers have been demonstrated to be underperforming in the geographical domain by Stokes et al. (2008). Therefore, some GIR systems use custom-built NER modules, such as TALP GeoIR by Ferrés and Rodríguez (2008), which employs a Maximum Entropy approach.

The second requirement consists in the resolution of the ambiguity of toponyms, *Toponym Disambiguation* or *Toponym Resolution*, which will be discussed in detail in

[1] http://gate.ac.uk/
[2] http://alias-i.com/lingpipe/

Chapter 4. The first two requirements could be considered part of the "Text Operations" module in the generic IR process (Figure 2.1). In Figure 2.2 it is shown how these modules are connected to the IR process.

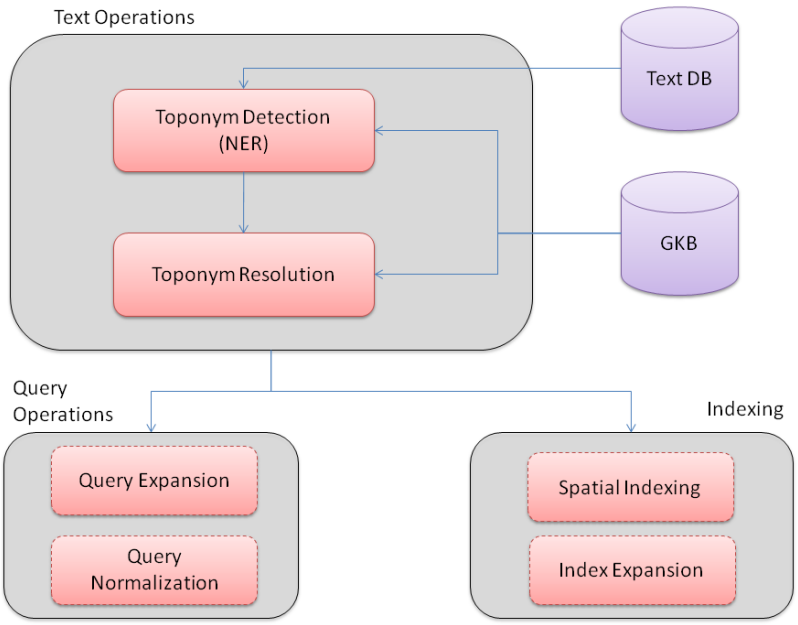


Figure 2.2: Modules usually employed by GIR systems and their position with respect to the generic IR process (see Figure 2.1). The modules with the dashed border are optional.

Storing information about locations and their relationships can be done using some database system which stores the geographic entities and their relationships. These databases are usually referred to as Geographical Knowledge Bases (GKB). Geographic entities could be cities or administrative areas, natural elements such as rivers, man-made structures. It is important not to confuse the databases used in GIS with GKBs. GIS systems store precise maps and the information connected to a geographic coordinate (for instance, how many people live in a place, how many fires have been in some area) in order to help humans in planning and take decisions. GKB are databases that determine a connection from a name to a geopolitical entity and how these entities are connected between them. Connections that are stored in GKBs are usually parent-child relations (e.g. Europe - Italy) or, sometimes, boundaries (e.g. Italy - France). Most approaches use gazetteers for this purpose. Gazetteers can be considered as dictionaries mapping names into coordinates. They will be discussed in detail in Chapter 3.

The search engines used in GIR do not differ significantly from the ones used in

standard IR. Gey et al. (2005) noted that most GeoCLEF participants based their systems on the vector space model with $tf \cdot idf$ weighting. Lucene[1], an open source engine written in Java, is used frequently, such as Terrier[2] and Lemur[3]. The combination of geographic and contextual relevance represents one of the most important challenges for GIR systems. The representation of geographic information needs with keywords and the retrieval with a general text-based retrieval system implies that a document may be geographically relevant for a given query but not thematically relevant, or that the geographic relevance is not specified adequately. Li (2007) identified the cases that could occur in the GIR scenario when users identify their geographic information needs using keywords. Here, we present a refinement of such classification. In the following, let G_d and G_q be the set of toponyms in the document and the query, respectively; let denote with $\alpha(q)$ the area covered by the toponyms included by the user in the query and $\alpha(d)$ the area that represent the geographic scope of the document. We use the $\Subset$ symbol to represent geographic inclusion (i.e., $a \Subset b$ means that area a is included in a broader region b), the $\Cap$ symbol to represent area overlap and the $\bowtie$ is used to indicate that two regions are near. Then, the following cases may occur in a GIR scenario:

a $G_q \subseteq G_d$ and $\alpha(q) = \alpha(d)$: this is the case in which both document and query contain the same geographic information;

b $G_q \cap G_d = \emptyset$ and $\alpha(q) \Cap \alpha(d) = \emptyset$: in this case the query and the document refer to different places and this is reflected in the toponyms they contain;

c $G_q \subseteq G_d$ and $\alpha(q) \Cap \alpha(d) = \emptyset$: in this case the query and the document refer to different places and this is *not* reflected by the terms they contain. This may occur if the toponyms that appear both in the document and the query are ambiguous and refer to different places;

d $G_q \cap G_d = \emptyset$ and $\alpha(q) = \alpha(d)$: in this case, the query and the document refer to the same places but the toponyms used are different; this may occur if some places can be identified by alternate names or *synonyms* (e.g. Netherlands $\Leftrightarrow$ Holland);

e $G_q \cap G_d = \emptyset$ and $\alpha(d) \Subset \alpha(q)$: in this case, the document contains toponyms that are not contained in the query but refer to places included in the relevance area specified by the query (for instance, a document containing "Detroit" maybe relevant for a query containing "Michigan");

[1] http://lucene.apache.org/
[2] http://ir.dcs.gla.ac.uk/terrier/
[3] http://www.lemurproject.org/

f $G_d \cap G_q \neq \emptyset$, with $|G_d \cap G_q| << |G_q|$ and $\alpha(d) \Subset \alpha(q)$: in this case, the query contain many toponyms of which only a small set is relevant with respect to the document: this could happen when the query contains a list of places that are all relevant (e.g. the user is interested in the same event taking place in different regions).

g $G_d \cap G_q = \emptyset$ and $\alpha(q) \Subset \alpha(d)$: then the document refers to a region that contains the places named in the query. For example, a document about the region of Liguria could be relevant to a query about "Genova", although this is not always true.

h $G_d \cap G_q = \emptyset$ and $\alpha(q) \bowtie \alpha(d)$: the document refers to a region close to the one defined by the places named in the query. This is the case of queries where users attempt to find information related to a fuzzy area around a certain region (e.g. "airports near London").

Of all the above cases, a general text-based retrieval system will only succeed in cases *a* and *b*. It may give an irrelevant document a high score in cases *c* and *f*. In the remaining cases it will fail to identify relevant documents. Case *f* could lead to *query overloading*, an undesirable effect that has been identified by Stokes et al. (2008). This effect occurs primarily when the query contains much more geographic terms than thematically-related terms, with the effect that the documents that are assigned the highest relevance are relevant to the query only under the geographic point of view.

Various techniques have been developed for GIR or adapted from IR in order to tackle this problem. Generally speaking, the combination of geographic relevance with thematic relevance such that no one surce dominates the other has been approached in two modes: the first one consist in the use of ranking fusion techniques, that is, to merge result lists obtained by two different systems into a single result list, eventually by taking advantage from the characteristics that are peculiar to each system. This technique has been implemented in the Cheshire (Larson (2008); Larson et al. (2005)) and GeoTextMESS (Buscaldi et al. (2008)) systems. The second approach used has been to combine geographic and thematic relevance into a single score, both using a combination of term weights or expanding the geographical terms used in queries and/or documents in order to catch the implicit information that is carried by such terms. The issue of whether to use ranking fusion techniques or a single score is still an open question, as reported by Mountain and MacFarlane (2007).

Query Expansion is a technique that has been applied in various works, Larson et al. (2005), Stokes et al. (2008) and Buscaldi et al. (2006c) among others. This technique consists in expanding the geographical terms in the query with geographically related

terms. The relations taken into account are those of inclusion, proximity and synonymy. In order to expand a query by inclusion, geographical terms that represent an area are expanded into terms that represent geographical entities within that area. For instance, "Europe" is expanded into a list of European countries. Expansion by proximity, used by Li et al. (2006b), is carried out by adding to the query toponyms that represent places near to the expanded terms (for instance, "near Southampton", where Southampton is the city located in the Hampshire county (UK) could be expanded into "Southampton, Eastleigh, Fareham") or toponyms that represent a broader region (in the previous example, "near Southampton" is transformed into "in Southampton and Hampshire"). Synonymy expansion is carried out by adding to a placename all terms that could be used to indicate the same place, according to some resource. For instance, "Rome" could be expanded into "Rome, eternal city, capital of Italy". Some times, "synonymy" expansion is used improperly to indicate "synecdoche" expansion: the synecdoche is a kind of metonymy in which a term denoting a part is used instead of the whole thing. An example is the use of the name of the capital to represent its country (e.g. "Washington" for "USA"), a figure of speech that is commonly used in news, especially to highlight the action of a government. The drawbacks of query expansion are the accuracy of the resources used (for instance, there is no resource indicating that "Bruxelles" is often used to indicate the "European Union") and the problem of query overloading. Expansion by proximity is also very sensible to the problem of catching the meaning of "near" as intended by the user: "near Southampton" may mean "within 30 Kms. from the centre of Southampton" but "near London" may mean a greater distance. The fuzzyness of the "near" queries is a problem that has been studied especially in GIS when natural language interfaces are used (see Robinson (2000) and Belussi et al. (2006)).

In order to contrast these effects, some researchers applied expansion on the terms contained in the index. In this way, documents are enriched with information that they did not contain originally. Ferrés et al. (2005); Li et al. (2006b) and Buscaldi et al. (2006b) add to the geographic terms in the index their containing entities, hierarchically: region, state, continent. Cardoso et al. (2007) focus on assigning a "geographic scope" or geographic signature to every document: that is, they attempt to identify the area covered by a document and add to the index the terms representing the geographic area for which the document could be relevant.

2.1.1 Geographical Diversity

Diversity Search is an IR paradigm that is somehow opposed to the classic IR vision of "Similarity Search", in which documents are ranked according to their similarity to the query. In the case of Diversity Search, users are interested in results that are relevant to the query but are different one from each other. This "diversity" could be of various kind: we may imagine a "temporal diversity", if we want to obtain documents that are relevant to an issue and show how this issue evolved in time (for instance, the query "Countries accepted into the European Union" should return documents where adhesions are grouped by year rather than a single document with a timeline of the adhesions to the Union), a "spatial" or "geographical diversity" if we are interested in obtaining relevant documents that refer to different places (in this case, the query "Countries accepted into the European Union" should return documents where adhesions are grouped by country). Diversity can be seen also as a sort of document clustering. Some clustering-based search engines like Clusty[1] and Carrot2[2] are currently available on the web, but hardly they can be considered as "diversity-based" search engines, and their results are far from being acceptable. The main reason for this failure depends on the fact that they are too general and they lack to catch diversity in any specific dimension (like the spatial or temporal dimensions).

The first mention of "Diversity Search" can be found in Carbonell and Goldstein (1998). In their paper, they proposed to use a Maximum Marginal Relevance (MMR) technique, aimed to reduce redundancy of the results obtained by an IR system, while keeping high the overall relevance of the set of results. This technique was also used with success in the document summarization task (Barzilay et al. (2002)). Recently, Diversity Search has been acquiring more importance in the work of various researchers. Agrawal et al. (2009) studied how best to diversify results in the presence of ambiguous queries and introduced some performance metrics that take into account diversity more effectively than classical IR metrics. Sanderson et al. (2009) carried out a study on diversity in the ImageCLEF 2008 task and concluded that "support for diversity is an important and currently largely overlooked aspect of information retrieval"; Paramita et al. (2009) proposed a spatial diversity algorithm that can be applied to image search; Tang and Sanderson (2010) showed that spatial diversity is greatly appreciated by users in a study carried out with the help of Amazon's Mechanical Turk[3]; finally, Clough et al. (2009) analysed query logs and found that in some ambiguity cases (person and place

[1] http://clusty.com
[2] http://search.carrot2.org
[3] https://www.mturk.com

names) users tend to reformulate queries more often.

How Toponym Disambiguation could affect Diversity Search? The potential contribution could be analyzed from two different viewpoints: in-query and in-document ambiguities. In the first case, TD may help in obtaining a better grouping of the results for those queries in which the toponym used is ambiguous. For instance, suppose that a user is looking for "Music festivals in Cambridge": the results could be grouped into two set of relevant documents, one related to music festivals in Cambridge, UK, and the other related to music festivals in Cambridge, Massachusetts. With regard to in-document ambiguities, a correct disambiguation of toponyms in the documents in the collection may help in obtaining the right results for a query where results have to be presented with spatial diversification: for instance, in the query "Universities in England", users are not interested in obtaining documents related to Cambridge, Massachusetts, which could occur if the "Cambridge" instances in the collection are incorrectly disambiguated.

2.1.2 Graphical Interfaces for GIR

An important point that is obtaining more importance recently is the development of techniques to allow users to visually explore on maps the results of queries submitted to a GIR system. For instance, results could be grouped according to place and displayed on a map, such as in the EMM NewsExplorer project[1] by Pouliquen et al. (2006), or in the SPIRIT project by Jones et al. (2002).

The number of news pages that include small maps which show the places related to some event is also increasing everyday. News from Associated Press[2] are usually found in Google News with a small map indicating the geographical scope of the news. In Fig. 2.4 we can see a mashup generated by merging data from Yahoo! Geocoding API, Google Maps and AP news (by `http://81nassau.com/apnews/`). Another example of news site providing geo-tagged news is the Italian newspaper "L'Eco di Bergamo"[3] (Fig. 2.5).

Toponym Disambiguation could result particularly useful in this task, allowing to improve the precision in geo-tagging and, consequently, the browsing experience by users. An issue with these systems is that geo-tagging errors are more evident than errors that could occur inside a GIR system.

[1] http://emm.newsexplorer.eu

[2] `http://www.ap.org/`

[3] `http://www.ecodibergamo.it/`

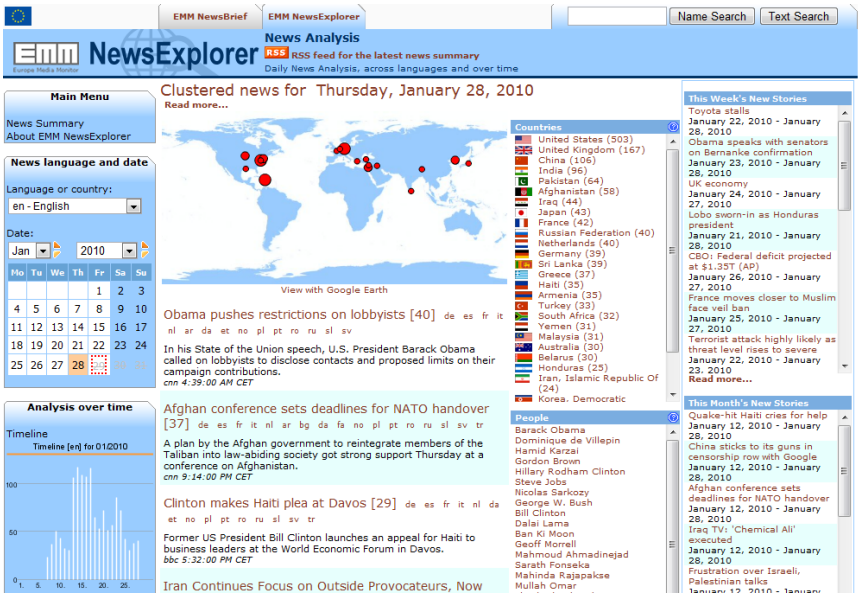


Figure 2.3: News displayed on a map in EMM NewsExplorer.

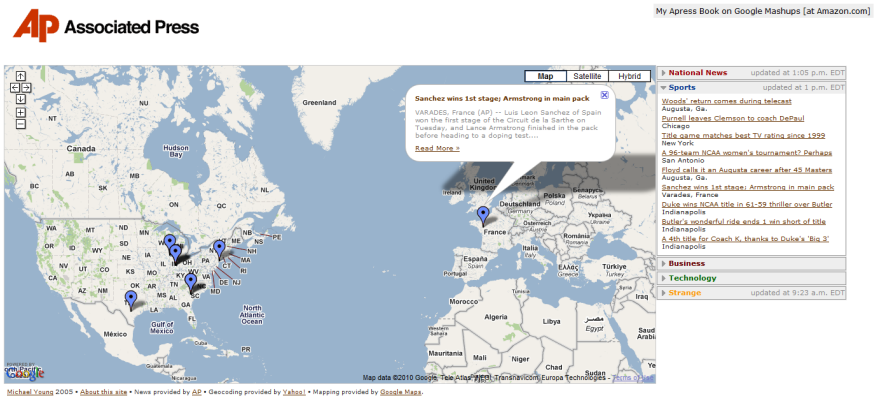


Figure 2.4: Maps of geo-tagged news of the Associated Press.

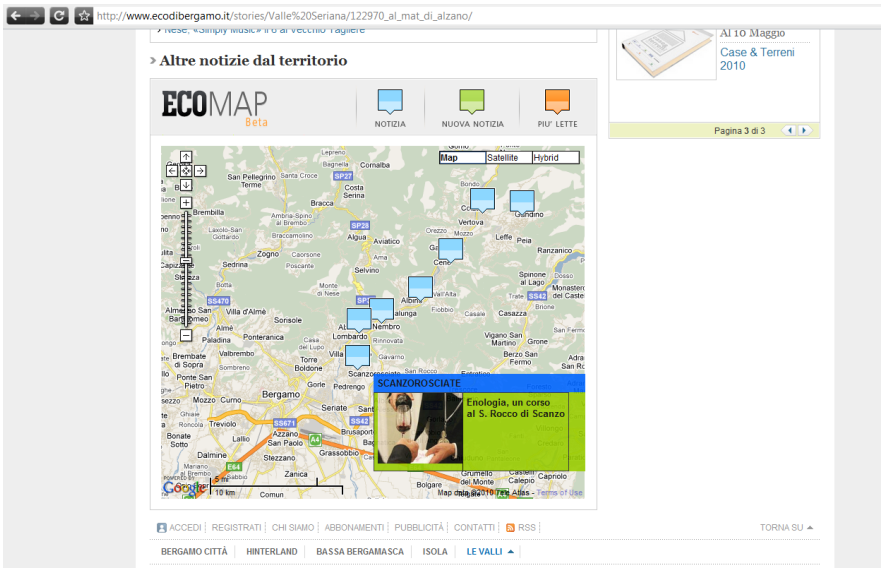


Figure 2.5: Geo-tagged news from the Italian "Eco di Bergamo".

2.1.3 Evaluation Measures

Evaluation in GIR is based on the same techniques and measures employed in IR. Many measures have been introduced in the past years; the most widely measures for the evaluation retrieval *Precision* and *Recall* NIS (2006). Let denote with R_q the set of documents in a collection that are relevant to the query q and A_s the set of documents retrieved by the system s.

The Recall $R(s,q)$ is the number of relevant documents retrieved divided by the number of relevant documents in the collection:

$$R(s,q) = \frac{|R_q \cap A_s|}{|R_q|} \tag{2.3}$$

It is used as a measure to evaluate the ability of a system to present all relevant items. The Precision ($P(s,q)$)is the fraction of relevant items retrieved over the number of items retrieved:

$$P(s,q) = \frac{|R_q \cap A_s|}{|A_s|} \tag{2.4}$$

These two measures evaluate the quality of an unordered set of retrieved documents. Ranked lists can be evaluated by plotting precision against recall. This kind of graphs is commonly referred to as *Precision-Recall* graph. Individual topic precision values are interpolated to a set of standard recall levels (0 to 1 in increments of .1):

$$P_{interp}(r) = \max_{r' \geq r} p(r') \tag{2.5}$$

Where r is the recall level. In order to better understand the relations between these measures, let us consider a set of 10 retrieved documents ($|A_s| = 10$) for a query q with $|R_q| = 12$ and let the relevance of documents be determined as in Table 2.1, with the recall and precision values calculated after examining each document.

Table 2.1: An example of retrieved documents with relevance judgements, precision and recall.

document	relevant?	Recall	Precision
d_1	y	0.08	1.00
d_2	n	0.08	0.50
d_3	n	0.08	0.33
d_4	y	0.17	0.50
d_5	y	0.25	0.60
d_6	n	0.25	0.50
d_7	y	0.33	0.57
d_8	n	0.33	0.50
d_9	y	0.42	0.55
d_{10}	n	0.42	0.50

For this example, recall and overall precision results to be $R(s,q) = 0.42$ and $P(s,q) = 0.5$ (half of the retrieved documents were relevant), respectively. The resulting Precision-Recall graph, considering the standard recall levels, is the one shown in Figure 2.6.

Another measure commonly used in the evaluation of retrieval systems is the R-Precision, defined as the precision after $|R_q|$ documents have been retrieved. One of the most used measures, especially among the TREC[1] community, is the Mean Average Precision (MAP), which provides a single-figure measure of quality across recall levels. MAP is calculated as the sum of the precision at each relevant document retrieved, divided by the total number of relevant documents in the collection. For the example in Table 2.1, MAP would be $\frac{1.00+0.50+0.60+0.57+0.55}{12} = 0.268$. MAP is considered to be an ideal measure of the quality of retrieval engines. To get an average precision of 1.0, the engine must retrieve all relevant documents (i.e., recall = 1.0) and rank them perfectly (i.e., R-Precision = 1.0).

The relevance judgments, a list of documents tagged with a label explaining whether they are relevant or not with respect to the given topic, is elaborated usually by hand

[1] `http://trec.nist.gov`

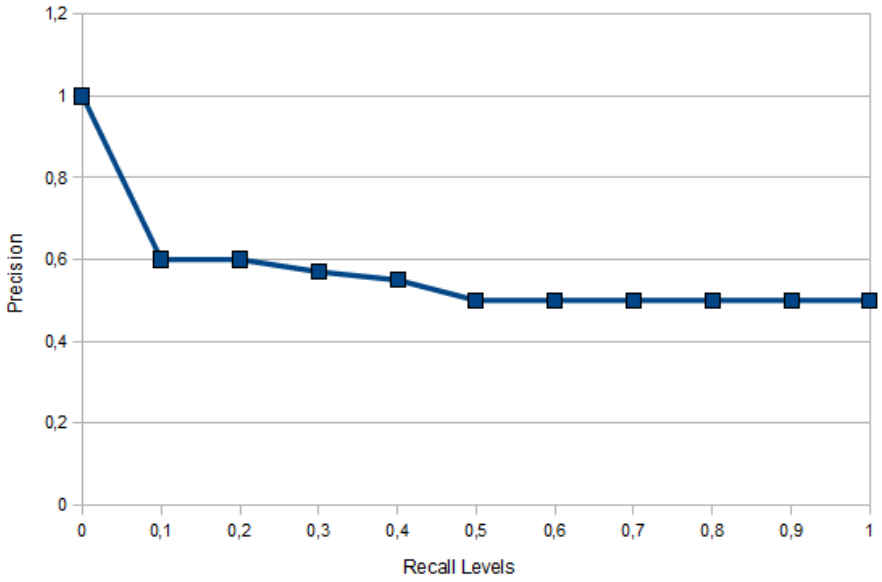


Figure 2.6: Precision-Recall Graph for the example in Table 2.1.

with human taggers. Sometimes it is not possible to prepare an exhaustive list of relevance judgments, especially in the cases where the text collection is not static (documents can be added or removed from this collection) and/or huge - like in IR on the web. In such cases, the Mean Reciprocal Rank (MRR) measure is used. MRR was defined by Voorhes in Voorhees (1999) as:

$$MRR(Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank(q)} \tag{2.6}$$

Where Q is the set of queries in the test set and $rank(q)$ is the rank at which the first relevant result is returned. Voorhees reports that the reciprocal rank has several advantages as a scoring metric and that it is closely related to the average precision measure used extensively in document retrieval.

2.1.4 GeoCLEF Track

GeoCLEF was a track dedicated to Geographical Information Retrieval that was hosted by the Cross Language Evaluation Forum (CLEF[1]) from 2005 to 2008. This track was established as an effort to evaluate comparatively systems on the basis of Geographic IR relevance, in a similar way to existing IR evaluation frameworks like TREC. The track included some cross-lingual sub-tasks together with the main English monolingual task.

[1] `http://www.clef-campaign.org`

The document collection for this task consists of 169,477 documents and is composed of stories from the British newspaper "The Glasgow Herald", year 1995 (GH95), and the American newspaper "The Los Angeles Times", year 1994 (LAT94) Gey et al. (2005). Each year, 25 "topics" were produced by the oganising groups, for a total of 100 topics covering the 4 years in which the track was held. Each topic is composed by an identifier, a title, a description and a narrative. An example of topic is presented in Figure 2.7.

```
<num>10.2452/89-GC</num>
<title>Trade fairs in Lower Saxony </title>
<desc>Documents reporting about industrial or
cultural fairs in Lower Saxony. </desc>
<narr>Relevant documents should contain
information about trade or industrial fairs which
take place in the German federal state of Lower
Saxony, i.e. name, type and place of the fair. The
capital of Lower Saxony is Hanover. Other cities
include Braunschweig, Osnabrck, Oldenburg and
Gttingen. </narr>
</top>
```

Figure 2.7: Example of topic from GeoCLEF 2008.

The title field synthesises the information need expressed by the topic, while description and narrative provides further details over the relevance criteria that should be met by the retrieved documents. Most queries in GeoCLEF present a clear separation between a thematic (or "non-geo") part and a geographic constraint. In the above example, the thematic part is "trade fairs" and the geographic constraint is "in Lower Saxony". Gey et al. (2006) presented a "tentative classification of GeoCLEF topics" based on this separation; a simpler classification is shown in Table 2.2.

Overell (2009) examined the constraints and presented a classification of the queries depending on their geographic constraint (or target location). This classification is shown in Table 2.3

Table 2.2: Classification of GeoCLEF topics based on Gey et al. (2006).

Freq.	Class
82	Non-geo subject restricted/associated to a place.
6	Geo subject with non-geographic restriction.
6	Geo subject restricted to a place.
6	Non-geo subject that is a complex function of a place.

Table 2.3: Classification of GeoCLEF topics according on their geographic constraint (Overell (2009)).

Freq.	Location	Example
9	Scotland.	Walking holidays in Scotland.
1	California.	Shark Attacks off Australia and California.
3	USA (excluding California).	Scientific research in New England Universities.
7	UK (excluding Scotland).	Roman cities in the UK and Germany.
46	Europe (excluding the UK).	Trade Unions in Europe.
16	Asia.	Solar or lunar eclipse in Southeast Asia.
7	Africa.	Diamond trade in Angola and South Africa.
1	Australasia.	Shark Attacks off Australia and California.
3	North America (excluding the USA).	Fishing in Newfoundland and Greenland.
2	South America.	Tourism in Northeast Brazil.
8	Other Specific Region.	Shipwrecks in the Atlantic Ocean.
6	Other.	Beaches with sharks.

2.2 Question Answering

A Question Answering (QA) system is an application that allows a user to question, in natural language, an unstructured document collection in order to look for the correct answer. QA is sometimes viewed as a particular form of Information Retrieval (IR) in which the amount of information retrieved is the minimal quantity of information that is required to satisfy user needs. It is clear from this definition that QA systems have to deal with more complicated problems than IR systems: first of all, what is the "minimal" quantity of information with respect to a given question? How should this information be extracted? How should it be presented to the user? These are just some of the many problems that may be encountered. The results obtained by the best QA systems are typically between 40 and 70 percent in accuracy, depending on the language and the type of exercise. Therefore, some efforts are being conducted in order to focus only on particular types of questions (restricted domain QA), including law, genomics and the geographical domain, among others.

A QA system can usually be divided into three main modules: Question Classification and Analysis, Document or Passage Retrieval, and Answer Extraction. These modules have to deal with different technical challenges, which are specific to each phase. The generic architecture of a QA system is shown in Figure 2.8.

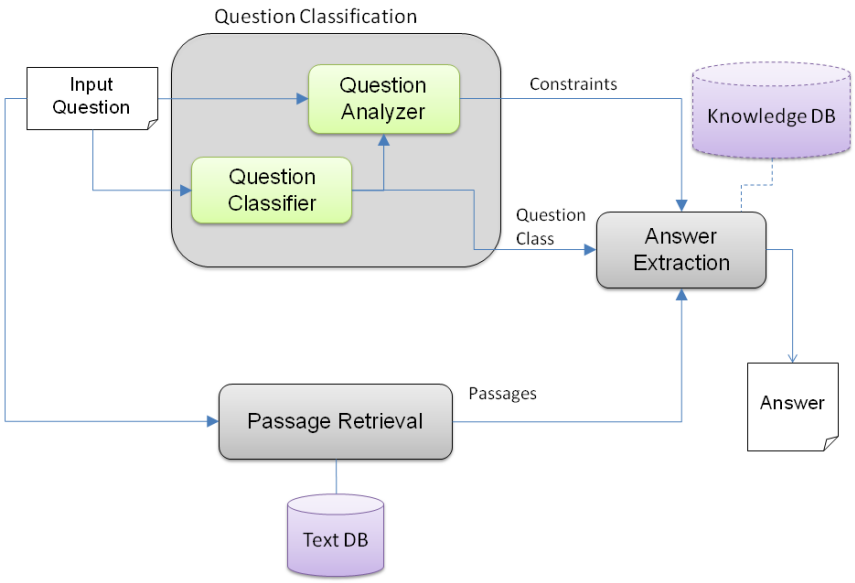


Figure 2.8: Generic architecture of a Question Answering system.

Question Classification (QC) is defined as the task of assigning a class to each question formulated to a system. Its main goals are to allow the answer extraction module to apply a different *Answer Extraction* (AE) strategy for each question type and to restrict the candidate answers. For example, extracting the answer to "What is Vicodin?", which is looking for a definition, is not the same as extracting the answer to "Who invented the radio?", which is asking for the name of a person. The class that can be assigned to a question affects greatly all the following steps of the QA process and, therefore, it is of vital importance to assign it properly. A study by Moldovan et al. (2003) reveals that more than 36% of the errors in QA are directly due to the question classification phase.

The approaches to question classification can be divided into two categories: pattern-based classifiers and supervised classifiers. In both cases, a major issue is represented by the taxonomy of classes that the question may be classified into. The design of a QC system always starts by determining what the number of classes is and how to arrange them. Hovy et al. (2000) introduced a QA typology made up of 94 question types. Most systems being presented at the TREC and CLEF-QA competitions use no more than 20 question types.

Another important task performed in the first phase is the extraction of the *focus* and the *target* of the question. The focus is the property or entity sought by the question. The target is represented by the event or object the question is about. For instance, in the question "*How many inhabitants are there in Rotterdam?*", the focus is "inhabitants" and the target "Rotterdam". Systems usually extract this information using light NLP tools, such as POS taggers and shallow parsers (chunkers).

Many questions contained in the test sets proposed in CLEF-QA exercises involve geographical knowledge (e.g. "Which is the capital of Croatia?"). The geographical information could be in the focus of the question (usually in questions asking "Where is ...") or in the target, or used as a constraint to contextualise the question. I carried out an analysis of CLEF QA questions, similarly to what Gey et al. (2006) did for GeoCLEF topics. 799 questions from the monolingual Spanish test sets from 2004 to 2007 were examined and a set of 205 questions (25.6% of the original test sets) were detected to have a geographic constraint (without discerning between target and not target) or a geographic focus, or both. The results of such classification are shown in Table 2.4. Ferrés and Rodríguez (2006) adapted an open-domain QA system to work on the geographical domain, demonstrating that geographical information could be exploited effectively in the QA task.

A *Passage Retrieval* (PR) system is an IR application that returns pieces of texts

Table 2.4: Classification of CLEF-QA questions from the monolingual Spanish test sets 2004-2007.

Freq.	Focus	Constraint	Example
45	Geo	Geo	*Which American state is San Francisco located in?*
65	Geo	non-Geo	*Which volcano did erupt in june 1991?*
95	Non-geo	Geo	*Who is the owner of the refinery in Leça da Palmeira?*

(passages) which are relevant to the user query instead of returning a ranked-list of documents. QA-oriented PR systems present some technical challenges that require an improvement of existing standard IR methods or the definition of new ones. First of all, the answer to a question may be unrelated to the terms used in the question itself, making classical term-based search methods useless. These methods usually look for documents characterised by a high frequency of query terms. For instance, in the question "What is BMW?", the only non-stopword term is "BMW", and a document that contains the term "BMW" many times probably does not contain a definition of the company. Another problem is to determine the optimal size of the passage: if it is too small, the answer may not be contained in the passage; if it is too long, it may bring in some information that is not related to the answer, requiring a more accurate Answer Extraction module. In Hovy et al. (2000); Roberts and Gaizauskas (2004) it is shown that standard IR engines often fail to find the answer in the documents (or passages) when presented with natural language questions. There are other PR approaches which are based on NLP in order to improve the performance of the QA task Ahn et al. (2004); Greenwood (2004); Liu and Croft (2002).

The *Answer Extraction* phase is responsible for extracting the answer from the passages. Every piece of information extracted during the previous phases is important in order to determine the right answer. The main problem that can be found in this phase is determining which of the possible answers is the right one, or the most informative one. For instance, an answer for "What is BMW?" can be "A car manufacturer"; however, better answers could be "A German car manufacturer" , or "A producer of luxury and sport cars based in Munich, Germany". Another problem that is similar to the previous one is related to the normalization of quantities: the answer to the question "What is the distance of the Earth from the Sun?" may be "$149,597,871$ km", "one AU", "$92,955,807$ miles" or "almost 150 million kilometers". These are descriptions of the same distance, and the Answer Extraction module should take this into account in order to exploit redundancy. Most of the Answer Extraction modules are usually based

on redundancy and on answer patterns Abney et al. (2000); Aceves et al. (2005).

2.2.1 Evaluation of QA Systems

Evaluation measures for QA are relatively simpler than the measures needed for IR, since systems are usually required to return only one answer per question. Therefore, *accuracy* is calculated as the number of "right" answers divided the number of questions answered in the test set. In QA, a "right" answer is a part of text that completely satisfies the information need of a user and represents the minimal amount of information needed to satisfy it. This requirement is necessary, otherwise it would be possible for systems to return whole documents. However, it is also difficult to determine in general what is the *minimal* amount of information that satisfies a user's information need.

CLEF-QA[1] was a task, organised within the CLEF evaluation campaign, which focused on the comparative evaluation of systems for mono- and multilingual QA. The evaluation rules of CLEF-QA were based on *justification*: systems were required to tell in which document they found the answer, and to return a snippet containing the retrieved answer. These requirements ensured that the QA system was effectively able to retrieve the answer from text and allowed the evaluators to understand whether the answer was fulfilling with the principle of minimal information needed or not. The organisers established four grades of correctness for the questions:

- R - right answer: the returned answer is correct and the document ID corresponds to a document that contains the justification for returning that answer.
- X - incorrect answer: the returned answer is missing part of the correct answer, or includes unnecessary information. For instance: Q:"What is the Atlantis?" -¿ A:"The launch of the space shuttle". The answer includes the right answer, but it also contains a sequence of words that is not needed in order to answer the question.
- U - unsupported answer: the returned answer is correct, but the source document does not contain any information allowing a human reader to deduce that answer. For instance, assuming the question is "Which company is owned by Steve Jobs?" and the document contains only "...Steve Jobs' latest creation, the Apple iPhone...", and the returned answer is "Apple", it is obvious that this passage does not state that Steve Jobs owns Apple.

[1] `http://nlp.uned.es/clef-qa/`

- W - wrong answer.

Another issue with the evaluation of QA systems is determined by the presence of *NIL* questions in test sets. A NIL question is a question for which it is not possible to return any answer. This happens when the required information is not contained in the text collection. For instance, the question "Who is Barack Obama?" posed to a system that is using the CLEF-QA 2005 collection, which used news collection from 1994 and 1995, had no answer, since "Barack Obama" is not cited in the collection (he was still an attorney in Chicago by that time). Precision over NIL questions is important, since a trustworthy system should achieve an high precision and not return NILs frequently, even when an answer exists. The Obama example is also useful to see that the answer to a same question may vary along time: "Who is the president of the United States?" has different answers if we look for in a text collection from 2010 or if we search in a text collection from 1994. The criterion used in CLEF-QA is that if the document justify the answer, then it is right.

2.2.2 Voice-activated QA

It is generally acknowledged that users prefer browsing results and checking the validity of a result by looking to contextual results rather than obtaining a short answer. Therefore, QA finds its application mostly in cases where such kind of interaction is not possible. The ideal application environment for QA systems is constituted by an environment where the user formulates the question using voice and receives the answer also vocally, via Text-To-Speech (TTS). This scenario requires the introduction of Speech Language Technologies (SLT) into QA systems.

The majority of the currently available QA systems are based on the detection of specific keywords, mostly Named Entities, in questions. For instance, a failure in the detection of the NE "Croatia" in the question "What is the capital of Croatia?" would make it impossible to find the answer. Therefore, the vocabulary of the Automated Speech Recognition (ASR) system must contain the set of NEs that can appear in the user queries to the QA system. However, the number of different NEs in a standard QA task could be huge. On the other hand, state-of-the-art speech recognition systems still need to limit the vocabulary size so that it is much smaller than the size of the vocabulary in a standard QA task. Therefore, the vocabulary of the ASR system is limited, and the presence of words in the user queries that were not in the vocabulary of the system (Out-Of-Vocabulary words) is a crucial problem in this context. Errors in keywords that are present in the queries such as *Who*, *When*, etc... can be very determinant in the question classification process. Thus, the ASR system should be

able to provide very good recognition rates on this set of words. Another problem that affects these systems is the incorrect pronunciation of NEs (such as names of persons or places) when the NE is in a language that is different from the user's. A mechanism that considers alternative pronunciations of the same word or acronym must be implemented.

In Harabagiu et al. (2002), the authors show the results of an experiment combining a QA system with an ASR system. The baseline performance of the QA system from text input was 76%, whereas when the same QA system worked with the output of the speech recogniser (which operated at s 30% WER), it was only 7%.

2.2.2.1 QAST: Question Answering on Speech Transcripts

QAST is a track that has been part of the CLEF evaluation campaign from 2007 to 2009. It is dedicated to the evaluation of QA systems that search answers in text collections composed of speech transcripts, which are particularly subject to errors. I was part of the organisation on the UPV side for the 2009 edition of QAST, in conjunction with the UPC (Universidad Politécnica de Catalunya) and LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur). In 2009, QAST aims were extended in order to provide a framework in which QA systems can be evaluated in a real scenario, where questions can be formulated as "spontaneous" oral questions. There were five main objectives to this evaluation Turmo et al. (2009):

- motivating and driving the design of novel and robust QA architectures for speech transcripts;
- measuring the loss due to the inaccuracies in state-of-the-art ASR technology;
- measuring this loss at different ASR performance levels given by the ASR word error rate;
- measuring the loss when dealing with spontaneous oral questions;
- motivating the development of monolingual QA systems for languages other than English.

Spontaneous questions may contain noise, hesitations and pronunciation errors that usually are absent in the written questions provided by other QA exercises. For instance, the manually transcribed spontaneous oral question *When did the bombing of Fallujah eee took take place?* corresponds to the written question *When did the bombing*

of Fallujah take place?. These errors make QAST probably the most realistic task for the evaluation of QA systems among the ones present in CLEF.

The text collection is constituted by the English and Spanish versions of the *TC-STAR05 EPPS English corpus*[1], containing 3 hours of recordings corresponding to 6 sessions of the European Parliament. Due to the characteristics of the document collection, questions were related especially to international issues, highlighting the geographical aspects of the questions. As part of the organisation of the task, I was responsible for the collection of questions for the Spanish test set, resulting in a set of 296 spontaneous questions. Among these questions, 79 (26.7%) required a geographic answer or were geographically constrained. In Table 2.5 a classification like the one presented in Table 2.4 is shown.

Table 2.5: Classification of QAST 2009 spontaneous questions from the monolingual Spanish test set.

Freq.	Focus	Constraint	Example
36	Geo	Geo	*en qué continente está la región de los grandes lagos.*
15	Geo	non-Geo	*dime un país del cual (%hesit) sus habitantes huyan del hambre.*
28	Non-geo	Geo	*cuántos habitantes hay en la Unión Europea.*

The QAST evaluation showed no significant difference between the use of written and spoken questions, indicating that the noise introduced in spontaneous questions does not represent a major issue for Voice-QA systems.

2.2.3 Geographical QA

The fact that many of the questions in open-domain QA tasks (25.6% and 26.7% in Spanish for CLEF-QA and QAST, respectively) have a focus related to geography or involve geographic knowledge is probably one of the most important factors that boosted the development of some tasks focused on geography. *GikiP*[2] was proposed in 2008 in the GeoCLEF framework as an exercise to "find Wikipedia entries / articles that answer a particular information need which requires geographical reasoning of some sort." (Santos and Cardoso (2008)). GikiP is some kind of an hybrid between an IR and a QA exercise, since the answer is constituted by a Wikipedia entry, like in IR, while the input query is a question like in QA. Example of GikiP questions: *Which waterfalls are used in the film "The Last of the Mohicans"?*, *Which plays of Shakespeare*

[1] `http://www.tc-star.org/`

[2] `http://www.linguateca.pt/GikiP/`

take place in an Italian setting?.

GikiCLEF[1] was a follow-up of the GikiP pilot task that took place in CLEF 2009. The test set was composed by 50 questions in 9 different languages, focusing on cross-lingual issues. The difficulty of questions was recognised to be higher than in GikiP or GeoCLEF (Santos et al. (2010)), with some questions involving complex geographical reasoning, like in *Find coastal states with Petrobras refineries* and *Austrian ski resorts with a total ski trail length of at least 100 km.*

In NTCIR[2], an evaluation workshop similar to CLEF focused on Japanese and Asian languages, a GIR-related task was proposed in 2010 under the name *GeoTime*[3]. This task is focused on questions that requires two answers, one about the place and another one about the time in which some event occurred. Examples of questions of the GeoTime task are: *When and where did Hurricane Katrina make landfall in the United States?*, *When and where did Chechen rebels take Russians hostage in a theatre?* and *When was the decision made on siting the ITER and where is it to be built?*. The document collection is composed of news stories extracted from the New York Times $2002-2005$ for the English language and news stories of the same time period extracted from the "Meinichi" newspaper for the Japanese language.

2.3 Location-Based Services

In the last years, mobile devices able to track their position by means of GPS have become increasingly common. These devices are also able to navigate in the web, making Location-Based Services (LBS) a reality. These services are information and/or entertainment services which can use the geographical position of the mobile device in order to provide the user with information that depends on its location. For instance, LBS can be used to find the nearest business or service (a restaurant, a pharmacy, or a banking cash machine), the whereabouts of a friend (such as Google latitude[4]), or even to track vehicles.

In most cases, the information to be presented to the user is static and geocoded (for instance, in GPS navigators business and services are stored with their position). Baldauf and Simon (2010) developed a service that, given a users whereabout, performs a location-based search for georeferenced Wikipedia articles using the coordinates of the user's device in order to show nearby places of interests. Most applications now

[1] `http://www.linguateca.pt/GikiCLEF/`
[2] `http://research.nii.ac.jp/ntcir/`
[3] `http://metadata.berkeley.edu/NTCIR-GeoTime`
[4] `http://www.google.com/mobile/latitude/`

allow users to upload contents, such as pictures or blog entries, and geo-tag them. Toponym Disambiguation could result useful when the content is not tagged and it is not practical to carry out the geo tagging by hand.

Chapter 3

Geographical Resources and Corpora

The concept of place is both a human and geographic concept. The cognition of place is vague: a crisp delineation of a place is not always possible. However, exactly in the same way as dictionaries exist for common names, representing an agreement that allows people to refer to the same concept using the same word, there are dictionaries that are dedicated to place names. These dictionaries are commonly referred to as *gazetteers*, and their basic function is to map toponyms to coordinates. They may also contain additional information regarding the place represented by a toponym, such as its area, height, or its population, if it is a populated place. Gazetteers can be seen as a "plain" list of pairs *name* → *geographical coordinates*, which is enough to carry out certain tasks (for instance, calculating distances between two places given their names); however, they lack the information about how places are organised or connected (i.e., the *topology*). GIS systems usually need this kind of topological information in order to be able to satisfy complex geographic information needs (such as "which river crosses Paris?" or "which motorway connects Rome to Milan?"). This information is usually stored in databases with specific geometric operators enabled. Some structured resources contain limited topological information, specifically the containment relationship, so we can say that Genova is a town inside Liguria that is a region of Italy. Basic gazetteers usually include the information about to which administrative entity a place belongs to, but other relationships, like "X borders Y", are usually not included.

The resources may be classified according to the following characteristics: *scope*, *coverage* and *detail*. The scope of a geographic resource indicates whether a resource is limited to a region or a country (GNIS for instance is limited to the United States),

or it is a broad resource covering all the parts of the world. Coverage is determined by the number of placenames listed in the resource. Obviously, scope determines also the coverage of the resource. Detail is related to how fine-grained is the resource with respect to the area covered. For instance, a local resource can be very detailed. On the other hand, a broad resource with low detail can cover only the most important places. This kind of resources may ease the toponym disambiguation task by providing a useful bias, filtering out placenames that are very rare, which may constitute 'noise'. The behaviour of people of seeing the world at a level of detail that decreases with distance is quite common. For instance, an "earthquake in L'Aquila" announced in Italian news becomes the "Italian earthquake" when the same event is reported by foreign news. This behaviour has been named the "Steinberg hypothesis" by Overell (2009), citing the famous cartoon "View of the world from 9th Avenue" by Saul Steinberg[1], which depicts the world as seen by self-absorbed New Yorkers.

In Table 3.1 we show the characteristics of the most used toponym resources with global scope, which are described in detail in the following sections.

Table 3.1: Comparative table of the most used toponym resources with global scope. *: coordinates added by means of Geo-WordNet. Coverage: number of listed places.

Type	Name	Coordinates	Coverage
Gazetteer	Geonames	y	$\sim 7,000,000$
	Wikipedia-World	y	$264,288$
Ontologies	Getty TGN	y	$1,115,000$
	Yahoo! GeoPlanet	n	$\sim 6,000,000$
	WordNet	y*	$2,188$

Resources with a less general scope are usually produced by national agencies for use in topographic maps. Geonames itself is derived from the combination of data provided by the National Geospatial Intelligence Agency (GNS[2] - GEOnet Names Server) and the United States Geological Service in cooperation with the U.S. Board of Geographic Names (GNIS[3] - Geographic Names Information System). The first resource (GNS) includes names from every part of the world except the United States, which are covered by the GNIS, which contains information about physical and cultural geographic features. Similar resources are produced by the agencies of the United Kingdom (Ord-

[1] http://www.saulsteinbergfoundation.org/gallery_24_viewofworld.html
[2] http://gnswww.nga.mil/geonames/GNS/
[3] http://geonames.usgs.gov/

nance Survey[1]), France (Institut Géographique National[2])), Spain (Instituto Geográfico Nacional[3]) and Italy (Istituto Geografico Militare[4]), among others. The resources produced by national agencies are usually very detailed but they present two drawbacks: they are usually not free, and sometimes they use geodetic systems that are different from the most commonly used (the World Geodetic System, or WGS). For instance, Ordnance Survey maps of Great Britain do not use latitude and longitude to indicate position but a special grid (British national grid reference system).

3.1 Gazetteers

Gazetteers are the main sources of geographical coordinates. A gazetteer is a dictionary where each toponym has associated its latitude and longitude. Moreover, they may include further information about the places indicated by toponyms, such as their feature class (e.g. city, mountain, lake, etc.).

One of the oldest gazetteer is the *Geography* of Ptolemy[5]. In this work, Ptolemy assigned to every toponym a pair of coordinates calculated using Erathostenes' coordinate system. In Table 3.2 we can see an excerpt of this gazetteer referring to Southeastern England:

Table 3.2: An excerpt of Ptolemy's gazetteer with modern corresponding toponyms and coordinates.

toponym	modern toponym	lon, lat (Erathostenes)	lat, lon (WGS84)
Londinium	London	$20 * 00$, $54°00$	$51°30'29"N, 0°7'29"W$
Daruernum	Canterbury	$21 * 00$, $54°00$	$51°16'30"N, 1°5'13.2"E$
Rutupie	Richborough	$21 * 45$, $54°00$	$51°17'47.4"N, 1°19'9.12"E$

The Geographic Coordinate Systems (GCS) used in ancient times were not particularly precise, due to the limits of the measurement methods. As it can be noted in Table 3.2, according to Ptolemy, all places laid at the same latitude, but now we know that this is not exact. A GCS is a coordinate system that allows to specify every location on Earth in three coordinates: latitude, longitude and height. For our purpose we will

[1] `http://www.ordnancesurvey.co.uk/oswebsite/`

[2] `http://www.ign.fr`

[3] `http://www.ign.es`

[4] `http://www.igmi.org`

[5] `http://penelope.uchicago.edu/Thayer/E/Gazetteer/Periods/Roman/_Texts/Ptolemy/home.html`

avoid talking about the third coordinate, focusing on 2-dimensional maps. *Latitude* is the angle from a point on the Earth's surface to the equatorial plane, measured from the center of the sphere. *Longitude* is the angle east or west of a reference meridian to another meridian that passes through an arbitrary point. In Ptolemy's Geography, the reference meridian passed through El Hierro island, in the Atlantic ocean, the (then) western-most position of the known world; in the WGS84 standard, the reference meridian passes about 100 meters west of the Greenwich meridian, which is used in the British national grid reference system. In order to be able to compute distances between places, it is necessary to approximate the shape of the Earth to a sphere or, more precisely, to an ellipsoid: the differences in standards are due to the choices made for the ellipsoid that approximates Earth's surface. Given a reference standard, is possible to calculate a distance between two points using spherical distance: given two points p and q with coordinates (ϕ_p, λ_p) and (ϕ_q, λ_q) respectively, with ϕ being the latitude and λ the longitude, then the spherical distance $r\Delta\hat{\sigma}$ between p and q can be calculated as:

$$r\Delta\hat{\sigma} = r \arccos\left(\sin\phi_p \sin\phi_q + \cos\phi_p \cos\phi_q \cos\Delta\lambda\right) \tag{3.1}$$

where r is the radius of the Earth $(6,371.01km)$ and $\Delta\lambda$ is the difference $\lambda_q - \lambda_p$.

As introduced before, place is not only a geographic concept, but also human: in fact, as it can be also observed in Table 3.2, most of the toponyms listed by Ptolemy were inhabited places. Modern gazetteers are also biased towards human usage: as it can be seen in Figure 3.2, most of Geonames locations are represented by buildings and populated places.

3.1.1 Geonames

Geonames[1] is an open project for the creation of a world geographic database. It contains more than 8 million geographical names and consists of 7 million unique features. All features are categorised into one out of nine feature classes (shown in Figure 3.2) and further subcategorised into one out of 645 feature codes. The most important data sources used by Geonames are the GEOnet Names Server (GNS) and the Geographic Names Information System (GNIS). The coverage of Geonames can be observed in Figure 3.1. The bright parts of the map show high density areas sporting a lot of features per km^2 and the dark parts show regions with no or only few GeoNames features.

To every toponym are associated the following information: alternate names, latitude, longitude, feature class, feature code, country, country code, four administrative entities that contain the toponym at different levels, population, elevation and time

[1] `http://www.geonames.org`

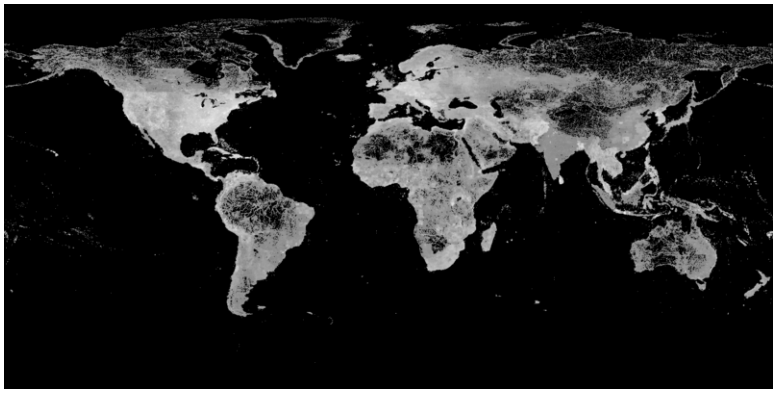


Figure 3.1: Feature Density Map with the Geonames data set.

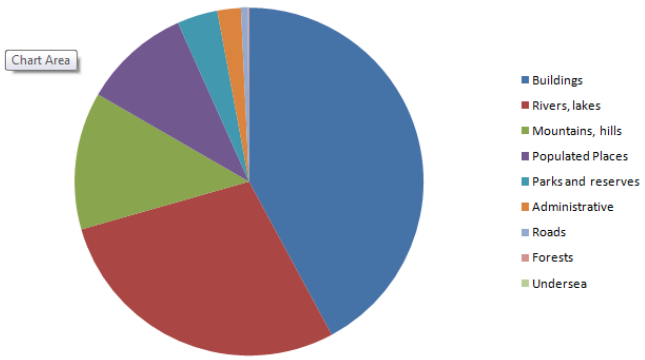


Figure 3.2: Composition of Geonames gazetteer, grouped by feature class.

zone. The database can also be queried online, showing the results on a map or as a list. The results of a query for the name "Genova" are shown in Figure 3.3. The Geonames database does not include zip codes, which can be downloaded separately.

	Name	Country	Feature class	Latitude	Longitude
1	Genoa Cenova,Dzenova,Dženova,Genes,Genoa,Genova,Genovo,Genua,Genuja,Genès,Genúa,Gènova,Génova,Gênes,Janov,...	Italy, Liguria	seat of a first-order administrative division population 601,951	N 44° 24' 22"	E 8° 56' 1"
2	Genoa Genes,Genoa,Genova,Genua,Genès,Génova,Gênes	Italy, Liguria	second-order administrative division	N 44° 30' 0"	E 9° 4' 0"
3	Genova Capital Genova,Capital Génova,Colon,Colón	Colombia, Nariño	populated place population 1,348	N 1° 38' 47"	W 77° 1' 27"
4	Génova	Colombia, Quindío	second-order administrative division	N 4° 15' 0"	W 75° 40' 0"
5	Golfo di Genova Genes,Genova,Genua,Genès,Gulf of Genoa,Génova	Italy,	gulf	N 44° 10' 0"	E 8° 55' 0"
6	Municipio de Génova	Guatemala, Quetzaltenango	second-order administrative division	N 14° 37' 0"	W 91° 50' 0"

Figure 3.3: Geonames entries for the name "Genova".

3.1.2 Wikipedia-World

The Wikipedia-World (WW) project[1] is a project aimed to label Wikipedia articles with geographic coordinates. The coordinates and the article data are stored in a SQL database that is available for download. The coverage of this resource is smaller than the one offered by Geonames, as it can be observed in Figure 3.4. By February 2010, the number of georeferenced Wikipedia pages is of 815,086. These data are included in the Geonames database. However, the advantage of using Wikipedia is that the entries included in Wikipedia represent the most discussed places on the Earth, constituting a good gazetteer for general usage.

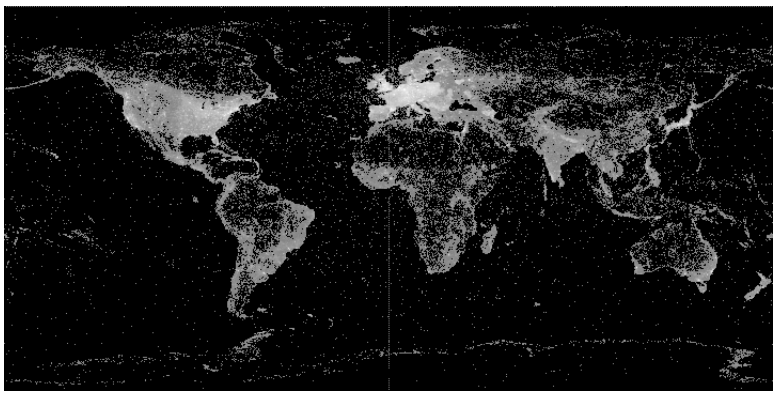


Figure 3.4: Place coverage provided by the Wikipedia World database (toponyms from the 22 covered languages).

[1] `http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_\Georeferenzierung/Wikipedia-World/en`

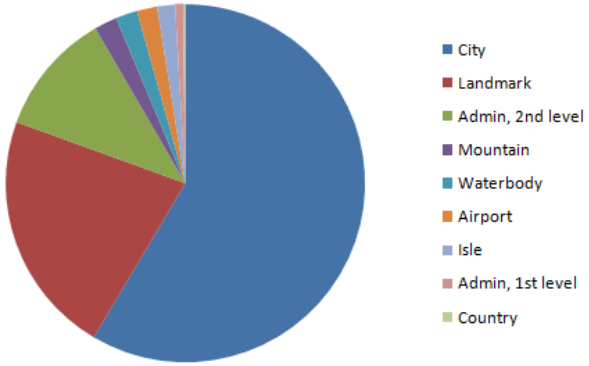


Figure 3.5: Composition of Wikipedia-World gazetteer, grouped by feature class.

Each entry of the Wikipedia-World gazetteer contains the toponym, alternate names for the toponym in 22 languages, latitude, longitude, population, height, containing country, containing region, and one of the classes shown in Figure 3.5. As it can be seen in this figure, populated places and human-related features, such as buildings and administrative names, constitute the great majority of the placenames included in this resource.

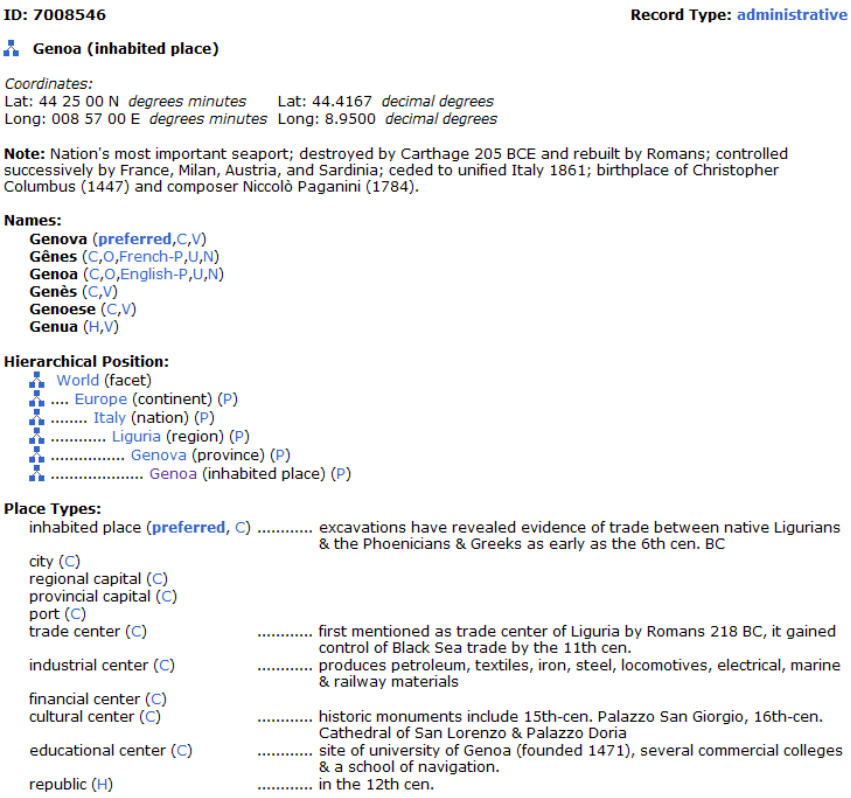
3.2 Ontologies

Geographic ontologies allow not only to know the coordinates and the physical characteristics of a place associated to a toponym, but also the relationships between toponyms. Usually these relationships are represented by containment relationships, indicating that a place is contained into another. However, some ontologies contain also information about neighbouring places.

3.2.1 Getty Thesaurus

The Getty Thesaurus of Geographic Names (TGN)[1] is a commercial structured vocabulary containing around 1,115,000 names. Names and synonyms are structured hierarchically. There are around 895,000 unique places in the TGN. In the database, each place record (also called a *subject*) is identified by a unique numeric ID or reference. In Figure 3.6 it is shown the result of the query "Genova" on the TGN online browser.

[1] `http://www.getty.edu/research/conductingresearch/vocabularies/tgn/`



ID: 7008546 **Record Type: administrative**

Genoa (inhabited place)

Coordinates:
Lat: 44 25 00 N *degrees minutes* Lat: 44.4167 *decimal degrees*
Long: 008 57 00 E *degrees minutes* Long: 8.9500 *decimal degrees*

Note: Nation's most important seaport; destroyed by Carthage 205 BCE and rebuilt by Romans; controlled successively by France, Milan, Austria, and Sardinia; ceded to unified Italy 1861; birthplace of Christopher Columbus (1447) and composer Niccolò Paganini (1784).

Names:
Genova (**preferred**,C,V)
Gênes (C,O,French-P,U,N)
Genoa (C,O,English-P,U,N)
Genès (C,V)
Genoese (C,V)
Genua (H,V)

Hierarchical Position:
World (facet)
.... Europe (continent) (P)
..... Italy (nation) (P)
..... Liguria (region) (P)
..... Genova (province) (P)
..... Genoa (inhabited place) (P)

Place Types:
inhabited place (**preferred**, C) ..... excavations have revealed evidence of trade between native Ligurians & the Phoenicians & Greeks as early as the 6th cen. BC
city (C)
regional capital (C)
provincial capital (C)
port (C)
trade center (C) ..... first mentioned as trade center of Liguria by Romans 218 BC, it gained control of Black Sea trade by the 11th cen.
industrial center (C) ..... produces petroleum, textiles, iron, steel, locomotives, electrical, marine & railway materials
financial center (C)
cultural center (C) ..... historic monuments include 15th-cen. Palazzo San Giorgio, 16th-cen. Cathedral of San Lorenzo & Palazzo Doria
educational center (C) ..... site of university of Genoa (founded 1471), several commercial colleges & a school of navigation.
republic (H) ..... in the 12th cen.

Figure 3.6: Results of the Getty Thesarurus of Geographic Names for the query "Genova".

3.2.2 Yahoo! GeoPlanet

Yahoo! GeoPlanet[1] is a resource developed with the aim of giving to developers the opportunity to geographically enable their applications, by including unique geographic identifiers in their applications and to use Yahoo! web services to unambiguously geotag data across the web. The data can be freely downloaded and provide the following information:

- WOEID or Where-On-Earth IDentifier: a number that uniquely identifies a place;
- Hierarchical containment of all places up to the "Earth" level;
- Zip codes are included as place names;
- Adjacencies: places neighbouring each WOEID;
- Aliases: synonyms for each WOEID.

As it can be seen, GeoPlanet focuses on structure rather than on the information about each toponym. In fact, the major drawback of GeoPlanet is that it does not list the coordinates associated at each WOEID. However, it is possible to connect to Yahoo! web services to retrieve them. In Figure 3.7 it is visible the composition of Yahoo! GeoPlanet, according the feature class used. It is notable that the great majority of the data is constituted by zip codes ($3,397,836$ zip codes), which, although not being usually considered toponyms, play an important role in the task of geo tagging data in the web. The number of towns listed in GeoPlanet is currently $863,749$, a figure close to the number of places in Wikipedia-World. Most of the data contained in GeoPlanet, however, is represented by the table of adjacencies, containing $8,521,075$ relations. From these data it is clear the vocation of GeoPlanet to be a resource for location-based and geographically-enabled web services.

3.2.3 WordNet

WordNet is a lexical database of English Miller (1995). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (*synsets*), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations, resulting in a network of meaningfully related words and concepts. Among the relations that connects synsets, the most important, under the geographic aspect, are the *hypernymy* (or *is-a* relationship), the *holonymy* (or *part-of* relationship) and the

[1] `http://developer.yahoo.com/geo/geoplanet/`

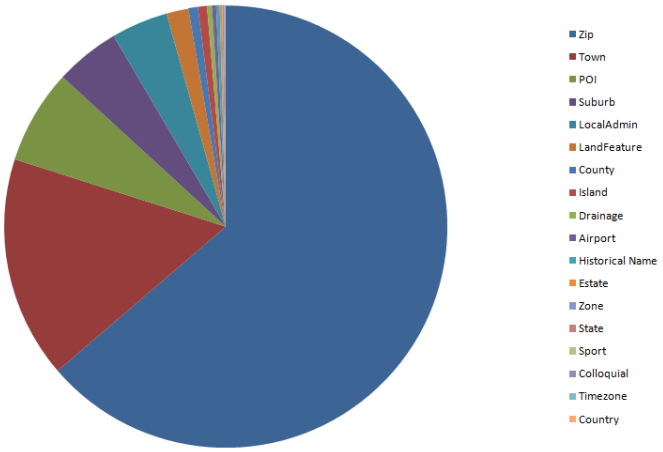


Figure 3.7: Composition of Yahoo! GeoPlanet, grouped by feature class.

instance_of relationship. For place names, instance_of allows to find the *class* of a given name (this relation was introduced in the 3.0 version of WordNet, in previous versions hypernymy was used in the same way). For example, "Armenia" is an instance of the concept "country", and "Mount St. Helens" is an instance of the concept "volcano". Holonymy can be used to find a geographical entity that contains a given place, such as "Washington (U.S. state)" that is holonym of "Mount St. Helens". By means of the holonym relationship it is possible to define hierarchies in the same way as in GeoPlanet or the TGN thesaurus. The inverse relationship of holonymy is *meronymy*: a place is meronym of another if it is included in this one. Therefore, "Mount St. Helens" is meronym of "Washington (U.S. state)". Synonymy in WordNet is coded by synsets: each synset comprises a set of lemmas that are synonyms and thus represent the same concept, or the same place, if the synset is referring to a location. For instance, "Paris", France appears in WordNet as "`Paris, City of Light, French capital, capital of France`". This information is usually missing from typical gazetteers, since "French capital" is considered a synonym for "Paris" (it is not an alternate name) which makes WordNet particularly useful for NLP tasks.

Unfortunately, WordNet presents some problems as a geographical information resource. First of all, the quantity of geographical information is quite small, especially if compared with any of the resources described in the previous sections. The number of geographical entities stored in WordNet can be calculated by means the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. The second problem is that WordNet is not *georef-*

erenced, that is, the toponyms are not assigned their actual coordinates on earth. Georeferencing WordNet can be useful for many reasons: first of all, it is possible to establish a semantics for synsets that is not vinculated only to a written description (the synset *gloss*, e.g.: "Marrakech, a city in western Morocco; tourist center"). In second place, it can be useful in order to enrich WordNet with information extracted from gazetteers, or to enrich gazetteers with information extracted from WordNet; finally, it can be used to evaluate toponym disambiguation methods that are based on geographical coordinates using resources that are usually employed for the evaluation of WSD methods like SemCor[1], a corpus of English text labelled with WordNet senses. The introduction of Geo-WordNet by Buscaldi and Rosso (2008b) allowed to overcome the issues related to the lack of georeferences in WordNet. This extension allowed to map the locations included in WordNet, as in Figure 3.8, from which it is notable the small coverage of WordNet compared to Geonames and Wikipedia-World. The development of Geo-WordNet is detailed in Section 3.3.

Figure 3.8: Feature Density Map with WordNet.

3.3 Geo-WordNet

In order to compensate the lack of geographical coordinates in WordNet, we developed Geo-WordNet as an extension of WordNet 2.0. Geo-WordNet should not be confused with another almost homonymous project, *GeoWordNet* (without the −) by Giunchiglia et al. (2010), which adds more geographical synsets to WordNet instead than adding information on the already included ones. This resource is not yet available at the time of writing. Geo-WordNet was obtained by mapping the locations included

[1] `http://www.cs.unt.edu/$\sim$rada/downloads.html#semcor`

in WordNet to locations in the Wikipedia-World gazetteer. This gazetteer was preferred with respect to the other resources because of its coverage. In Figure 3.9 we can see a comparison between the coverage of toponyms by the resources previously presented. WordNet is the resource covering the least amount of toponyms, followed by TGN and Wikipedia-World which are similar in size although they do not cover exactly the same toponyms. Geonames is the largest resource, although GeoPlanet contains zip codes that are not included in Geonames (however, they are available separately).

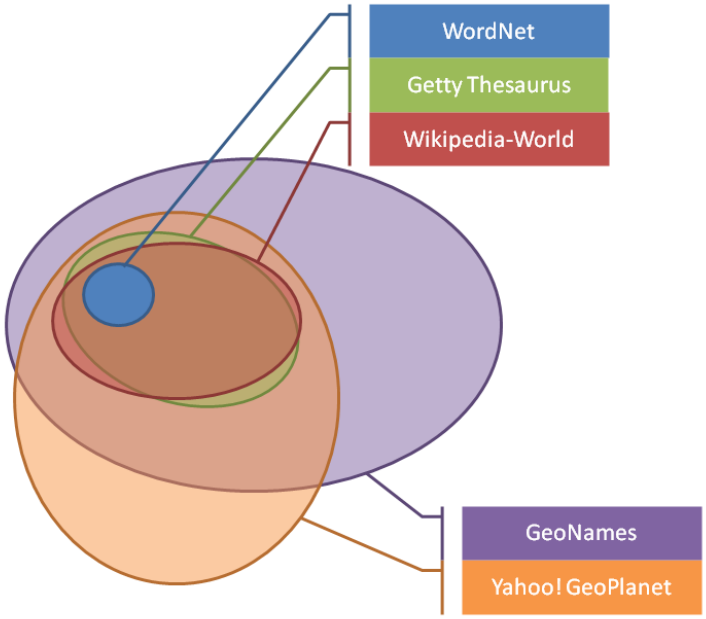


Figure 3.9: Comparison of toponym coverage by different gazetteers.

Therefore, the selection of Wikipedia-World allowed to reduce the number of possible referents for each WordNet locations with respect to a broader gazetteer such as Geonames, simplifying the task. For instance, "Cambridge" has only 2 referents in WordNet, 68 referents in Geonames and 26 in Wikipedia-World. TGN was not taken into account because it is not freely available.

The heuristic developed to assign an entry in Wikipedia-World to a geographic entry in WordNet is pretty simple and is based on the following criteria:

- Match between a synset wordform and a database entry;

- Match between the holonym of a geographical synset and the containing entity of the database entry;
- Match between a second level holonym and a second level containing entity in the database;
- Match between holonyms and containing entities at different levels (0.5 weight); this corresponds to a case in which WordNet or the WW lacks the information about the first level containing entity.
- Match between the hypernym and the class of the entry in the database (0.5 weight);
- A class of the database entry is found in the gloss (i.e., the description) of the synset (0.1 weight).

The reduced weights were introduced for cases where an exact match could lead to a wrong assignment. This is true especially for gloss comparison, since WordNet glosses usually include example sentences that are not related with the definition of the synset, but instead provide a "use case" example.

The mapping algorithm is the following one:

1. Pick a synset s in WordNet and extract all of its wordforms $w_1, \ldots, w_n$ (i.e., the name and its synonyms)
2. Check whether a wordform w_i is in the WW database
3. If w_i appears in WW: find the holonym h_s of the synset s. Else: goto 1.
4. If $h_s = \{\}$: goto 1. Else: find the holonym h_{hs} of h_s
5. Find the hypernym H_s of the synset s.
6. $L = \{l_1, \ldots, l_m\}$ is the set of locations in WW that correspond to the synset s
7. A weight is assigned to each l_i depending on the weighting function f
8. The coordinates related to $\max_{l_i \in L} f(l_i)$ are assigned to the synset s
9. Repeat until the last synset in WordNet

A final step was carried out manually and consisted in reviewing the labelled synsets, removing those which were mistakenly identified as locations.

The weighting function is defined as:

$$f(l) = m(w_i, l) + m(h_s, c(l)) + m(h(h_s), c(c(l))) + \\ +0.5 \cdot m(h_s, c(c(l))) + 0.5 \cdot m(h(h_s), c(l)) + \\ +0.1 \cdot g(D(l)) + 0.5 \cdot m(H_s, D(l))$$

where $m : \Sigma^* \times \Sigma^* \rightarrow \{1, 0\}$ is a function returning 1 if the string x matches l from the beginning to the end or from the beginning to a comma, and 0 in the other cases. $c(x)$ returns the containing entity of x, for instance it can be $c(\text{“}Abilene\text{”}) = \text{“}Texas\text{”}$ and $c(\text{“}Texas\text{”}) = \text{“}US\text{”}$. In a similar way, $h(x)$ retrieves the holonym of (x) in WordNet. $D(x)$ returns the class of location x in the database (e.g. a mountain, a city, an island, etc.). $g : \Sigma^* \rightarrow \{1, 0\}$ returns 1 if the string is contained in the gloss of synset s. Country names obtain an extra +1 if they match with the database entry name and the country code in the database is the same as the country name.

For instance, consider the following synset from WordNet: *(n) Abilene (a city in central Texas)*; in Figure 3.10 we can see its first level and second level holonyms (“Texas” and “USA”, respectively) and its direct hypernym (“city”).

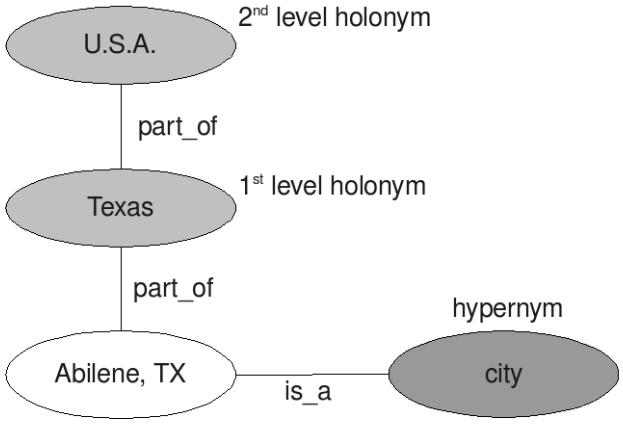


Figure 3.10: Part of WordNet hierarchy connected to the “Abilene” synset.

A search in the WW database with the query `SELECT Titel_en, lat, lon, country, subregion, style FROM pub_CSV_test3 WHERE Titel_en like ‘‘Abilene%"` returns the results in Figure 3.11. The fields have the following meanings: *Titel_en* is the English name of the place, *lat* is the latitude, *lon* the longitude, *country* is the country the place belongs to, *subregion* is an administrative division of a lower level than country.

Titel_en	lat	lon	country	subregion	style
Abilene Municipal Airport	38.9041666666667	-97.2358333333333	US		airport
Abilene Regional Airport	32.4113888888889	-99.6819444444445	US		airport
Abilene, Kansas	38.9166666666667	-97.2166666666667	US	KS	city k
Abilene, Texas	32.45	-99.7333333333333	US		city e

Figure 3.11: Results of the search for the toponym "Abilene" in Wikipedia-World.

Subregion and *country* fields are processed as first level and second level containing entities, respectively. In the case the *subregion* field is empty, we use the specialisation in the *Titel_en* field as first level containing entity. Note that styles fields (in this example *city k* and *city e*) were normalised to fit with WordNet classes. In this case, we transformed *city k* and *city e* into *city*. The calculated weights can be observed in Table 3.3.

Table 3.3: Resulting weights for the mapping of the toponym "Abilene".

Entity	Weight
Abilene Municipal Airport	1.0
Abilene Regional Airport	1.0
Abilene, Kansas	2.0
Abilene, Texas	3.6

The weight of the two airports derive from the match for "US" as the second level containing entity ($m(h(h_s), c(c(l))) = 1$). "Abilene, Kansas" benefits also from an exact name match ($m(w_i, l) = 1$). The highest weight is obtained for "Abilene, Texas" since there are the same matches as before, but also they share the same containing entity ($m(h_s, c(l)) = 1$) and there are matches in the class part both in gloss (a *city* in central Texas) and in the direct hypernym.

The final resource is constituted by two plain text files: the most important is a single text file that contains 2,012 labeled synsets, where each row is constituted by an offset (WordNet version 2.0) together with its latitude and longitude, separated by tabs. This file is named `WNCoord.dat`. A small sample of the content of this file, corresponding to the synsets *Marshall Islands*, *Kwajalein* and *Tuvalu* can be found in Figure 3.12.

The other file contains a human-readable version of the database, where each line contains the synset description and the entry in the database: *Acapulco a port and fash-*

```
08294059 7.06666666667 171.266666667
08294488 9.19388888889 167.459722222
08294965 -7.475 178.005555556
```

Figure 3.12: Sample of Geo-WordNet corresponding to the Marhsall Islands, Kwajalein and Tuvalu.

ionable resort city on the Pacific coast of southern Mexico; known for beaches and water sports (including cliff diving) ('Acapulco', 16.851666666666699, -99.9097222222222, 'MX', 'GRO', 'city c').

An advantage of Geo-WordNet is that the WordNet meronymy relationship can be used to approximate area shapes. One of the critics moved from GIS researchers to gazetteers is that they usually associate a single pair of coordinates to areas, with a loss of precision with respect to GIS databases where areas (like countries) are stored as shapes, rivers as lines, etc. With Geo-WordNet this problem can be partially solved, using meronyms coordinates to build a Convex Hull (CH)[1] that approximates the boundaries of the area. For instance, in Figure 3.13 a), "South America" is represented by the point associated in Geo-WordNet to the "South America" synset. In Figure 3.13 b), the meronyms of "South America", corresponding to countries, were added in red, obtaining an approximated CH that covers partially the area occupied by South America. Finally, in Figure 3.13 c) were used the meronyms of countries (cities and administrative divisions), obtaining a CH that covers almost completely the area of South America.

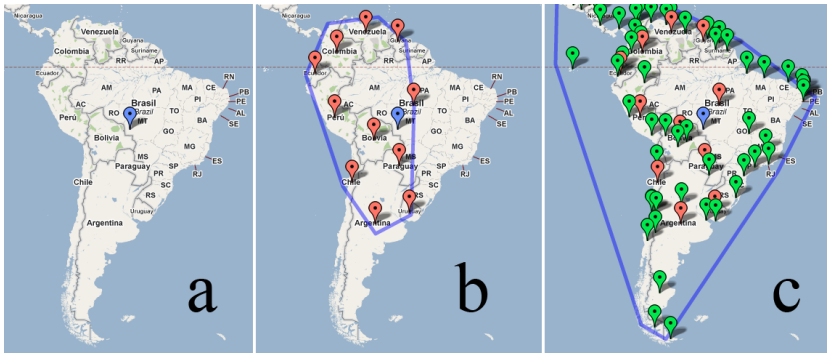


Figure 3.13: Approximation of South America boundaries using WordNet meronyms.

Geo-WordNet can be downloaded from the Natural Language Engineering Lab web-

[1] the minimal convex polygon that includes all the points in a given set

site: *http://www.dsic.upv.es/grupos/nle*.

3.4 Geographically Tagged Corpora

The lack of a disambiguated corpus has been a major obstacle to the evaluation of the effect of word sense ambiguity in IR. Sanderson (1996) had to introduce ambiguity creating pseudo-words; Gonzalo et al. (1998) adapted the SemCor corpus, which is not usually used to evaluate IR systems. In toponym disambiguation, this represented a major problem, too. Currently, few text corpora can be used to evaluate toponym disambiguation methods or the effects of TD on IR. In this section we present some text corpora in which toponyms have been labelled with geographical coordinates or with some unique identifier that allows to assign a toponym its coordinates. These resources are GeoSemCor, the CLIR-WSD collection, the TR-CoNLL collection and the ACE 2005 SpatialML corpus. The first two were used in this work: GeoSemCor, in particular, was tagged in the framework of this Ph.D. thesis work and made it publicly available at the NLE Lab web page. CLIR-WSD was developed for the CLIR-WSD and QA-WSD tasks and made available to CLEF participants. Although it was not created explicitely for TD, it was large enough to carry out GIR experiments. TR-CoNLL unfortunately seems to be not so easily accessible[1] and it was not considered. The ACE 2005 Spatial ML corpus is an annotation of data used in the 2005 Automatic Content Extraction evaluation exercise[2]. We did not use it because of its limited size, as it can be observed in Table 3.4, where the characteristics of the different corpora are shown. Only CLIR-WSD is large enough to carry out GIR experiments, whereas both GeoSemCor and TR-CoNLL represent good choices for TD evaluation experiments, due to their size and the manual labelling of the toponyms. We chose GeoSemCor for the evaluation experiments because of its availability.

Table 3.4: Comparison of evaluation corpora for Toponym Disambiguation.

name	geo label source	availability	labelling	# of instances	# of docs
GeoSemCor	WordNet 2.0	free	manual	1,210	352
CLIR-WSD	WordNet 1.6	CLEF part.	automatic	354,247	169,477
TR-CoNLL	Custom (TextGIS)	not-free	manual	6,980	946
SpatialML	Custom (IGDB)	LDC	manual	4,783	104

[1] We made several attempts to obtain it, without success.

[2] `http://www.itl.nist.gov/iad/mig//tests/ace/2005/index.html`

3.4.1 GeoSemCor

GeoSemCor was obtained from SemCor, the most used corpus for the evaluation of WSD methods. SemCor is a collection of texts extracted from the Brown Corpus of American English, where each word has been labelled with a WordNet sense (synset). In GeoSemCor, toponyms were automatically tagged with a *geo* attribute. The toponyms were identified with the help of WordNet itself: if a synset (corresponding to the combination of the word – the *lemma* tag – with its sense label – *wnsn*) had the synset *location* among its hypernyms, then the respective word was labelled with a *geo* tag (for instance: `<wf geo=true cmd=done pos=NN lemma=dallas wnsn=1 lexsn=1:15:00::>Dallas</wf>`). The resulting *GeoSemCor* collection contains 1,210 toponym instances and is freely available from the NLE Lab web page: `http://www.dsic.upv.es/grupos/nle/`. Sense labels are those of WordNet 2.0 . The format is based on the SGML used for SemCor. Details of GeoSemCor are shown in Table 3.5. Note that the polysemy count is based on the number of senses in WordNet and not on the number of places that a name can represent. For instance, "London" in WordNet has two senses, but only the first of them corresponds to the city, because the second one is the surname of the American writer "Jack London". However, only the instances related to toponyms have been labelled with the *geo* tag in GeoSemCor.

Table 3.5: GeoSemCor statistics.

total toponyms	1,210
polysemous toponyms	709
avg. polysemy	2.151
labelled with MF sense	1,140(94.2%)
labelled with 2nd sense	53
labelled with a sense > 2	17

In Figure 3.14 a section of text from the `br-m02` file of GeoSemCor is displayed.

The *cmd* attribute indicates whether the tagged word is a stop-word (ignore) or not (done). The *wnsn* and *lexsn* attributes indicate the senses of the tagged word. The attribute *lemma* indicates the base form of the tagged word. Finally, *geo=true* tells us that the word represents a geographical location. The 's' tag indicates the sentence boundaries.

```
<s snum=74>
<wf cmd=done pos=RB lemma=here wnsn=1 lexsn=4:02:00::>Here</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=people wnsn=1 lexsn=1:14:00::>peoples</wf>
<wf cmd=done pos=VB lemma=speak wnsn=3 lexsn=2:32:02::>spoke</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=tongue wnsn=2 lexsn=1:10:00::>tongue</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf geo=true cmd=done pos=NN lemma=iceland wnsn=1 lexsn=1:15:00::>Iceland</wf>
<wf cmd=ignore pos=IN>because</wf>
<wf cmd=ignore pos=IN>that</wf>
<wf cmd=done pos=NN lemma=island wnsn=1 lexsn=1:17:00::>island</wf>
<wf cmd=done pos=VBD ot=notag>had</wf>
<wf cmd=done pos=VB ot=idiom>gotten_the_jump_on</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=hawaiian wnsn=1 lexsn=1:10:00::>Hawaiian</wf>
<wf cmd=done pos=NN lemma=american wnsn=1 lexsn=1:18:00::>Americans</wf>
[...]
</s>
```

Figure 3.14: Section of the `br-m02` file of GeoSemCor.

3.4.2 CLIR-WSD

Recently, the lack of disambiguated collections has been compensated by the CLIR-WSD task[1], a task introduced in CLEF 2008. The CLIR-WSD collection is a disambiguated collection developed for the CLIR-WSD and QA-WSD tasks organised by Eneko Agirre of the University of Basque Country. This collection contains 104,112 toponyms labeled with WordNet 1.6 senses. The collection is composed by the 169,477 documents of the GeoCLEF collection: the Glasgow Herald 1995 (GH95) and the Los Angeles Times 1994 (LAT94). Toponyms have been automatically disambiguated using k-Nearest Neighbour and Singular Value Decomposition developed at the University of Basque Country (UBC) by Agirre and Lopez de Lacalle (2007). Another version, where toponyms were disambiguated using a method based on parallel corpora by Ng et al. (2003), was also offered to participants but, since it was not posssible to know the exact performance in disambiguation of the two methods on the collection, we opted to

[1] `http://ixa2.si.ehu.es/clirwsd/`

carry out the experiments only with the UBC tagged version. Below we show a portion of the labelled collection, corresponding to the text "*Old Dumbarton Road, Glasgow*" in document `GH951123-000164`.

```
<TERM ID="GH951123-000164-221" LEMA="old" POS="NNP">
<WF>Old</WF>
<SYNSET SCORE="1" CODE="10849502-n"/>
</TERM>
<TERM ID="GH951123-000164-222" LEMA="Dumbarton" POS="NNP">
<WF>Dumbarton</WF>
</TERM>
<TERM ID="GH951123-000164-223" LEMA="road" POS="NNP">
<WF>Road</WF>
<SYNSET SCORE="0" CODE="00112808-n"/>
<SYNSET SCORE="1" CODE="03243979-n"/>
</TERM>
<TERM ID="GH951123-000164-224" LEMA="," POS=",">
<WF>,</WF>
</TERM>
<TERM ID="GH951123-000164-225" LEMA="glasgow" POS="NNP">
<WF>Glasgow</WF>
<SYNSET SCORE="1" CODE="06505249-n"/>
</TERM>
```

The sense repository used for these collections is WordNet 1.6. Senses are coded as pairs "*offset-POS*", where POS can be n, v, r, or a, standing for *noun*, *verb*, *adverb* and *adjective*, respectively. During the indexing phase, we assumed the synset with the highest score to be the "right" sense for the toponym. Unfortunately, WordNet 1.6 contains less geographical synsets than WordNet 2.0 and WordNet 3.0 (see Table 3.6). For instance, "Aberdeen" has only one sense in WordNet 1.6, whereas it appears in WordNet 2.0 with 4 possible senses (one from Scotland and three from the US). Therefore, some errors appear in the labelled data, such as "Valencia, CA", a community located in Los Angeles county labelled as "Valencia, Spain". However, since a gold standard does not exists for this collection, it was not possible to estimate the disambiguation accuracy.

Table 3.6: Comparison of the number of geographical synsets among different WordNet versions.

feature	WordNet 1.6	WordNet 2.0	WordNet 3.0
cities	328	619	661
capitals	190	191	192
rivers	113	180	200
mountains	55	66	68
lakes	19	41	43

3.4.3 TR-CoNLL

The TR-CoNLL corpus, developed by Leidner (2006), consists in a collection of documents of the Reuters news agency labelled with toponym referents. It was announced in 2006, but it was made available only in 2009. This resource is based on the Reuters Corpus Volume I (RCV1)[1], a document collection containing all English language news stories produced by Reuters journalists between August 20, 1996 and August 19, 1997. Among other uses, the RCV1 corpus is frequently used for benchmarking automatic text classification methods. A subset of 946 documents was manually annotated with coordinates from a custom gazetteer derived from Geonames, using a XML-based annotation scheme named TRML. The resulting resource contains 6,980 toponym instances, with 1,299 unique toponyms.

3.4.4 SpatialML

The ACE 2005 SpatialML corpus by Mani et al. (2008) is a manually tagged (inter-annotator agreement: 77%) collection of documents from the corpus used in the Automatic Content Extraction evaluation held in 2005. This corpus, drawn mainly from broadcast conversation, broadcast news, news magazine, newsgroups, and weblogs, contains 4,783 toponyms instances, of which 915 are unique. Each document is annotated using SpatialML, an XML-based language which allows the recording of toponyms and their geographically relevant attributes, such as their lat/lon position and feature type. The 104 documents are news wire, which are focused on broadly distributed geographic audience. This is reflected on the geographic entities that can be found in the corpus: 1,685 countries, 255 administrative divisions, 454 capital cities and 178 populated places. This corpus can be obtained at the Linguistic Data Consortium (LDC)[2] for a

[1] `about.reuters.com/researchandstandards/corpus/`

[2] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T03`

fee of 500 or $1,000 US$\$$.

Chapter 4

Toponym Disambiguation

Toponym Disambiguation or *Resolution* can be defined as the task of assigning to an ambiguous place name the reference to the actual location that it represents in a given context. It can be seen as a specialised form of Word Sense Disambiguation (WSD). The problem of WSD is defined as the task of automatically assigning the most appropriate meaning to a polysemous (i.e., with more than one meaning) word within a given context. Many research works attempted to deal with the ambiguity of human language, under the assumption that ambiguity does worsen the performance of various NLP tasks, such as machine translation and information retrieval. The work of Lesk (1986) was based on the textual definitions of dictionaries: given a word to disambiguate, he looked to the context of the word to find partial matching with the definitions in the dictionary. For instance, suppose that we have to disambiguate "Cambridge"; if we look at the definitions of "Cambridge" in WordNet:

1. Cambridge: a city in Massachusetts just to the north of Boston; site of Harvard University and the Massachusetts Institute of Technology

2. Cambridge: a city in eastern England on the River Cam; site of Cambridge University

the presence of "Boston", "Massachussetts" or "Harvard" in the context of "Cambridge" would assign to it the first sense. The presence of "England" and "Cam" would assign to "Cambridge" the second sense. The word "university" in context is not discriminating since it appears in both definitions. This method was refined later by Banerjee and Pedersen (2002), who searched also in the textual definitions of synsets connected to the synsets of the word to disambiguate. For instance, for the previous example they would have included the definitions of the synsets related to the two

meanings of "Cambridge" shown in Figure 4.1.

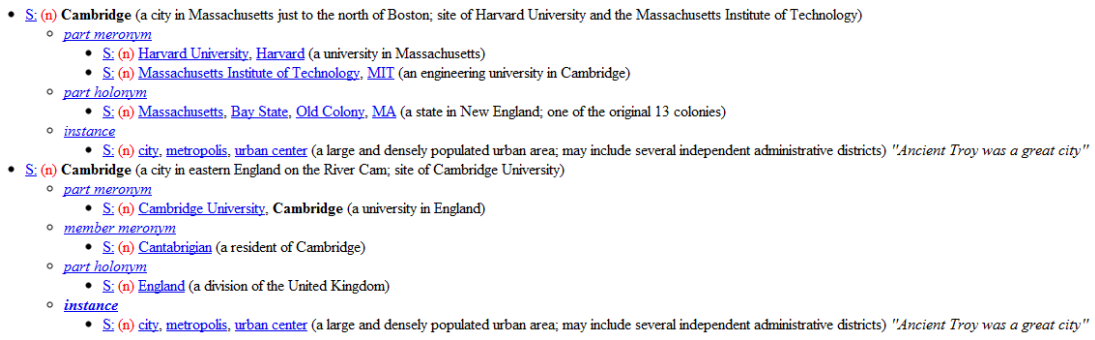


Figure 4.1: Synsets corresponding to "Cambridge" and their relatives in WordNet 3.0.

Lesk algorithm was prone to disambiguation errors, but marked an important step in WSD research since it opened the way to the creation of resources like WordNet and Semcor which were later used to carry out comparative evaluations of WSD methods, especially in the Senseval[1] and Semeval[2] workshops. In these evaluation frameworks emerged a clear distinction between method that were based only on dictionaries or ontologies (*knowledge-based* methods) and those which used machine learning techniques (*data-driven* methods), with the second ones often obtaining better results, although labelled corpora are usually not commonly available. Particularly interesting are the methods developed by Mihalcea (2007), which used Wikipedia as a training corpus, and Ng et al. (2003), which exploited parallel texts, on the basis that some words are ambiguous in a language but not in another one (for instance, "calcio" in Italian may mean both "Calcium" and "football").

The measures used for the evaluation of Toponym Disambiguation methods are also the same used in the WSD task. There are four measures that are commonly used: *Precision* or *Accuracy*, *Recall*, *Coverage* and *F-measure*. *Precision* is calculated as the number of correctly disambiguated toponyms divided by the number of disambiguated toponyms. *Recall* is the number of correctly disambiguated toponyms divided by the total number of toponyms in the collection. *Coverage* is the number of disambiguated toponyms, either correctly or wrongly, divided the total number of toponyms. Finally, the *F-measure* is a combination of precision and recall, calculated as their harmonic mean:

$$\frac{2 * precision * recall}{precision + recall} \tag{4.1}$$

[1] `http://www.senseval.org/`
[2] `http://semeval2.fbk.eu`

A taxonomy for TD methods that extends the taxonomy for WSD methods has been proposed in Buscaldi and Rosso (2008a). According to this taxonomy, existing methods for the disambiguation of toponyms may be subdivided in three categories:

- *map-based*: methods that use an explicit representation of places on a map;
- *knowledge-based*: they exploit external knowledge sources such as gazetteers, Wikipedia or ontologies;
- *data-driven* or *supervised*: based on standard machine learning techniques.

Among the first ones, Smith and Crane (2001) proposed a method for toponym resolution based on the geographical coordinates of places: the locations in the context are arranged in a map, weighted by the number of times they appear. Then a centroid of this map is calculated and compared with the actual locations related to the ambiguous toponym. The location closest to the 'context map' centroid is selected as the right one. They report precisions of between 74% and 93% (depending on test configuration), where precision is calculated as the number of correctly disambiguated toponyms divided by the number of toponyms in the test collection. The GIPSY subsystem by Woodruff and Plaunt (1994) is also based on spatial coordinates, although in this case they are used to build polygons. Woodruff and Plaunt (1994) report issues with noise and runtime problems. Pasley et al. (2007) also used a map-based method to resolve toponyms at different scale levels, from a regional level (Midlands) to a Sheffield suburbs of 1/2km by 1/2km. For each geo-reference, they selected the possible coordinates closest to the context centroid point as the most plausible location of that geo-reference for that specific document.

The majority of the TD methods proposed in literature are based on rules that exploits some specific kind of information included in a knowledge source. Gazetteers were used as knowledge sources in the methods of Olligschlaeger and Hauptmann (1999) and Rauch et al. (2003). Olligschlaeger and Hauptmann (1999) disambiguated toponyms using a cascade of rules. First, toponym occurrences that are ambiguous in one place of the document are resolved by propagating interpretations of other occurrences in the same document based on the "one referent per discourse" assumption. For example, using this heuristic together with a set of unspecified patterns, Cambridge can be resolved to Cambridge, MA, USA, in case Cambridge, MA occurs elsewhere in the same discourse. Besides the discourse heuristic, the information about states and countries contained in the gazetteer (a commercial global gazetteer of $80,000$ places) is used in the form of a "superordinate mention" heuristic. For instance, Paris is taken to refer to

Paris, France, if France is mentioned elsewhere. Olligschlaeger and Hauptmann (1999) report a precision of 75% for their rule-based method, correctly disambiguating 269 out of 357 instances. In the work by Rauch et al. (2003), population data are used in order to disambiguate toponyms, exploiting the fact that references to populous places are most frequent that to less populated ones, to the presence of postal addresses. Amitay et al. (2004) integrated the population heuristic together with a path of prefixes extracted from a spatial ontology. For instance, given the following two candidates for the disambiguation of "Berlin": *Europe/Germany/Berlin*, *NorthAmerica/USA/CT/Berlin* and the context "Potsdam" (*Europe/Germany/Potsdam*) they assign to "Berlin" in the document the place *Europe/Germany/Berlin*. They report an accuracy of 73.3% on a random 200-page sample from a $1,200,000$ TREC corpus of U.S. government Web pages.

Wikipedia was used in Overell et al. (2006) to develop WikiDisambiguator, which takes advantage from article templates, categories and referents (links to other articles in Wikipedia). They evaluated disambiguation over a set of manually annotated "ground truth" data ($1,694$ locations from a random article sample of the online encyclopedia Wikipedia), reporting 82.8% in resolution accuracy. Andogah et al. (2008) combined the "one referent per discourse" heuristic with place type information (city, administration division, state), selecting the toponym having the same type of neighbouring toponyms (if "New York" appears together with "London", then it is more probable that the document is talking about the city of New York and not the state), and the resolution of the geographical scope of a document, limiting the search for candidates within the geographical area interested by the theme of the document. Their results over Leidner's TR-CoNLL corpus are of a precision of 52.3% if scope resolution is used, and 77.5% in the case it is not used.

Data-driven methods, although being widely used in WSD, are not commonly used in TD. The weakness of supervised methods consists in the need for a large quantity of training data in order to obtain a high precision, data that currently are not available for the TD task. Moreover, the inability to classify unseen toponyms is also a major problem that affects this class of methods. A Naïve Bayes classifier is used by Smith and Mann (2003) to classify place names with respect to the U.S. state or foreign country. They report precisions between 21.8% and 87.4%, depending on the test collection used. Garbin and Mani (2005) used a rule-based classifier, obtaining precisions between 65.3% and 88.4%, also depending on the test corpus. Li et al. (2006a) developed a probabilistic TD system which used the following features: local contextual information (geo-term pairs that occur in close proximity to each other in the text,

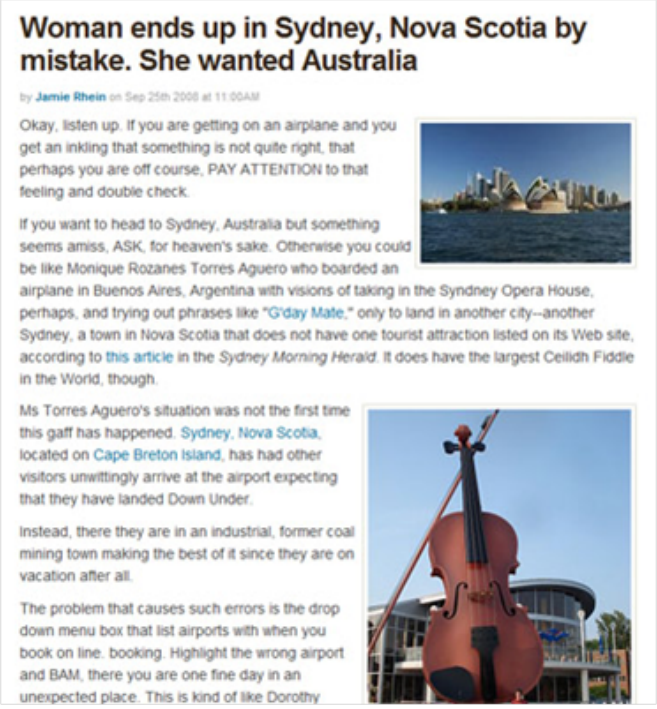
such as "Washington, D.C.", population statistics, geographical trigger words such as "county" or "lake") and global contextual information (the occurrence of countries or states can be used to boost location candidates if the document makes reference to one of its ancestors in the hierarchy). A peculiarity of the TD method by Li et al. (2006a) is that toponyms are not completely disambiguated: improbable candidates for disambiguation end up with non-zero, but small, weights, meaning that although in a document "England" has been found near to "London", there exists still a small probability that the author of the document is referring instead to "London" in Ontario, Canada. Martins et al. (2010) used a stacked learning approach, in which a first learner based on a Hidden Markov Model is used to annotate place references, and then a second learner implementing a regression through a Support Vector Machine is used to rank the possible disambiguations for the references that were initially annotated. Their method compares favorably against commercial state-of-the-art systems such as Yahoo! Placemaker[1] over various collections in different languages (Spanish, English and Portuguese). They report F1 measures between 22.6% and 67.5%, depending on the language and the collection considered.

4.1 Measuring the Ambiguity of Toponyms

How big is the problem of toponym ambiguity? As for the ambiguity of other kinds of word in natural languages, the ambiguity of toponym is closely related to the use people make of them. For instance, a musician may ignore that "bass" is not only a musical instrument, but also a type of fish. In the same way, many people in the world ignores that *Sydney* is not only the name of one of the most important cities in Australia, but also a city in Nova Scotia, Canada, which in some cases lead to errors like the one in Figure 4.2.

Dictionaries may be used as a reference for the senses that may be assigned to a word, or, in this case, to a toponym. An issue with toponyms is that the granularity of the gazetteers may vary greatly from one resource to another, with the result that the ambiguity for a given toponym may not be the same in different gazetteers. For instance, Smith and Mann (2003) studied the ambiguity of toponyms at continent level with the Getty TGN, obtaining that almost the 60% of names used in North and Central America were ambiguous (i.e., for each toponym there exist at least 2 places with the same name). However, if toponym ambiguity is calculated on Geonames, these values change significantly. The comparison of the average ambiguity values is shown in Table

[1] `http://developer.yahoo.com/geo/placemaker/`



Woman ends up in Sydney, Nova Scotia by mistake. She wanted Australia

by Jamie Rhein on Sep 25th 2008 at 11:00AM

Okay, listen up. If you are getting on an airplane and you get an inkling that something is not quite right, that perhaps you are off course, PAY ATTENTION to that feeling and double check.

If you want to head to Sydney, Australia but something seems amiss, ASK, for heaven's sake. Otherwise you could be like Monique Rozanes Torres Aguero who boarded an airplane in Buenos Aires, Argentina with visions of taking in the Syndney Opera House, perhaps, and trying out phrases like "G'day Mate," only to land in another city--another Sydney, a town in Nova Scotia that does not have one tourist attraction listed on its Web site, according to this article in the *Sydney Morning Herald*. It does have the largest Ceilidh Fiddle in the World, though.

Ms Torres Aguero's situation was not the first time this gaff has happened. Sydney, Nova Scotia, located on Cape Breton Island, has had other visitors unwittingly arrive at the airport expecting that they have landed Down Under.

Instead, there they are in an industrial, former coal mining town making the best of it since they are on vacation after all.

The problem that causes such errors is the drop down menu box that list airports with when you book on line. booking. Highlight the wrong airport and BAM, there you are one fine day in an unexpected place. This is kind of like Dorothy

Figure 4.2: Flying to the "wrong" Sydney.

4.1. In Table 4.2 are listed the most ambiguous toponyms, according to Geonames, GeoPlanet and WordNet respectively. From this table it can be appreciated the level of detail of the various resources, since there are $1,536$ places named "San Antonio" in Geonames, almost 7 times as many as in GeoPlanet, while in WordNet the most ambiguous toponym has only 5 possible referents.

The top 10 territories ranked by the percentage of ambiguous toponyms, calculated on Geonames, are listed in Table 4.3. Total indicates the total number of places in each territory, unique the number of distinct toponyms used in that territory, ambiguity ratio is the ratio $total/unique$, ambiguous toponyms indicates the number of toponyms that may refer to more than one place. The ambiguity ratio is not a precise measure of ambiguity, but it could be used as an estimate of how many referents exist for each ambiguous toponym on average. The percentage of ambiguous toponyms measures how many toponyms are used for more than one place.

In Table 4.2 we can see that "San Francisco" is one of the most ambiguous toponyms, according both to Geonames and GeoPlanet. However, is it possible to state that "San Francisco" is an highly ambiguous toponym? Most people in the world probably know only the "San Francisco" in California. Therefore, it is important to consider ambiguity

Table 4.1: Ambiguous toponyms percentage, grouped by continent.

Continent	% ambiguous (TGN)	% ambiguous (Geonames)
North and Central America	57.1%	9.5%
Oceania	29.2%	10.7%
South America	25.0%	10.9%
Asia	20.3%	9.4%
Africa	18.2%	9.5%
Europe	16.6%	12.6%

Table 4.2: Most ambiguous toponyms in Geonames, GeoPlanet and WordNet.

Geonames		GeoPlanet		WordNet	
Toponym	# of Places	Toponym	# of Places	Toponym	# of Places
San Antonio	1,536	Rampur	319	Victoria	5
Mill Creek	1,529	Fairview	250	Aberdeen	4
Spring Creek	1,483	Midway	233	Columbia	4
San José	1,360	San Antonio	227	Jackson	4
Dry Creek	1,269	Benito Juarez	218	Avon	3
Santa Rosa	1,185	Santa Cruz	201	Columbus	3
Bear Creek	1,086	Guadalupe	193	Greenville	3
Mud Lake	1,073	San Isidro	192	Bangor	3
Krajan	1,030	Gopalpur	186	Salem	3
San Francisco	929	San Francisco	177	Kingston	3

Table 4.3: Territories with most ambiguous toponyms, according to Geonames.

Territory	Total	Unique	Amb. ratio	Amb. toponyms	% ambiguous
Marshall Islands	3,250	1,833	1.773	983	53.63%
France	118,032	71,891	1.642	35,621	49.55%
Palau	1,351	925	1.461	390	42.16%
Cuba	1,7820	1,2316	1.447	4,185	33.98%
Burundi	8,768	4,898	1.790	1,602	32.71%
Italy	46,380	34,733	1.335	9,510	27.38%
New Zealand	63,600	43,477	1.463	11,130	25.60%
Micronesia	5,249	4,106	1.278	1,051	25.60%
Brazil	78,006	44,897	1.737	11,128	24.79%

not only from an absolute perspective, but also from the point of view of usage. In Table 4.4 the top 15 toponyms, ranked by frequency, extracted from the GeoCLEF collection, which is composed by news stories from the Los Angeles Times (1994) and Glasgow Herald (1995), as described in Section 2.1.4. From the table, it seems that the toponyms reflect the context of the readers of the selected news sources, following the "Steinberg hypothesis". Figures 4.4 and 4.5 have been processed by examining the GeoCLEF collection labelled with WordNet synsets developed by the University of Basque Country for the CLIR-WSD task. The histograms represents the number of toponyms found in the Los Angeles Times (LAT94) and Glasgow Herald (GH95) portions of the collection within a certain distance from Los Angeles (California) and Glasgow (Scotland). In Figure 4.4 it could be observed that in `LAT94` there are more toponyms within 6,000 km from Los Angeles than in `GH95`, and in Figure 4.5 the number of toponyms observed within 1,200 km from Glasgow is higher in `GH95` than in `LAT94`. It should be noted that the scope of WordNet is mostly on United States and Great Britain, and in general the English-speaking part of the world, resulting in higher toponym density for the areas corresponding to the USA and the UK.

Table 4.4: Most frequent toponyms in the GeoCLEF collection.

Toponym	Count	Amb. (WN)	Amb. (Geonames)
United States	63,813	n	n
Scotland	35,004	n	y
California	29,772	n	y
Los Angeles	26,434	n	y
United Kingdom	22,533	n	n
Glasgow	17,793	n	y
Washington	13,720	y	y
New York	13,573	y	y
London	11,676	n	y
England	11,437	n	y
Edinburgh	11,072	n	y
Europe	10,898	n	n
Japan	9,444	n	y
Soviet Union	8,350	n	n
Hollywood	8,242	n	y

In Table 4.4 it can be noted that only 2 out of 15 toponyms are ambiguous according

to WordNet, whereas 11 out of 15 are ambiguous according to Geonames. However, "Scotland" in LAT94 or GH95 never refers to, e.g. "Scotland", the county in North Carolina, although "Scotland" and "North Carolina" appear together in 25 documents. "Glasgow" appears together with "Delaware" in 3 documents, but it is always referring to the Scottish Glasgow and not the Delaware one. On the other hand, there are at least 25 documents where "Washington" refers to the State of Washington and not to the U.S. capital. Therefore, choosing WordNet as a resource for toponym ambiguity to work on the GeoCLEF collection seems to be reasonable, given the scope of the news stories. Of course, it would be completely inappropriate to use WordNet on a news collection from Delaware: in the caption of the `http://www.delawareonline.com/` online news of Figure 4.3 we can see that the Glasgow named in this source is not the Scottish one. A solution to this issue is to "customise" gazetteers depending on the collection they are going to be used for. A case study using an Italian newspaper and a gazetteer that includes details up to the level of street names is described in Section 4.4.

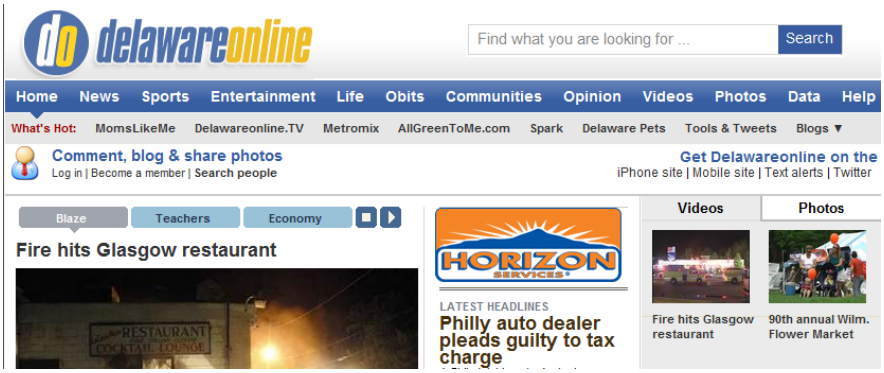


Figure 4.3: Capture from the home page of Delaware online.

4.2 Toponym Disambiguation using Conceptual Density

Using WordNet as a resource for GIR is not limited to using it as a "sense repository" for toponyms. Its structured data can be exploited to adapt WSD algorithms based on WordNet to the problem of Toponym Disambiguation. One of such algorithms is the *Conceptual Density* (CD) algorithm, introduced by Agirre and Rigau (1996) as a measure of the correlation between the sense of a given word and its context. It is computed on WordNet sub-hierarchies, determined by the hypernymy relationship. The disambiguation algorithm by means of CD consists of the following steps:

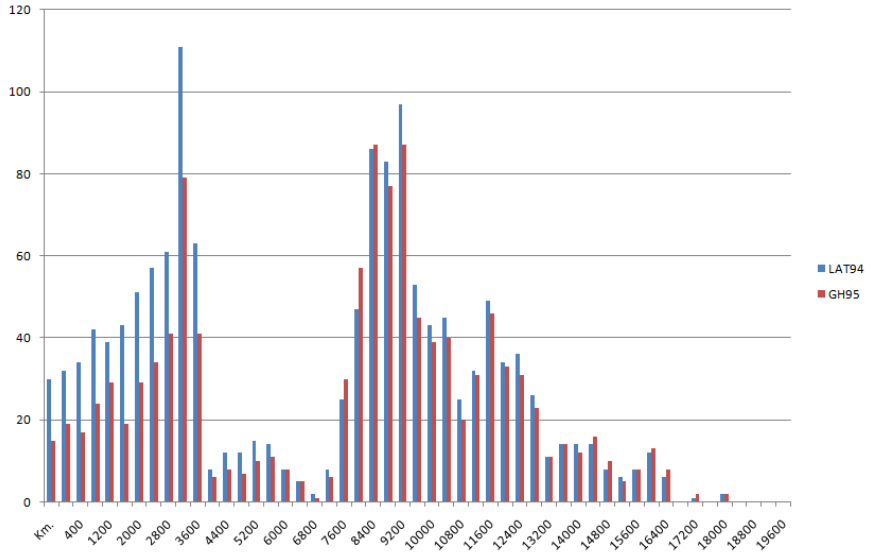


Figure 4.4: Number of toponyms in the GeoCLEF collection, grouped by distances from Los Angeles, CA.

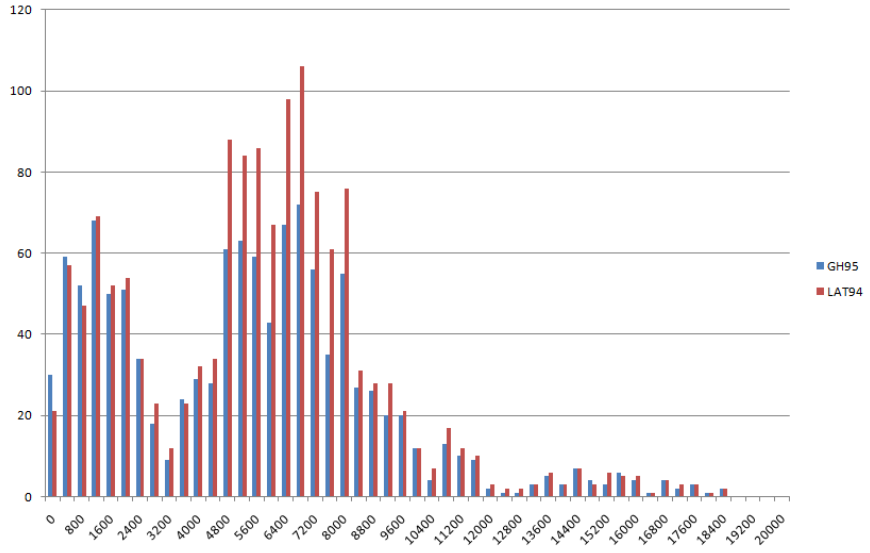


Figure 4.5: Number of toponyms in the GeoCLEF collection, grouped by distances from Glasgow, Scotland.

1. Select the next ambiguous word w, with $|w|$ senses;
2. Select the context $\bar{c}_w$, i.e. a sequence of words, for w;
3. Build $|w|$ subhierarchies, one for each sense of w;
4. For each sense s of w, calculate CD_s;
5. Assign to w the sense which maximises CD_s.

We modified the original Conceptual Density formula used to calculate the density of a WordNet sub-hierarchy s in order to take into account also the rank of frequency f (Rosso et al. (2003)):

$$CD(m, f, n) = m^{\alpha} \left(\frac{m}{n}\right)^{\log f}, \tag{4.2}$$

where m represents the count of *relevant* synsets that are contained in the sub-hierarchy, n is the total number of synsets in the sub-hierarchy, and f is the rank of frequency of the word sense related to the sub-hierarchy (e.g. 1 for the most frequent sense, 2 for the second one, etc.). The inclusion of the frequency rank means that less frequent senses are selected only when $m/n \geq 1$. Relevant synsets are both the synsets corresponding to the meanings of the word to disambiguate and of the context words.

The WSD system based on this formula obtained 81.5% in precision over the nouns in the SemCor (baseline: 75.5%, calculated by assigning to each noun its most frequent sense), and participated at the Senseval-3 competition as the CIAOSENSO system (Buscaldi et al. (2004)), obtaining 75.3% in precision over nouns in the all-words task (baseline: 70.1%). These results were obtained with a context window of only two nouns, the one preceding and the one following the word to disambiguate.

With respect to toponym disambiguation, the hypernymy relation cannot be used, since both instances of the same toponym share the same hypernym: for instance *Cambridge(1)* and *Cambridge(2)* are both instances of the '*city*' concept and, therefore, they share the same hypernyms (this has been changed in WordNet 3.0 where now Cambridge is connected to the 'city' concept by means of the '*instance of*' relation). The result, applying the original algorithm, would be that the sub-hierarchies would be composed only by the synsets of the two senses of '*Cambridge*', and the algorithm would leave the word undisambiguated because the sub-hierarchies density are the same (in both cases it is 1).

The solution is to consider the *holonymy* relationship instead of hypernymy. With this relationship it is possible to create sub-hierarchies that allow to discern different locations having the same name. For instance, the last three holonyms for '*Cambridge*' are:

(1) `Cambridge` → `England` → `UK`

(2) `Cambridge` → `Massachusetts` → `New England` → `USA`

The best choice for context words is represented by other place names, because holonymy is always defined through them and because they constitute the actual 'geographical' context of the toponym to disambiguate. In Figure 4.6 we can see an example of a holonym tree obtained for the disambiguation of '*Georgia*' with the context '*Atlanta*', '*Savannah*' and '*Texas*', from the following fragment of text extracted from the `br-a01` file of SemCor:

> "Hartsfield has been mayor of **Atlanta**, with exception of one brief interlude, since 1937. His political career goes back to his election to city council in 1923. The mayor's present term of office expires Jan. 1. He will be succeeded by Ivan Allen Jr., who became a candidate in the Sept. 13 primary after Mayor Hartsfield announced that he would not run for reelection. **Georgia** Republicans are getting strong encouragement to enter a candidate in the 1962 governor's race, a top official said Wednesday. Robert Snodgrass, state GOP chairman, said a meeting held Tuesday night in Blue Ridge brought enthusiastic responses from the audience. State Party Chairman James W. Dorsey added that enthusiasm was picking up for a state rally to be held Sept. 8 in **Savannah** at which newly elected **Texas** Sen. John Tower will be the featured speaker."

According to WordNet, *Georgia* may refer to 'a state in southeastern United States' or a 'republic in Asia Minor on the Black Sea separated from Russia by the Caucasus mountains'.

As one would expect, the holonyms of the context words populate exclusively the sub-hierarchy related to the first sense (the area filled with a diagonal hatching in Figure 4.6); this is reflected in the CD formula, which returns a CD value 4.29 for the first sense ($m = 8$, $n = 11$, $f = 1$) and 0.33 for the second one ($m = 1$, $n = 5$, $f = 2$). In this work, we considered as relevant also those synsets which belong to the paths of the context words that fall into a sub-hierarchy of the toponym to disambiguate.

4.2.1 Evaluation

The WordNet-based toponym disambiguator described in the previous section was tested over a collection of $1,210$ toponyms. Its results were compared with the *Most Frequent* (MF) baseline, obtained by assigning to each toponym its most frequent sense,

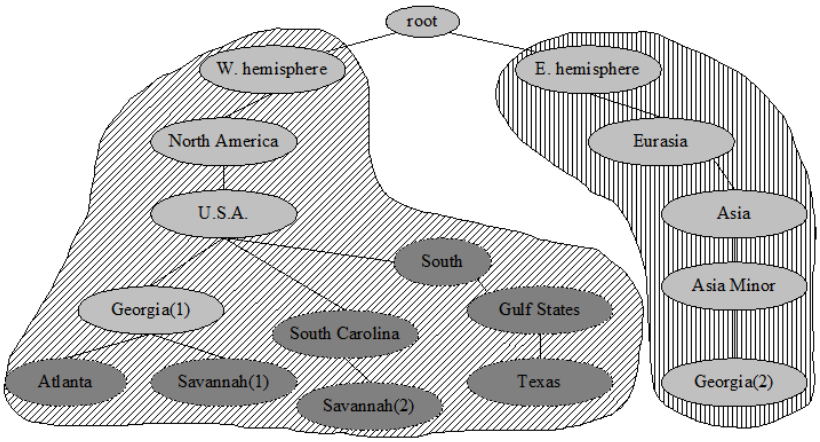


Figure 4.6: Example of subhierarchies obtained for *Georgia* with context extracted from a fragment of the `br-a01` file of SemCor.

and with another WordNet-based method which uses its glosses, and those of its context words, to disambiguate it. The corpus used for the evaluation of the algorithm was the GeoSemCor corpus.

For comparison, the method by Banerjee and Pedersen (2002) was also used. This method represent an enhancement of the well-known dictionary-based algorithm proposed by Lesk (1986) and is also based on WordNet. This enhancement consists in taking into account also the glosses of concepts related to the word to disambiguate by means of various WordNet relationships. Then the similarity between a sense of the word and the context is calculated by means of *overlaps*. The word is assigned the sense which obtains the best overlap match with the glosses of the context words and their related synsets. In WordNet (version 2.0) there can be 7 relations for each word, this means that for every pair of words up to 49 relations have to be considered. The similarity measure based on Lesk has been demonstrated as one of the best measures for the semantic relatedness of two concepts by Patwardhan et al. (2003).

The experiments were carried out considering three kinds of contexts:

1. *sentence* context: the context words are all the toponyms within the same sentence;

2. *paragraph* context: all toponyms in the same paragraph of the word to disambiguate;

3. *document* context: all toponyms contained in the document are used as context.

Most WSD methods use a context window of a fixed size (e.g. two words, four words,

etc.). In the case of a geographical context composed only by toponyms, it is difficult to find more than two or three geographical terms in a sentence, and setting a larger context size would be useless. Therefore, a variable context size was used instead. The average sizes obtained by taking into account the above context types are displayed in Table 4.5.

Table 4.5: Average context size depending on context type.

context type	avg. context size
sentence	2.09
paragraph	2.92
document	9.73

It can be observed that there is a small difference between the use of sentence and paragraph, whereas the context size when using the entire document is more than 3 times the one obtained by taking into account the paragraph. In Tables 4.6, 4.7 and 4.8 are summarised the results obtained by the Conceptual Density disambiguator and the enhanced Lesk for each context type. In the tables, CD-1 indicates the CD disambiguator, CD-0 a variant that improves coverage by assigning a density 0 to all the sub-hierarchies composed by a single synset (in Formula 4.2 these sub-hierarchies would obtain 1 as weight); *Enh.Lesk* refers to the method by Banerjee and Pedersen (2002).

The obtained results show that the CD-based method is very precise when the smallest context is used, but there are many cases in which the context is empty and, therefore, it is impossible to calculate the CD. On the other hand, as one would expect, when the largest context is used coverage and recall increase, but precision drops below the most frequent baseline. However, we observed that 100% coverage cannot be achieved by CD due to some issues with the structure of WordNet. In fact, there are some 'critical' situations where CD cannot be computed, even when a context is present. This occurs when the same place name can refer to a place and another one it contains: for instance, '*New York*' is used to refer both to the city and the state it is contained in (i.e., its holonym). The result is that two senses fall within the same subhierarchy, thus not allowing to assign an unique sense to '*New York*'.

Nevertheless, even with this problem, the CD-based methods obtain a greater coverage than the enhanced Lesk method. This is due to the fact that few overlaps can be found in the glosses, because the context is composed exclusively of toponyms (for instance, the gloss of "city", the hypernym of "Cambridge" is : "`a large and densely`

`populated urban area; may include several independent administrative districts; ‘‘Ancient Troy was a great city"` – this means that an overlap will be found only if 'Troy' is in the context). Moreover, the greater is the context, the higher is the probability to obtain the same overlaps for different senses, with the consequence that the coverage drops. By knowing the number of monosemous (that is, with only one referent) toponym in GeoSemCor (501) we are able to calculate the minimum coverage that a system can obtain (41.4%), close to the value obtained with the enhanced lesk and document context (45.9%). This explains also the correlation of high precision with low coverage, due to the monosemous toponyms.

4.3 Map-based Toponym Disambiguation

In the previous section, it was shown how the structured information of the WordNet ontology can be used to effectively disambiguate toponyms. In this section, a Map-based method will be introduced. This method, inspired by the method of Smith and Crane (2001), takes advantage from Geo-WordNet to disambiguate toponyms using their coordinates, comparing the distance of the candidate referents to the centroid of the context locations. The main differences are that in Smith and Crane (2001) the context size is fixed and the centroid is calculated using only unambiguous or already disambiguated toponyms. In this version, all possible referents are used and the context size depends from the number of toponyms contained in a sentence, paragraph or document.

The algorithm is as follows: start with an ambiguous toponym t and the toponyms in the context C, $c_i \in C, 0 \leq i < n$, where n is the context size. The context is composed by the toponyms occurring in the same document, paragraph or sentence (depending on the setup of the experiment) of t. Let us call $t_0, t_1, \ldots t_k$ the locations that can be assigned to the toponym t. The map-based disambiguation algorithm consists of the following steps:

1. Find in Geo-WordNet the coordinates of each c_i. If c_i is ambiguous, consider all its possible locations. Let us call the set of the retrieved points P_c;

2. Calculate the centroid $\hat{c} = (c_0 + c_1 + \ldots + c_n)/n$ of P_c;

3. Remove from P_c all the points being more than 2σ away from $\hat{c}$, and recalculate $\hat{c}$ over the new set of points ($\bar{P}_c$). σ is the standard deviation of the set of points;

4. Calculate the distances from $\hat{c}$ of $t_0, t_1, \ldots, t_k$;

5. Select the location t_j having minimum distance from $\hat{c}$. This location corresponds to the actual location represented by the toponym t.

For instance, let us consider the following text, extracted from the *br-d03* document in the GeoSemCor:

> *One hundred years ago there existed in **England** the Association for the Promotion of the Unity of Christendom. ... A **Birmingham** newspaper printed in a column for children an article entitled "The True Story of Guy Fawkes," ... An Anglican clergyman in **Oxford** sadly but frankly acknowledged to me that this is true. ... A notable example of this was the discussion of Christian unity by the Catholic Archbishop of **Liverpool**, Dr. Heenan...*

We have to disambiguate the toponym "Birmingham", which according to WordNet can have two possible senses (each sense in WordNet corresponds to a *synset*, set of synonyms):

1. Birmingham, Pittsburgh of the South – (the largest city in Alabama; located in northeastern Alabama)

2. Birmingham, Brummagem – (a city in central England; 2nd largest English city and an important industrial and transportation center)

The toponyms in the context are "Oxford", "Liverpool" and "England". "Oxford" is also ambiguous in WordNet, having two possible senses: "Oxford, UK" and "Oxford, Mississippi". We look for all the locations in Geo-WordNet and we find the coordinates in Table 4.9, which correspond to the points of the map in Figure 4.7.

The resulting centroid is $\hat{c} = (47.7552, -23.4841)$; the distances of all the locations from this point are shown in Table 4.10. The standard deviation σ is 38.9258. There are no locations more distant than $2\sigma = 77,8516$ from the centroid, therefore no point is removed from the context.

Finally, "Birmingham (UK)" is selected, because it is nearer to the centroid $\hat{c}$ than "Birmingham, Alabama".

4.3.1 Evaluation

The experiments were carried out on the GeoSemCor corpus (Buscaldi and Rosso (2008a)), using the context divisions introduced in the previous Section, with the same average context sizes shown in Table 4.5. For the above example the context was extracted from the entire document.

Table 4.6: Results obtained using sentence as context.

system	precision	recall	coverage	F-measure
CD-1	94.7%	56.7%	59.9%	.709
CD-0	92.2%	78.9%	85.6%	**0.850**
Enh. Lesk	96.2%	53.2%	55.3%	0.685

Table 4.7: Results obtained using paragraph as context.

system	precision	recall	coverage	F-measure
CD-1	94.0%	63.9%	68.0%	0.761
CD-0	91.7%	76.4%	83.4%	0.833
Enh. Lesk	95.9%	53.9%	56.2%	0.689

Table 4.8: Results obtained using document as context.

system	precision	recall	coverage	F-measure
CD-1	92.2%	74.2%	80.4%	0.822
CD-0	89.9%	77.5%	86.2%	0.832
Enh. Lesk	99.2%	45.6%	45.9%	0.625

Table 4.9: Geo-WordNet coordinates (decimal format) for all the toponyms of the example.

	lat	lon
Birmingham (UK)	52.4797	−1.8975
Birmingham, Alabama	33.5247	−86.8128

Context locations		
	lat	lon
Oxford (UK)	51.7519	−1.2578
Oxford, Mississippi	34.3598	−89.5262
Liverpool	53.4092	−2.9855
England	51.5	−0.1667

Figure 4.7: "Birmingham"s in the world, together with context locations: "Oxford", "England", "Liverpool", according to WordNet data, and position of the context centroid.

Table 4.10: Distances from the context centroid $\hat{c}$.

location	distance from centroid (degrees)
Oxford (UK)	22.5828
Oxford, Mississippi	67.3870
Liverpool	21.2639
England	23.6162
Birmingham (UK)	**22.2381**
Birmingham, Alabama	64.9079

The results can be found in Table 4.11. Results were compared to the CD disambiguator introduced in the previous section. We also considered a map-based algorithm that does not remove from the context all the points farther than 2σ from the context centroid (i.e., does not perform step 3 of the algorithm). The results obtained with this algorithm are indicated in the Table with Map-2σ.

The results show that CD-based methods are very precise when the smallest context is used. On the other hand, for the map-based method holds the following rule: the greater the context, the better the results. Filtering with 2σ does not affect results when the context is extracted at sentence or paragraph level. The best result, in terms of F-measure, is obtained with the enhanced coverage CD method and sentence-level context.

Table 4.11: Obtained results with. p: precision, r: recall, c: coverage, F: F-measure. Map-2σ refers to the map-based algorithm previously described, and Map is the algorithm without the filtering of points farther than 2σ from the context centroid.

context	system	p	r	c	F
Sentence	CD-1	94.7%	56.7%	59.9%	0.709
	CD-0	92.2%	78.9%	85.6%	**0.850**
	Map	83.2%	27.8%	33.5%	0.417
	Map-2σ	83.2%	27.8%	33.5%	0.417
Paragraph	CD-1	94.0%	63.9%	68.0%	0.761
	CD-0	91.7%	76.4%	83.4%	0.833
	Map	84.0%	41.6%	49.6%	0.557
	Map-2σ	84.0%	41.6%	49.6%	0.557
Document	CD-1	92.2%	74.2%	80.4%	0.822
	CD-0	89.9%	77.5%	86.2%	0.832
	Map	87.9%	70.2%	79.9%	0.781
	Map-2σ	86.5%	69.2%	79.9%	0.768

From these results we can deduce that the map-based method needs more information (intended as context size) than the WordNet based method in order to obtain the same performance. However, both methods are outperformed by the first sense baseline that obtains an F-measure of .942. This may indicate that GeoSemCor is excessively biased towards the first sense. It is a well-known fact that human annotations, taken as a gold standard, are biased in favor of the first WordNet sense, which corresponds to the most frequent (Fernández-Amorós et al. (2001)).

4.4 Disambiguating Toponyms in News: a Case Study[1]

Given a news story with some toponyms in it, draw their position on a map. This is the typical application for which Toponym Disambiguation is required. This seemingly simple setup hides a series of design issues: which level of detail is required? What is the source of news stories? Is it a local news source? Which toponym resource to use? Which TD method to use? The answers to most of these questions depends on the news source. In this case study, the work was carried out on a static news collection constituted by the articles of the "L'Adige" newspaper, from 2002 to 2006. The target audience of this newspaper is constituted mainly by the population of the city of Trento, in Northern Italy, and its province. The news stories are classified in 11 sections; some are thematically closed, such as "sport" or "international", while other sections are dedicated to important places in the province: "Riva del Garda", "Rovereto", for instance.

The toponyms we extracted from this collection using *EntityPRO*, a Support Vector Machine-based tool, part of a broader suite named TextPRO, that obtained 82.1% in precision over Italian named entities Pianta and Zanoli (2007). EntityPRO may labels toponyms using one of the following labels: *GPE* (Geo-Political Entities) or *LOC* (LOCations). According to the ACE guidelines Lin (2008), "GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people. Location (LOC) entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations". The precision of EntityPRO over GPE and LOC entities has been estimated, respectively, in 84.8% and 77.8% in the EvalITA-2007[2] exercise. In the collection there are $70,025$ entities labelled as GPE or LOC, with a majority of them (58.9%) occurring only once. In the data, names of countries and cities were labelled with GPE, whereas LOC was used to label everything that can be considered a place, including *street names*. The presence of this kind of toponyms automatically determines the detail level of the resource to be used at the highest level.

As can be seen in Figure 4.8, toponyms follow a zipfian distribution, independently from the section they belong to. This is not particularly surprising, since the toponyms in the collection represent a corpus of natural language, for which Zipf law holds ("in

[1]The work presented in this section was carried out during a three months stage at the FBK-IRST under the supervision of Bernardo Magnini. Part of this section has been published as Buscaldi and Magnini (2010).

[2]`http://evalita.fbk.eu/2007/index.html`

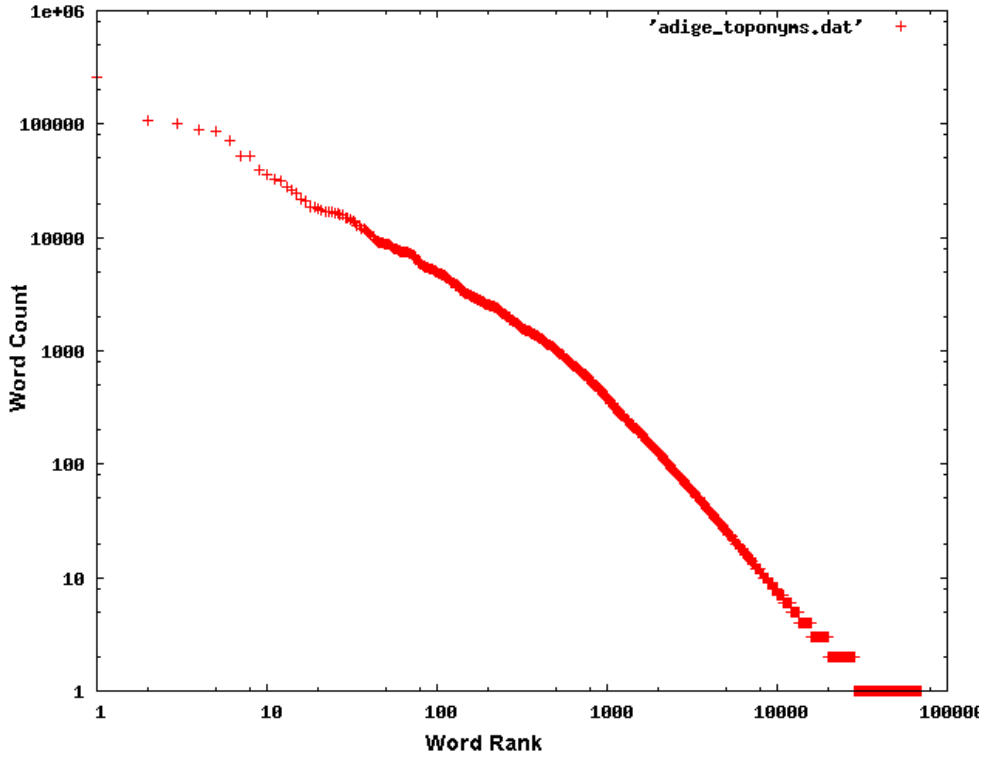


Figure 4.8: Toponyms frequency in the news collection, sorted by frequency rank. Log scale on both axes.

any large enough text, the frequency ranks of wordforms or lemmas are inversely proportional to the corresponding frequencies" Zipf (1949)). We can also observe that the set of most frequent toponyms change depending on the section of the newspaper being examined (see Table 4.12). Only 4 of the most frequent toponyms in the "international" section are included in the 10 most frequent toponyms in the whole collection, and if we look just at the articles contained in the local "Riva del Garda" section, only 2 of the most frequent toponyms are also the most frequent in the whole collection. "Trento" is the only frequent toponym that appears in all lists.

Table 4.12: Frequencies of the 10 most frequent toponyms, calculated in the whole collection ("all") and in two sections of the collection ("international" and "Riva del Garda").

all		international		Riva del Garda	
toponym	frequency	toponym	frequency	toponym	frequency
Trento	260,863	Roma	32,547	Arco	25,256
provincia	109,212	Italia	19,923	Riva	21,031
Trentino	99,555	Milano	9,978	provincia	6,899
Rovereto	88,995	Iraq	9,010	Dro	6,265
Italia	86,468	USA	8,833	Trento	6,251
Roma	70,843	Trento	8,269	comune	5,733
Bolzano	52,652	Europa	7,616	Riva del Garda	5,448
comune	52,015	Israele	4,908	Rovereto	4,241
Arco	39,214	Stati Uniti	4,667	Torbole	3,873
Pergine	35,961	Trentino	4,643	Garda	3,840

In order to build a resource providing a mapping from place names to their actual geographic coordinates, the Geonames gazetteer alone cannot be used, since this resource do not cover street names, which count for 9.26% of the total number of toponyms in the collection. The adopted solution was to build a repository of possible referents by integrating the data in the Geonames gazetteer with those obtained by querying the Google maps API geocoding service[1]. For instance, this service returns 9 places corresponding to the toponym "Piazza Dante", one in Trento and the other 8 in other cities in Italy (see Figure 4.9). The results of Google API are influenced by the region (typically the country) from which the request is sent. For example, searches for "San Francisco" may return different results if sent from a domain within the United States than one sent from Spain. In the example in Figure 4.9 there are some places

[1] `http://maps.google.com/maps/geo`

missing (for instance, piazza Dante in Genova), since the query was sent from Trento. A problem with street names is that they are particularly ambiguous, especially if the

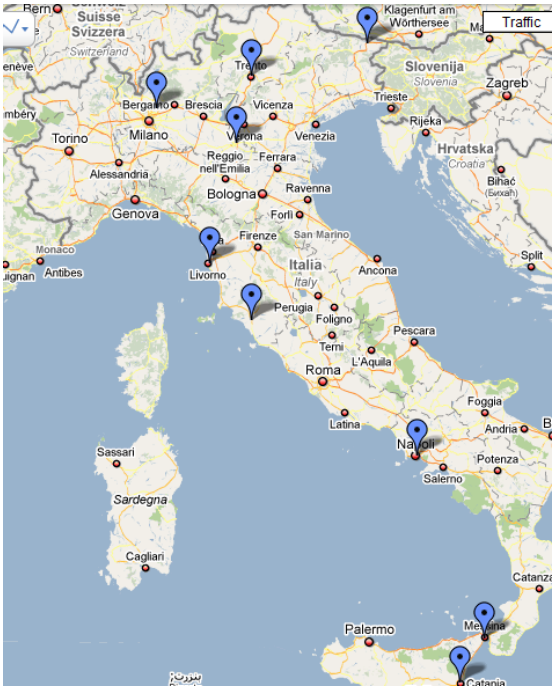


Figure 4.9: Places corresponding to "Piazza Dante", according to the Google geocoding service (retrieved Nov. 26 2009).

name of the street indicates the city pointed by the axis of the road: for instance, there is a "via Brescia" both in Mantova and Cremona, in both cases pointing towards the city of Brescia. Another common problem occurs when a street crosses different municipalities while keeping the same name. Some problems were detected during the use of the Google geocoding service, in particular with undesired automatic spelling corrections (such as "Ravina", near Trento, that is converted to "Ravenna", in the Emilia Romagna region) and with some toponyms that are spelled differently in the database used by the API and by the local inhabitants (for instance, "Piazza Fiera" was not recognised by the geocoding service, which indicated it with the name "Piazza *di* Fiera"). These errors were left unaltered in the final sense repository.

Due to the usage limitations of the Google maps geocoding service, the size of the sense repository had to be limited in order to obtain enough coverage in a reasonable time. Therefore, we decided to include only the toponyms that appeared at least 2 times in the news collection. The result was a repository containing 13,324 unique toponyms and 62,408 possible referents. This corresponds to 4.68 referents per toponym, a degree

of ambiguity considerably higher if compared to other resources used in the toponym disambiguation task, as can be seen in Table 4.13. The higher degree of ambiguity is

Table 4.13: Average ambiguity for resources typically used in the toponym disambiguation task.

Resource	Unique names	Referents	ambiguity
Wikipedia (Geo)	180,086	264,288	1.47
Geonames	2,954,695	3,988,360	1.35
WordNet2.0	2,069	2,188	1.06

due to the introduction of street names and "partial" toponyms such as "provincia" (province) or "comune" (community). Usually these names are used to avoid repetitions if the text previously contains another (complete) reference to the same place, such as in the case "provincia di Trento", or "comune di Arco", or when the context is not ambiguous.

Once the resource has been fixed, it is possible to study how ambiguity is distributed with respect to frequency. Let define the probability of finding an ambiguous toponym at frequency F by means of Formula 4.3.

$$P(F) = \frac{|T_{amb_F}|}{|T_F|} \tag{4.3}$$

Where $f(t)$ is the frequency of toponym t, T is the set of toponyms with frequency $\leq$ F: $T_F = \{t|f(t) \leq F\}$ and T_{amb_F} is the set of ambiguous toponyms with frequency $\leq$ F, i.e. $T_{amb_F} = \{t|f(t) \leq F \wedge s(t) > 1\}$, with $s(t)$ indicating the number of senses for toponym t.

In Figure 4.10 is plotted $P(F)$ for the toponyms in the collection, taking into account all the toponyms, only street names and all toponyms except street names. As can be seen from the figure, less frequent toponyms are particularly ambiguous: the probability of a toponym with frequency $f(t) \leq 100$ of being ambiguous is between 0.87 and 0.96 in all cases, while the probability of a toponym with frequency $1,000 < f(t) \leq 100,000$ of being ambiguous is between 0.69 and 0.61. It is notable that street names are more ambiguous than other terms: their overall probability of being ambiguous is 0.83, compared to 0.58 of all other kind of toponyms.

In the case of common words, the opposite phenomenon is usually observed: the most frequent words (such as "have", "be") are also the most ambiguous ones. The reason of this behaviour is that the more a word is frequent, the more are the chances it could appear in different contexts. Toponyms are used somehow in a different way:

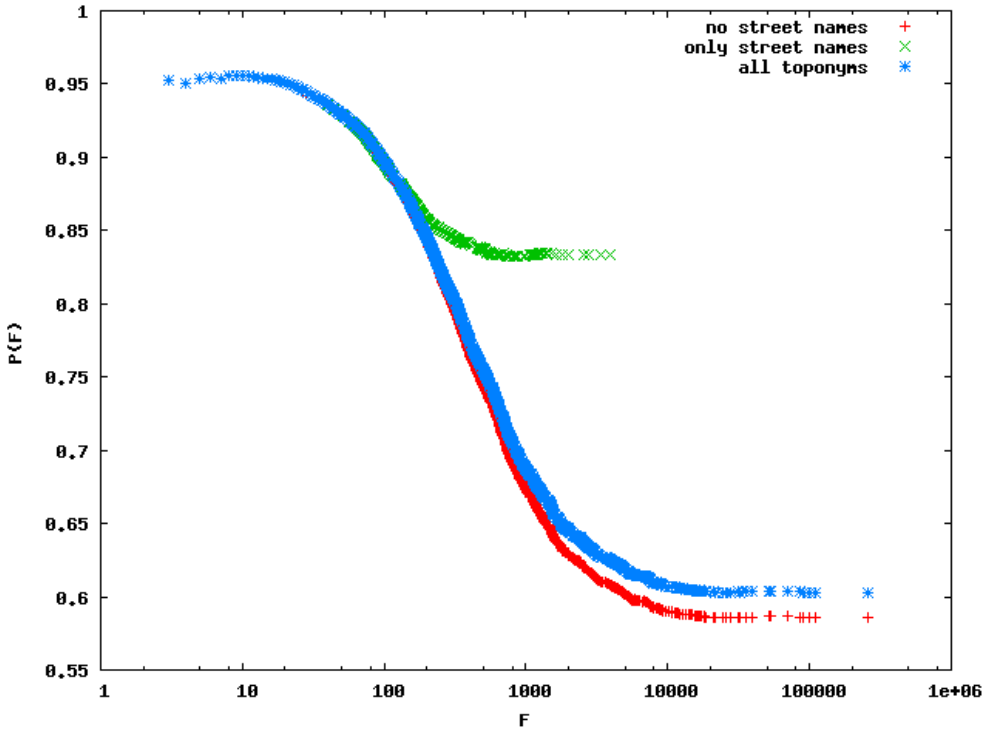


Figure 4.10: Correlation between toponym frequency and ambiguity, taking into account only street names, all toponyms, and all toponyms except street names (*no street names*). Log scale applied to x-axis.

frequent toponyms usually refer to well-known location and have a definite meaning, although used in different contexts.

The spatial distribution of toponyms in the collection with respect to the "source" of the news collection follows the "Steinberg" hypothesis as described by Overell (2009). Since "L'Adige" is based in Trento, we counted how many toponyms are found within a certain range from the center of the city of Trento (see Figure 4.11). It can be observed that the majority of place names are used to reference places within 400 km of distance from Trento.

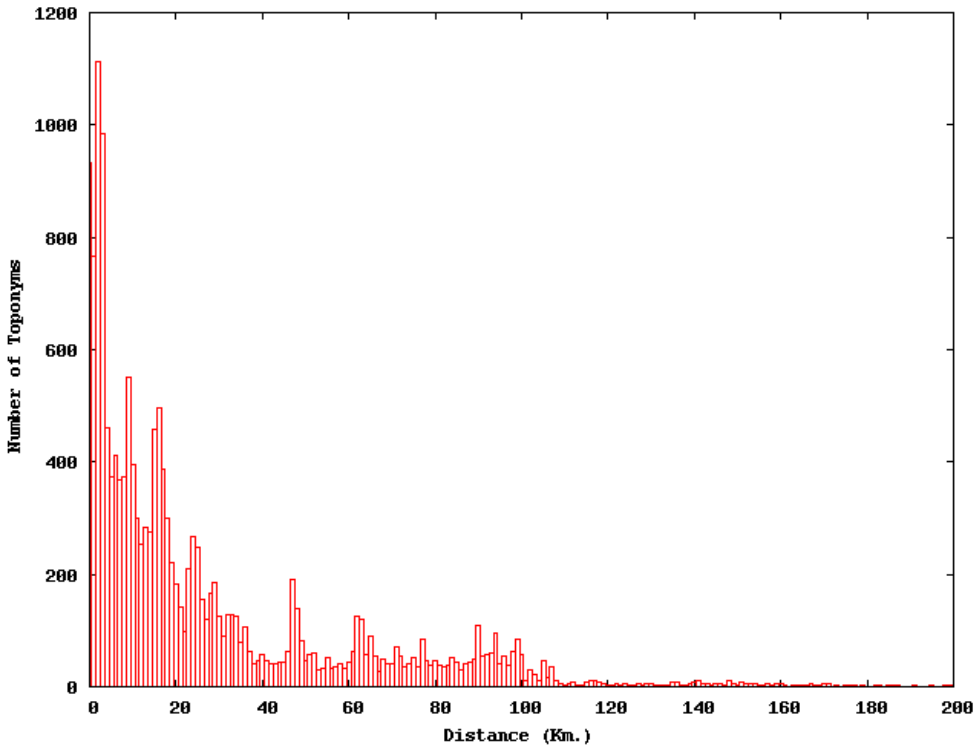


Figure 4.11: Number of toponyms found at different distances from Trento. Distances are expressed in km divided by 10.

Both knowledge-based methods and machine learning methods were not applicable to the document collection. In the first case, it was not possible to discriminate places at an administrative level lower than province, since it is the lowest administrative level provided by the Geonames gazetteer. For instance, it is possible to distinguish "via Brescia" in Mantova from "via Brescia" in Cremona (they are in two different provinces), but it is not possible to distinguish "via Mantova" in Trento from "via Mantova" in Arco, because they are in the same province. Google does actually provide

data at municipality level, but they were incompatible for merging them with those from the Geonames gazetteer. In the case of machine learning, we discarded this possibility because we had no availability of a large enough quantity of labelled data.

Therefore, the adopted solution was to improve the map-based disambiguation method described in Section 4.3, by taking into account the relation between places and distance from Trento observed in Figure 4.11, and the frequency of toponyms in the collection. The first kind of knowledge was included by adding to the context of the toponym to be resolved the place related to the news source: "Trento" for the general collection, "Riva del Garda" for the Riva section, "Rovereto" for the related section and so on. The base context for each toponym is composed by every other toponym that can be found in the same document. The size of this context window is not fixed: the number of toponyms in the context depends on the toponyms contained in the same document of the toponym to be disambiguated. From Table 4.4 and Figure 4.10 we can assume that toponyms that are frequently seen in news may be considered as not ambiguous and they could be used to specify the position of ambiguous toponyms located nearby in the text. In other words, we can say that frequent place names have a higher resolving power than place names with low frequency. Finally, we considered that word distance in text is key to solve some ambiguities: usually, in text, people writes a disambiguating place just besides the ambiguous toponyms (e.g. Cambridge, Massachusetts).

The resulting improved map-based algorithm is as follows:

1. Identify the next ambiguous toponym t with senses $S = (s_1, \ldots, s_n)$;
2. Find all toponyms t_c in context;
3. Add to the context all senses $C = (c_1, \ldots, c_m)$ of the toponyms in context (if a context toponym has been already disambiguated, add to C only that sense);
4. $\forall c_i \in C, \forall s_j \in S$ calculate the map distance $d_M(c_i, s_j)$ and text distance $d_T(c_i, s_j)$;
5. Combine frequency count ($F(c_i)$) with distances in order to calculate, for all s_j: $F_i(s_j) = \sum_{c_i \in C} \frac{F(c_i)}{(d_M(c_i,s_j) \cdot d_T(c_i,s_j))^2}$;
6. Resolve t by assigning it the sense $s = \arg_{s_j \in S} \max F_i(s_j)$.
7. Move to next toponym; if there are no more toponyms: stop.

Text distance was calculated using the number of word separating the context toponym from t. Map distance is the great-circle distance calculated using formula 3.1. It

could be noted that the part $\frac{F(c_i)}{(d_M(c_i,s_j)}$ of the weighting formula resembles the Newton's gravitation law, where the mass of a body has been replaced by the frequency of a toponym. Therefore, we can say that the formula represents a kind of "attraction" between toponyms, where most frequent toponyms have a higher "attraction" power.

4.4.1 Results

If we take into account that TextPRO identified the toponyms and labelled them with their position in the document, greatly simplifying step 1,2 and the calculation of text distance, the complexity of the algorithm is in $O(n^2 \cdot m)$, where n is the number of toponyms and m the number of senses (or possible referents). Given that the most ambiguous toponym in the database has 32 senses, we can rewrite the complexity in terms only of the number of toponyms as $O(n^3)$. Therefore, the evaluation was carried out only on a small test set and not on the entire document collection. $1,042$ entities of type GPE/LOC were labelled with the right referent, selected among the ones contained in the repository. This test collection was intended to be used to estimate the accuracy of the disambiguation method. In order to understand the relevance of the obtained results, they were compared to the results obtained by assigning to the ambiguous toponyms the referent with minimum distance from the context toponyms (that is, without taking into account neither the frequency nor the text distance) and to the results obtained without adding the context toponyms related to the news source. The $1,042$ toponyms were extracted from a set of 150 randomly selected documents.

In Table 4.14 we show the result obtained using the proposed method, compared to the results obtained with the baseline method and a version of the proposed method that did not use text distance. In the table, *complete* is used to indicate the method that includes text distance, map distance, frequency and local context; $map + freq + local$ indicates the method that do not use text distance; $map + local$ is the method that uses only local context and map distance.

Table 4.14: Results obtained over the "L'Adige" test set composed of $1,042$ ambiguous toponyms.

method	precision	recall	F-measure
complete	88.43%	88.34%	0.884
map+freq+local	88.81%	88.73%	0.888
map+local	79.36%	79.28%	0.793
baseline (only map)	78.97%	78.90%	0.789

The difference between recall and precision is due to the fact that the methods were able to deal with $1,038$ toponyms instead of the complete set of $1,042$ toponyms, because it was not possible to disambiguate 4 toponyms for the lack of context toponyms in the respective documents. The average context size was 6.96 toponyms per document, with a maximum and a minimum of 40 and 0 context toponyms in a document, respectively.

Chapter 5

Toponym Disambiguation in GIR

Lexical ambiguity and its relationship to IR has been object of many studies in the past decade. One of the most debated issues has been whether Word Sense Disambiguation could be useful to IR or not. Mark Sanderson thoroughly investigated the impact of WSD on IR. In Sanderson (1994, 2000), he experimented with pseudo-words (artificially created ambiguous words), demonstrating that when the introduced ambiguity is disambiguated with an accuracy of 75% (25% error), the effectiveness is actually worse than if the collection is left undisambiguated. He argued that only high accuracy (above 90%) in WSD could allow to obtain performance benefits and showed also that the use of disambiguation was useful only in the case of short queries, due to the lack of context. Later, Gonzalo et al. (1998) carried out some IR experiments on the SemCor corpus, finding that error rates below 30% produce better results than standard word indexing. More recently, according to this prediction, Stokoe et al. (2003) were able to obtain increased precision in IR, using a disambiguator with a WSD accuracy of 62.1%. In their conclusions, they affirm that the benefits of using WSD in IR may be present within certain types of retrieval, or in specific retrieval scenarios. GIR may constitute such a retrieval scenario, given that assigning a wrong referent to a toponym may alter significantly the results of a given query (e.g. returning results referring to "Cambridge, MA" when we were searching for results related to "Cambridge, UK").

Some research work on the the effects of various NLP errors on GIR performance has been carried out by Stokes et al. (2008). Their experimental setup used the Zettair[1] search engine with an expanded index, adding hierarchical-based geo-terms into the index as if they were "words", a technique for which it is not necessary to introduce spatial data structures. For example, they represented "Melbourne, Victoria" in the

[1] `http://www.seg.rmit.edu.au/zettair/`

index with the term "`@OC-Australia-Victoria-Melbourne`" (OC means "Oceania"). In their work, they studied the effects of NERC and toponym resolution errors over a subset of 302 manually annotated documents from the GeoCLEF collection. Their experiments showed that low NERC recall has a greater impact on retrieval effectiveness than low NERC precision does, and that statistically significant decreases in MAP scores occurred when disambiguation accuracy is reduced from 80% to 40%. However, the custom character and small size of the collection do not allow to generalize the results.

5.1 The GeoWorSE GIR System

This system is the development of a series of GIR systems that were designed in the UPV to compete in the GeoCLEF task. The first GIR system presented at GeoCLEF 2005 consisted in a simple Lucene adaptation where the input query was expanded with synonyms and meronyms of the geographical terms included in the query, using WordNet as a resource (Buscaldi et al. (2006c)). For instance, in query `GC-02`: "Vegetables exporter in Europe", Europe would be expanded to the list of countries in Europe according to WordNet. This method did not prove particularly successful and was replaced by a system that used index terms expansion, in a similar way to the approach described by Stokes et al. (2008). The evolution of this system is the *GeoWorSE* GIR System that was used in the following experiments. The core of GeoWorSE is constituted by the Lucene open source search engine. Named Entity Recognition and classification is carried out by the Stanford NER system based on Conditional Random Fields Finkel et al. (2005).

During the indexing phase, the documents are examined in order to find location names (*toponym*) by means of the Stanford NER system. When a toponym is found, the disambiguator determines the correct reference for the toponym. Then, a geographical resource (WordNet or Geonames) is examined in order to find holonyms (recursively) and synonyms of the toponym. The retrieved holonyms and synonyms are put in another separate index (*expanded* index), together with the original toponym. For instance, consider the following text from the document `GH950630-000000` in the Glasgow Herald 95 collection:

> ...The *British* captain may be seen only once more here, at next month's world championship trials in **Birmingham**, where all athletes must compete to win selection for *Gothenburg*...

Let us suppose that the system is working using WordNet as a geographical resource.

Birmingham is found in WordNet both as "*Birmingham, Pittsburgh of the South (the largest city in Alabama; located in northeastern Alabama)*", and "*Birmingham, Brummagem (a city in central England; 2nd largest English city and an important industrial and transportation center)*". "Gothenburg" is found only as "*Goteborg, Goeteborg, Gothenburg (a port in southwestern Sweden; second largest city in Sweden)*". Let us suppose that the disambiguator correctly identifies "Birmingham" with the English referent, then its holonyms are *England*, *United Kingdom*, *Europe* and their synonyms. All these words are added to the *expanded* index for "Birmingham". In the case of "Gothenburg" we obtain *Sweden* and *Europe* as holonyms, the original Swedish name of Gothenburg (*Goteborg*) and the alternate spelling "Goetenborg" as synonyms. These words are also added to the expanded index, such that the index terms corresponding to the above paragraph contained in the expanded index are: *Birmingham*, *Brummagem*, *England*, *United Kingdom*, *Europe*,*Gothenburg*, *Goteborg*, *Goeteborg*, *Sweden*.

Then, a modified Lucene indexer adds to the *geo* index the toponym coordinates (retrieved from Geo-WordNet); finally, all document terms are stored in the *text* index. In Figure 5.1 we show the architecture of the indexing module.

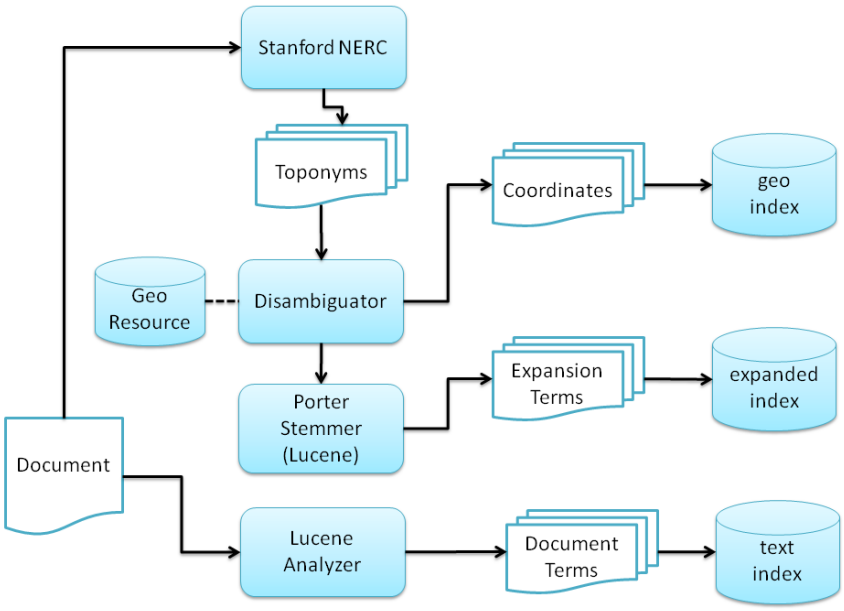


Figure 5.1: Diagram of the Indexing module

The *text* and expanded indices are used during the search phase; the *geo* index is not used explicitly for search, since its purpose is to store the coordinates of the

toponyms contained in the documents. The information contained in this index is used for ranking with Geographically Adjusted Ranking (see Subsection 5.1.1).

The architecture of the search module is shown in Figure 5.2.

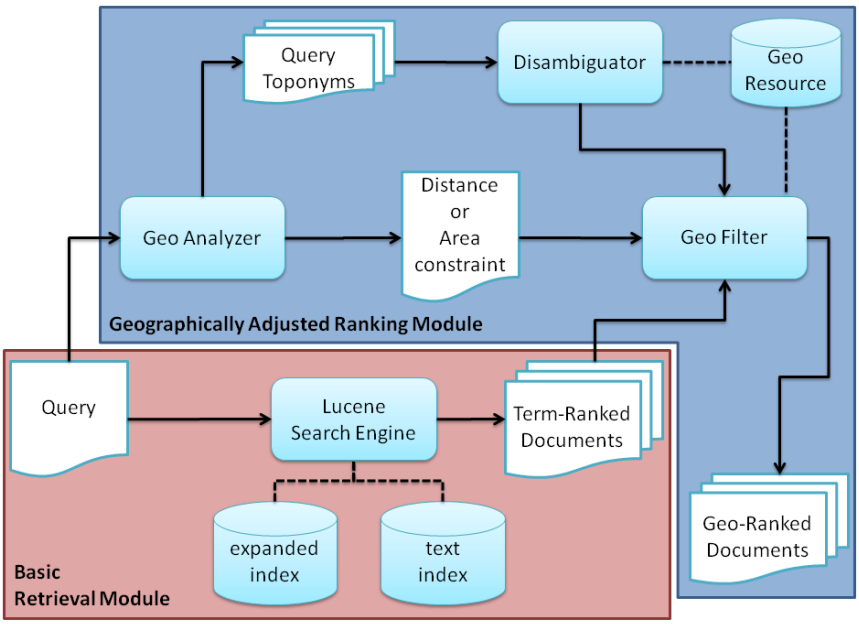


Figure 5.2: Diagram of the Search module

The topic text is searched by Lucene in the text index. All the toponyms are extracted by the Stanford NER and searched for by Lucene in the *expanded* index with a weight 0.25 with respect to the content terms. This value has been selected on the basis of the results obtained in GeoCLEF 2007 with different weights for toponyms, shown in Table 5.1. The results were calculated using the two default GeoCLEF run settings: only Title and Description and "All Fields" (see Section 2.1.4 or Appendix B for examples of GeoCLEF topics).

The result of the search is a list of documents ranked using the $tf \cdot idf$ weighting scheme, as implemented in Lucene.

5.1.1 Geographically Adjusted Ranking

Geographically Adjusted Ranking (GAR) is an optional ranking mode used to modify the final ranking of the documents by taking into account the coordinates of the places named in the documents. In this mode, at search time, the toponyms found in the query

Table 5.1: MAP and Recall obtained on GeoCLEF 2007 topics, varying the weight assigned to toponyms.

Title and Description runs		
weight	MAP	Recall
0.00	0.226	0.886
0.25	0.239	0.888
0.50	0.239	0.886
0.75	0.231	0.877
"All Fields" runs		
0.00	0.247	0.903
0.25	0.263	0.926
0.50	0.256	0.915

are passed to the GeoAnalyzer, which creates a geographical constraint that is used to re-rank the document list. The GeoAnalyzer may return two types of geographical constraints:

- a *distance* constraint, corresponding to a point in the map: the documents that contain locations closer to this point will be ranked higher;
- an *area* constraint, correspoinding to a polygon in the map: the documents that contain locations included in the polygon will be ranked higher;

For instance, in topic $10.2452/58 - GC$ there is a distance constraint: "Travel problems at major airports near to London". Topic $10.2452/76 - GC$ contains an area constraint: "Riots in South American prisons". The GeoAnalyzer determines the area using WordNet meronyms: *South America* is expanded to its meronyms: *Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Uruguay, Venezuela*. The area is obtained by calculating the convex hull of the points associated to the meronyms using the Graham algorithm Graham (1972).

The topic narrative allows to increase the precision of the considered area, since the toponyms in the narrative are also expanded to their meronyms (when possible). Figure 5.3 shows the convex hulls of the points corresponding to the meronyms of "South America", using only topic and description (left) or all the fields, including narrative (right).

The objective of the GeoFilter module is to re-rank the documents retrieved by Lucene, according to geographical information. If the constraint extracted from the

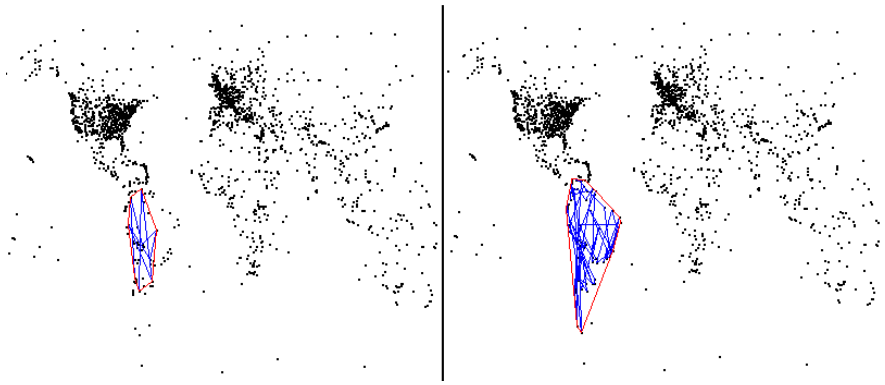


Figure 5.3: Areas corresponding to "South America" for topic 10.2452/76 – *GC*, calculated as the convex hull (in red) of the points (connected by blue lines) extracted by means of the WordNet meronymy relationship. On the left, the result using only topic and description; on the right, also the narrative has been included. Black dots represents the locations contained in Geo-WordNet.

topic is a *distance* constraint, the weights of the documents are modified according to the following formula:

$$w(doc) = w_L(doc) * (1 + \exp(-\min_{p \in P} d(q,p))) \tag{5.1}$$

Where w_L is the weight returned by Lucene for the document *doc*, P is the set of points contained in the document, and q is the point extracted from the topic.

If the constraint extracted from the topic is an *area* constraint, the weights of the documents are modified according to Formula 5.2:

$$w(doc) = w_L(doc) * \left(1 + \frac{|P_q|}{|P|}\right) \tag{5.2}$$

where P_q is the set of points in the document that are contained in the area extracted from the topic.

5.2 Toponym Disambiguation vs. no Toponym Disambiguation

The first question to be answered is whether Toponym Disambiguation allows to obtain better results that just adding to the index all the candidate referents. In order to answer this question, the GeoCLEF collection was indexed in four different configurations with the GeoWorSE system:

Table 5.2: Statistics of GeoCLEF topics.

conf.	avg. query length	# toponyms	# amb. toponyms
Title Only	5.74	90	25
Title Desc	17.96	132	42
All Fields	52.46	538	135

- *GeoWN*: Geo-WordNet and the Conceptual Density were used as gazetteer and disambiguation method,respectively, for the disambiguation of toponyms in the collection;
- *GeoWN noTD*: Geo-WordNet was used as gazetteer but no disambiguation was carried out;
- *Geonames*: Geonames was used as gazetteer and the map-based method described in Section 4.3 was used for toponym disambiguation;
- *Geonames noTD*: Geonames was used as gazetteer,no disambiguation;

The test set was composed by the 100 topics from GeoCLEF $2005-2008$ (see Appendix B for details). When TD was used, the index was expanded only with the holonyms related to the disambiguated toponym; when no TD was used, the index was expanded with all the holonyms that were associated to the toponym in the gazetter. For instance, when indexing "Aberdeen" using Geo-WordNet in the "no TD" configuration, the following holonyms were added to the index: "Scotland", "Washington, Evergreen State, WA", "South Dakota, Coyote State, Mount Rushmore State, SD", "Maryland, Old Line State, Free State, MD". Figure 5.4 and Figure 5.5 show the Precision/Recall graphs obtained using Geonames or Geo-WordNet, respectively, compared to the "no TD" configuration. Results are presented for the two basic CLEF configurations ("Title and Description" and "All Fields") and the "Title Only" configuration, where only the topic title is used. Although the evaluation in the "Title Only" configuration is not standard in CLEF competitions, it is interesting to study these results because this configuration reflects the way people usually queries search engines: Baeza-Yates et al. (2007) highlighted that the average length of queries submitted to the Yahoo! search engine between 2005 and 2006 was of only 2.5 words. In Table 5.2 it can be noticed how the average length of the queries is considerably greater in modes different from "Title Only".

In Figure 5.6 are displayed the average MAP obtained by the systems in the different run configurations.

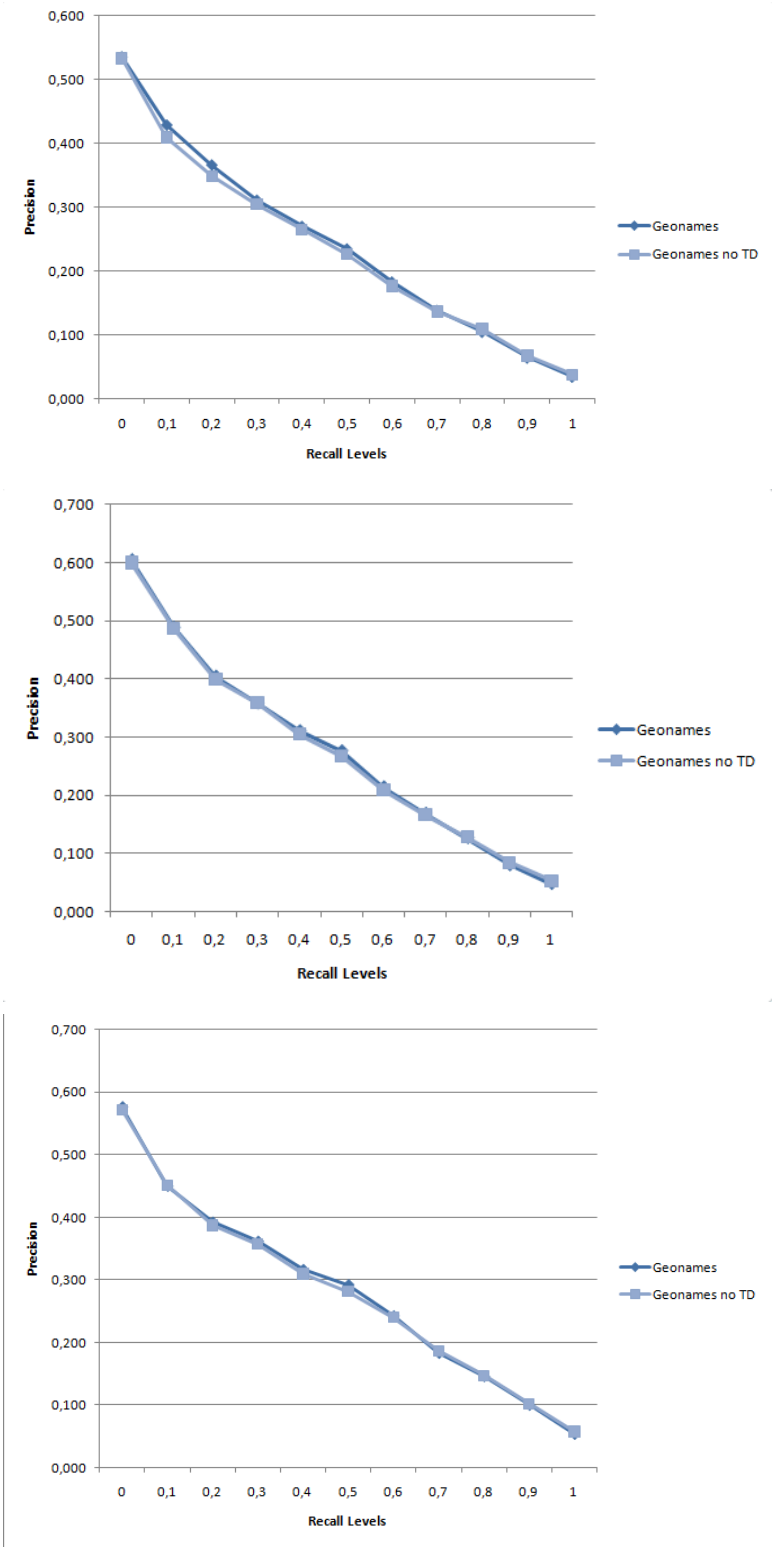


Figure 5.4: Comparison of the Precision/Recall graphs obtained using Toponym Disambiguation or not, using Geonames as a resource. From top to bottom: "Title Only", "Title and Description" and "All Fields" runs.

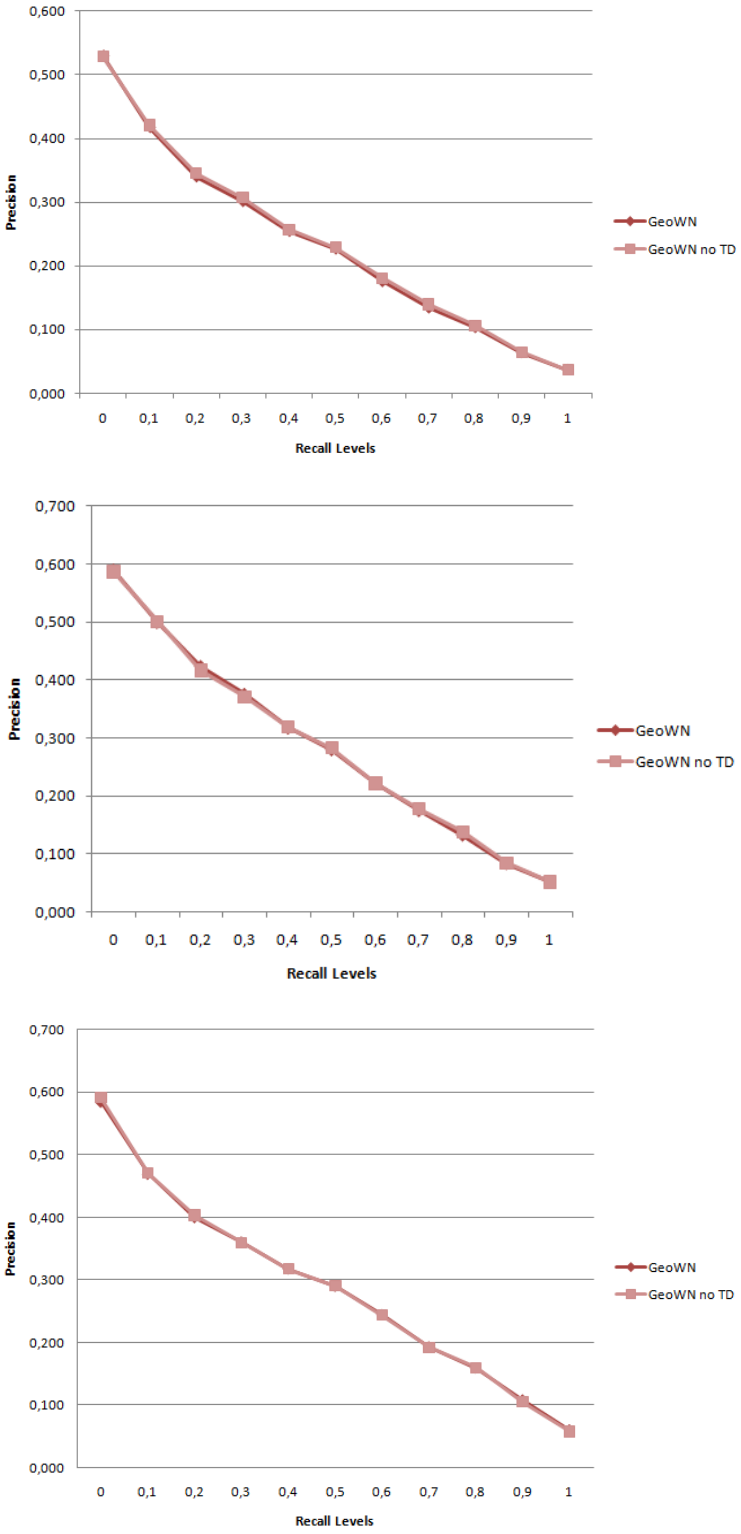


Figure 5.5: Comparison of the Precision/Recall graphs obtained using Toponym Disambiguation or not, using Geo-WordNet as a resource. From top to bottom: "Title Only", "Title and Description" and "All Fields" runs.

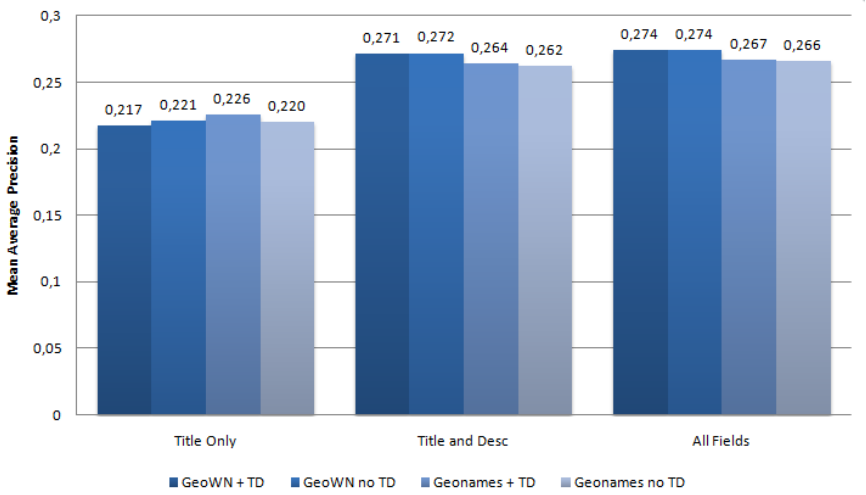


Figure 5.6: Average MAP using Toponym Disambiguation or not.

5.2.1 Analysis

From the results it can be observed that Toponym Disambiguation was useful only in Geonames runs (Figure 5.4), especially in the "Title Only" configuration, while in the Geo-WordNet runs, not only it did not allow any improvement, but resulted in a decrease in precision, especially for the "Title Only" configuration. The only statistical significant difference is between the Geonames and the Geo-WordNet "Title Only" runs. An analysis of the results topic-by-topic showed that the greatest difference between the Geonames and Geonames noTD runs was observed in topic `84-GC`: "*Bombings in Northern Ireland*". In Figure 5.7 are shown the differences in MAP for each topic between the disambiguated and not disambiguated runs using Geonames.

A detailed analysis of the results obtained for topic `84-GC` showed that one of the relevant documents, `GH950819-000075` ("Three petrol bomb attacks in Northern Ireland") was ranked in third position by the system using TD and was not present in the top 10 results returned by the "no TD" system. In the document left undisambiguated, "*Belfast*" was expanded to "*Belfast*", "Saint Thomas", "Queensland", "Missouri", "Northern Ireland", "California", "Limpopo", "Tennessee", "Natal", "Maryland", "Zimbabwe", "Ohio", "Mpumalanga", "Washington", "Virginia", "Prince Edward Island", "Ontario", "New York", "North Carolina", "Georgia", "Maine", "Pennsylvania", "Nebraska", "Arkansas". In the disambiguated document, "Northern Ireland" was correctly selected as the only holonym for Belfast.

On the other hand, in topic `GC-010` ("Flooding in Holland and Germany") the re-

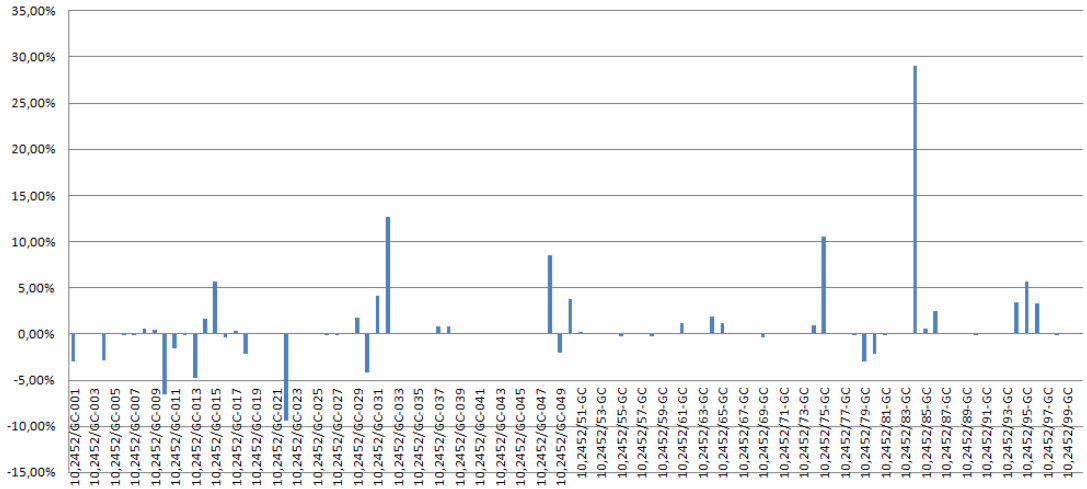


Figure 5.7: Difference, topic-by-topic, in MAP between the Geonames and Geonames "no TD" runs.

sults obtained by the system that did not use disambiguation were better thanks to document `GH950201-000116` ("Floods sweep across northern Europe"): this document was retrieved at the 6^{th} place by this system and was not included in the top 10 documents retrieved by the TD-based system. The reason in this case was that the toponym "*Zeeland*" was incorrectly disambiguated and assigned to its referent in "North Brabant" (it is the name of a small village in this region of the Netherlands), instead of the correct Zeeland province in the "Netherlands", whose "Holland" synonym was included in the index created without disambiguation.

It should be noted that in Geo-WordNet there is only one referent for "Belfast" and no referent for "Zeeland" (although there is one referent for "Zealand", corresponding to the region in Denmark). However, Geo-WordNet results were better in "Title and Description" and "All Fields" runs, as it can be seen in Figure 5.6. The reason for this is that in longer queries, such the ones derived from the use of the additional topic fields, the geographical context is better defined, if more toponyms are added to those included in the "Title Only" runs; on the other hand, if more non-geographical terms are added, the importance of toponyms is scaled down.

Correct disambiguation is not always ensuring that the results can be improved: in topic `GC-022`, "Restored buildings in Southern Scotland", the relevant document `GH950902-000127` ("stonework restoration at Culzean Castle") is ranked only in 9^{th} position by the system that uses toponym disambiguation, while the system that does not use disambiguation retrieves it in the first position. This difference is determined

by the fact that the documents ranked $1-8$ by the system using TD are all referring to places in Scotland, and they were expanded only to this holonym. The system that do not use TD ranked them lower, because their toponyms were expanded to all the referents, and, according to the $tf \cdot idf$ weighting, "Scotland" obtained a lower weight because it was not the only term in the expansion.

Therefore, disambiguation seems to help to improve retrieval accuracy only in the case of short queries and if the detail of the geographic resource used is high. Even in these cases, disambiguation errors can actually improve the results if they alter the weighting of a non-relevant document such that it is ranked lower.

5.3 Retrieving with Geographically Adjusted Ranking

In this section we compare the results obtained by the systems using Geographically Adjusted Ranking to those obtained without using GAR. In Figure 5.8 and Figure 5.9 are presented the Precision/Recall graphs obtained for GAR runs, using both disambiguation or not, compared to the base runs, with the system that used TD and standard term-based ranking.

From the comparison of Figure 5.8 and Figure 5.9, and the average MAP results shown in Figure 5.10, it can be observed how the Geo-WordNet-based system does not obtain any benefit from the Geographically Adjusted Ranking, except in the "no TD", title only run. On the other hand, the following results can be observed when Geonames is used as toponym resource (Figure 5.8):

- The use of GAR allows to improve MAP if disambiguation is applied (Geonames + GAR);
- Applying GAR to the system that do not use TD results in lower MAP.

These results strengthen the previous findings that the detail of the resource used is crucial to obtain improvements by means of Toponym Disambiguation.

5.4 Retrieving with Artificial Ambiguity

The objective of this section is to study the relation between the number of errors in TD and the accuracy in IR. In order to carry out this study, it was necessary to work on a disambiguated collection. The experiments were carried out by introducing errors on 10%, 20%, 30%, 40%, 50% and 60% of the monosemic (i.e., with only one meaning) toponyms instances contained in the CLIR-WSD collection. An error is

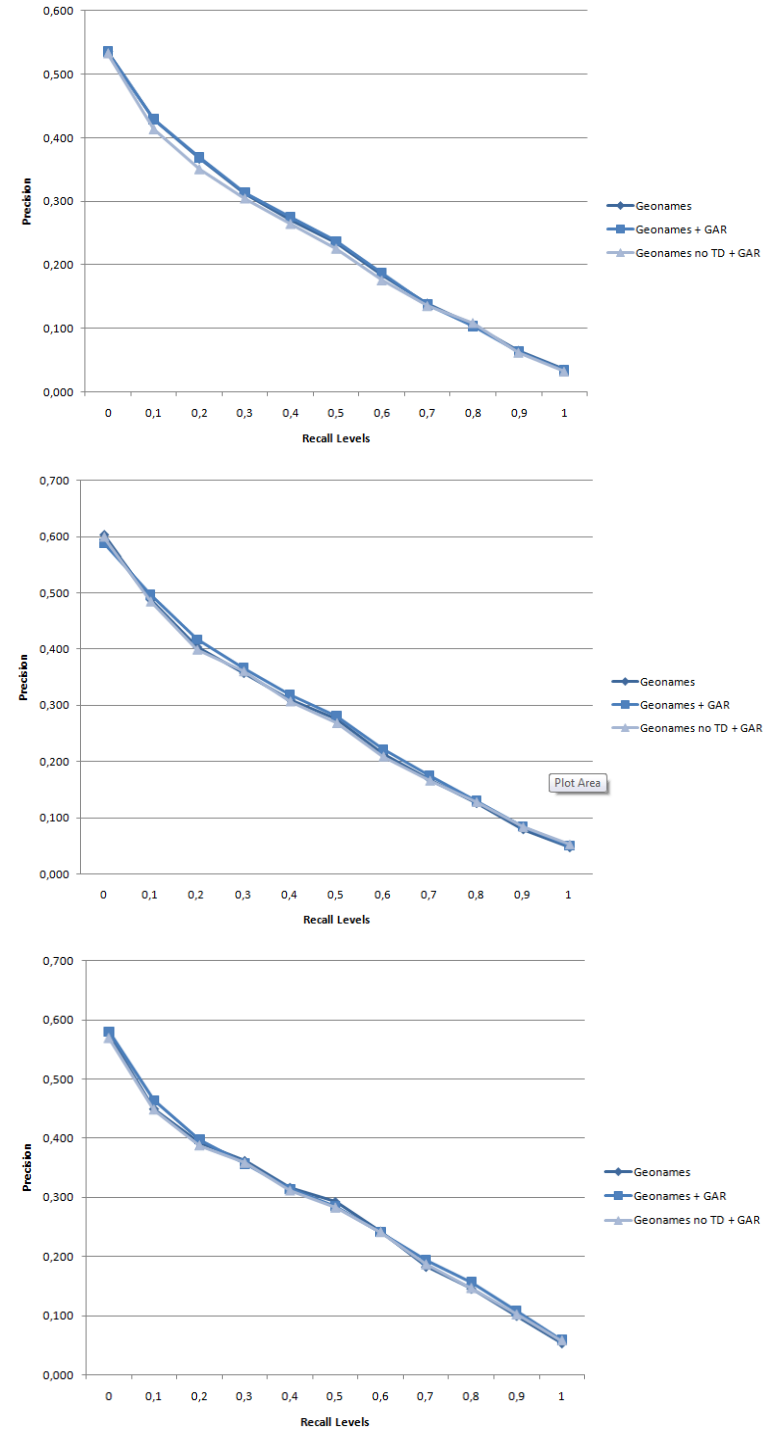


Figure 5.8: Comparison of the Precision/Recall graphs obtained using Geographically Adjusted Ranking or not, using Geonames. From top to bottom: "Title Only", "Title and Description" and "All Fields" runs.

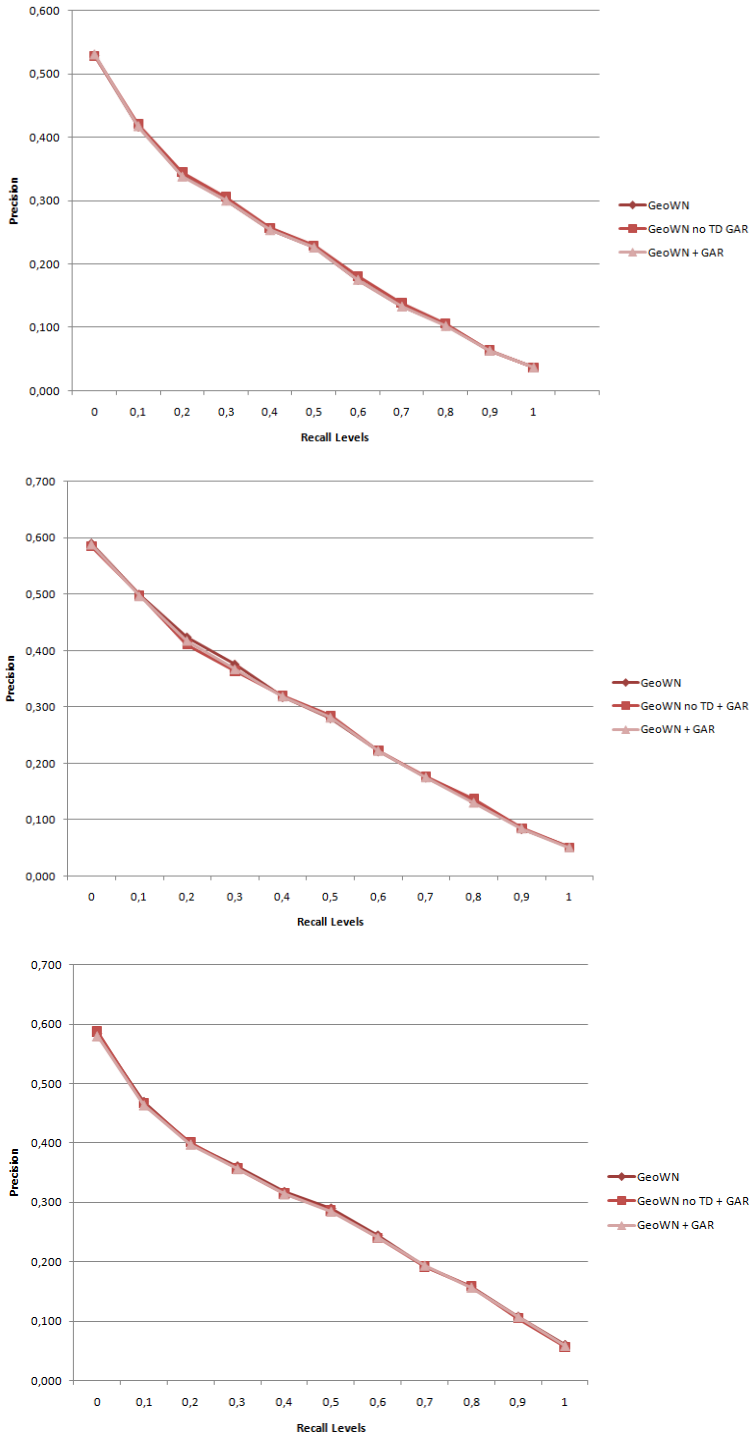


Figure 5.9: Comparison of the Precision/Recall graphs obtained using Geographically Adjusted Ranking or not, using Geo-WordNet. From top to bottom: "Title Only", "Title and Description" and "All Fields" runs.

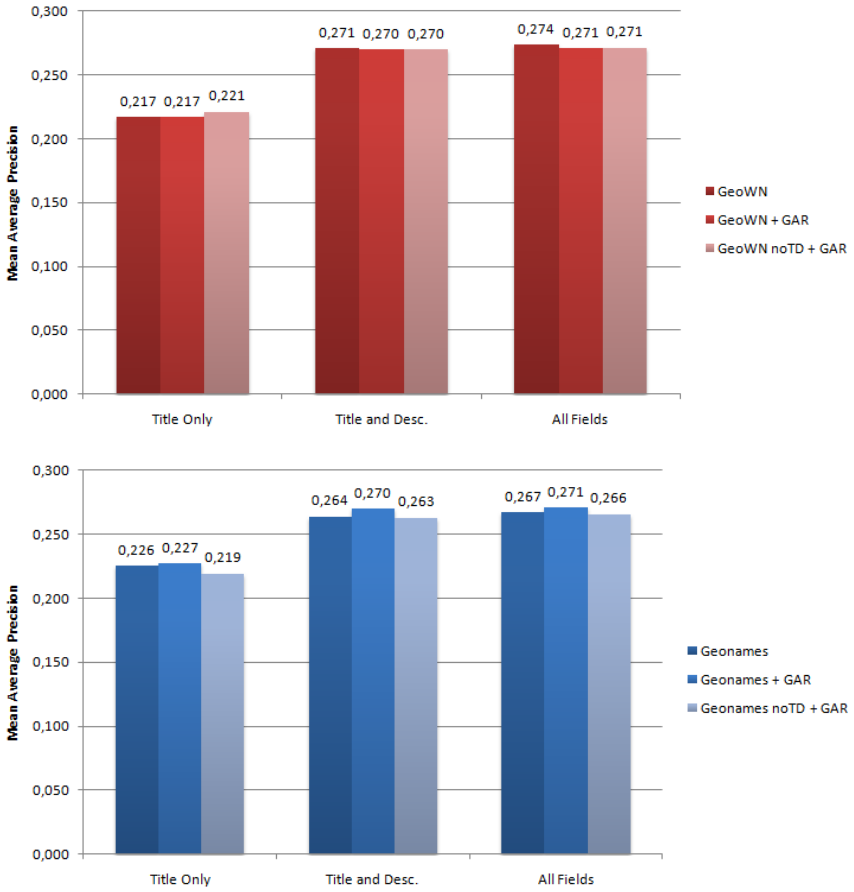


Figure 5.10: Comparison of MAP obtained using Geographically Adjusted Ranking or not. Top: Geo-WordNet, Bottom: Geonames.

introduced by changing the holonym from the one related to the sense assigned in the collection to a "sister term" of the holonym itself. "Sister term" in this case is used to indicate a toponym that shares the same holonym with another toponym (i.e., they are meronyms of the same synset). For instance, to introduce an error in "Paris, France", the holonym "France" can be changed to "Italy", because they are both meronyms of "Europe". Introducing errors on the monosemic toponyms allows to ensure that the errors are "real" errors. In fact, the disambiguation accuracy over toponyms in the CLIR-WSD collection is not perfect (100%). Changing the holonym on an incorrectly disambiguated toponym may result in actually correcting en existing error, instead than introducing a new one. The developers were not able to give a figure of the overall accuracy on the collection, however the accuracy of the method reported in Agirre and Lopez de Lacalle (2007) is of 68.9% in precision and recall over the Senseval-3 All-Words task, and 54.4% in the Semeval-1 All-Words task. These numbers seem particularly low, but they are in line with the accuracy levels obtained by the best systems in WSD competitions. We expect a similar accuracy level over toponyms.

Figure 5.11 shows the Precision/Recall graphs obtained in the various run configurations ("Title Only", "Title and Description", "All Fields"), and at the above defined TD error levels. Figure 5.12 shows the MAP for each experiment, grouped by run configuration. Errors were generated randomly, independently from the errors generated at the previous levels. In other words, the disambiguation errors in the 10% collection were not preserved into the 20% collection: the increment of the number of errors does not constitute an increment over previous errors.

The differences in MAP between the runs in the same configuration are not statistically meaningful (t-test 44% in the best case), however it is noteworthy that the MAP obtained at 0% error level is always higher than the MAP obtained at 60% error level. One of the problems with the CLIR-WSD collection is that, despite the precautions taken by introducing errors only on monosemic toponyms, some of the introduced errors could actually fix an error. This is the case in which WordNet does not contain referents that are used in text. For instance, toponym "Valencia" was labelled as *Valencia/Spain/Europe* in CLIR-WSD, although most of the "Valencias" named in the documents of collection (especially the Los Angeles Times collection) are representing a suburb of Los Angeles in California. Therefore, a toponym that is monosemic for WordNet may not be actually monosemic, and the random selection of a different holonym may end in picking the right holonym. Another problem is that changing the holonym may not alter the result of queries that cover an area at continent level: "Springfield" in WordNet 1.6 has only one possible holonym, "Illinois". Changing the holonym to

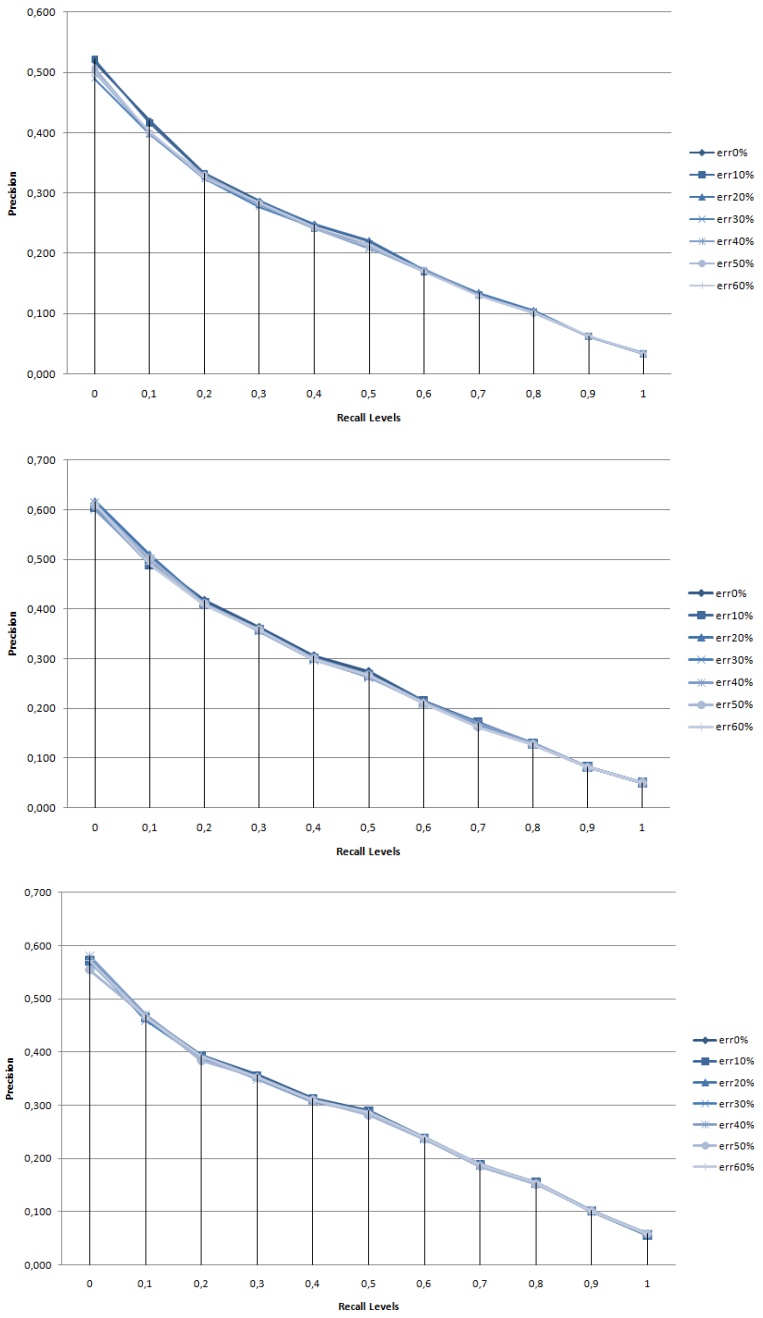


Figure 5.11: Comparison of the Precision/Recall graphs obtained using different TD error levels. From above to bottom: "Title Only", "Title and Description", "All Fields" runs.

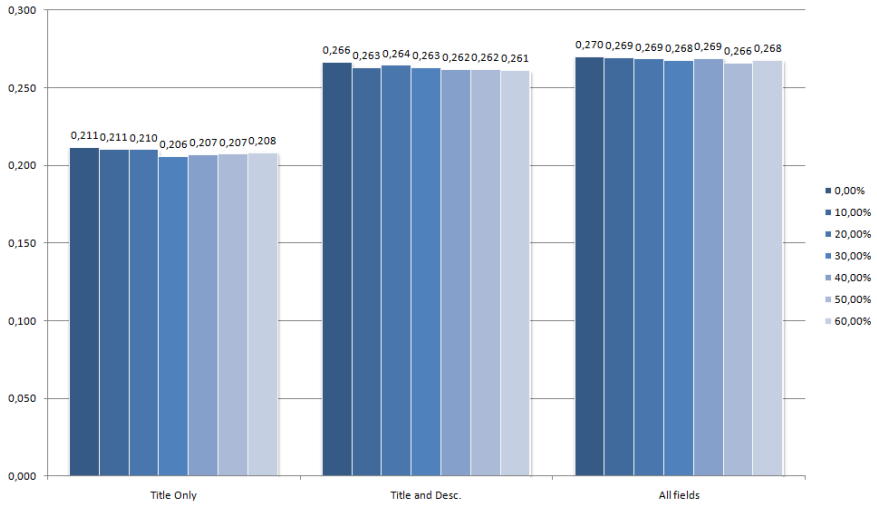


Figure 5.12: Average MAP at different artificial toponym disambiguation error levels.

"Massachusetts", for instance, does not change the scope to outside the United States and would not affect the results for a query about the United States or North America.

5.5 Final Remarks

In this chapter we presented the results obtained by applying Toponym Disambiguation or not to a GIR system we developed, GeoWorSE. These results show that disambiguation is useful only if the query length is short and the resource is detailed enough, while no improvements can be observed if a resource with low detail is used, like WordNet, or queries are long enough to provide context to the system. The use of the GAR technique also proved to be effective under the same conditions. We also carried out some experiments by introducing artificial ambiguity on a GeoCLEF disambiguated collection, CLIR-WSD. The results show that no statistically significant variation in MAP is observed between a 0% and a 60% error rate.

Chapter 6

Toponym Disambiguation in QA

6.1 The SemQUASAR QA System

QUASAR (Buscaldi et al. (2009)) is a QA system that participated in CLEF-QA 2005, 2006 and 2007 (Buscaldi et al. (2006a, 2007); Gómez et al. (2005)) in Spanish, French and Italian. The participations ended with relatively good results, especially in Italian (best system in 2006 with 28.2% accuracy) and Spanish (third system in 2005 with 33.5% accuracy). In this section we present a version that was slightly modified in order to work on disambiguated documents instead of the standard text documents, using WordNet as sense repository. QUASAR was developed following the idea that, in a large enough document collection, it is possible to find an answer formulated in a similar way to the question. The architecture of most QA system that participated in the CLEF-QA tasks is similar, consisting in an analysis subsystem, which is responsible to check the type of the questions, a Passage Retrieval (PR) module, which is usually a standard IR search engine adapted to work on short documents, and an analysis module, which uses the information extracted in the analysis phase to look for the answer in the retrieved passages. The JIRS PR system constitutes the most important advance introduced by QUASAR, since it is based on n-grams similarity measures instead of classical weighting schemes that are usually based on term frequency, such as $tf \cdot idf$. Most QA systems are based on IR methods that have been adapted to work on passages instead of the whole documents (Magnini et al. (2001); Neumann and Sacaleanu (2004); Vicedo (2000)). The main problems with these QA systems derive from the use of methods which are adaptations of classical document retrieval systems, which are not specifically oriented to the QA task and, therefore, do not take into account its characteristics: the style of questions is different from the style of IR

queries, and relevance models that are useful on long documents may fail when the size of documents is small, as introduced in Section 2.2. The architecture of SemQUASAR is very similar to the architecture of QUASAR and is shown in Figure 6.1.

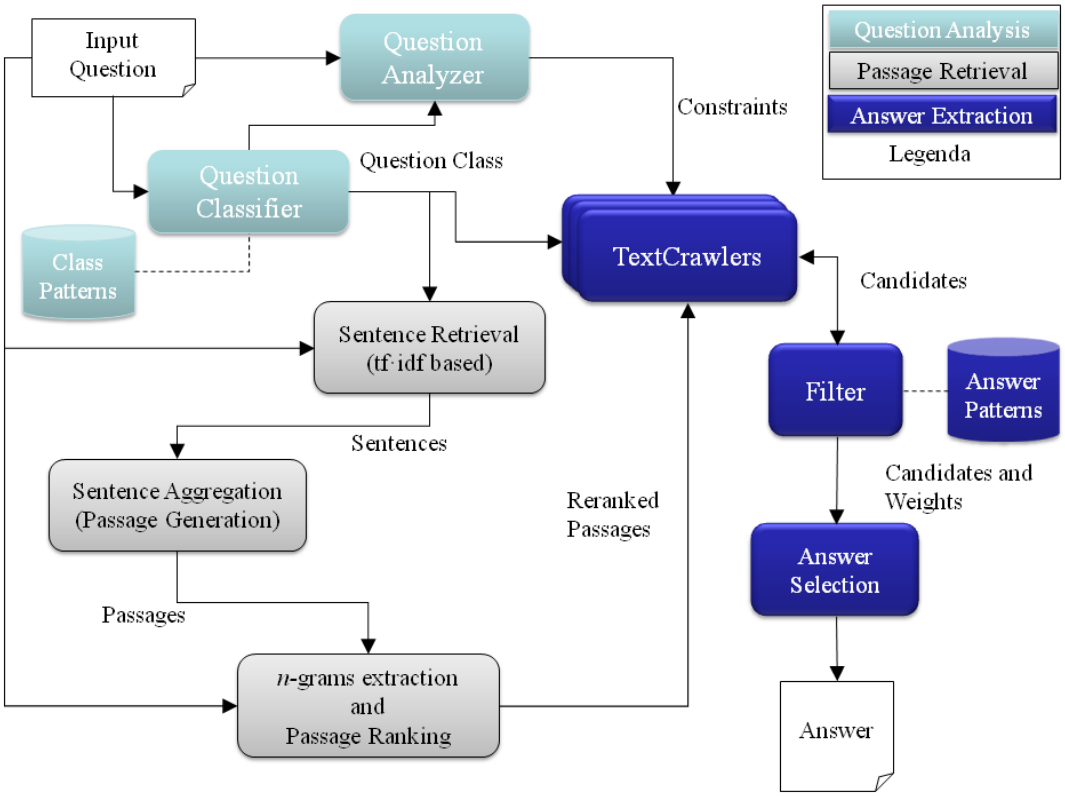


Figure 6.1: Diagram of the SemQUASAR QA system

Given a user question, this will be handed over to the *Question Analysis* module, which is composed by a *Question Analyzer* that extracts some constraints to be used in the answer extraction phase, and by a *Question Classifier* that determines the class of the input question. At the same time, the question is passed to the *Passage Retrieval* module, which generates the passages used by the *Answer Extraction* module, together with the information collected in the question analysis phase, in order to extract the final answer. In the following subsections we detail each of the modules.

6.1.1 Question Analysis Module

This module obtains both the expected answer type (or *class*) and some constraints from the question. The different answer types that can be treated by our system are shown in Table 6.1.

Table 6.1: QC pattern classification categories.

L0	L1	L2
NAME	ACRONYM	
	PERSON	
	TITLE	
	FIRSTNAME	
	LOCATION	COUNTRY
		CITY
		GEOGRAPHICAL
DEFINITION	PERSON	
	ORGANIZATION	
	OBJECT	
DATE	DAY	
	MONTH	
	YEAR	
	WEEKDAY	
QUANTITY	MONEY	
	DIMENSION	
	AGE	

Each category is defined by one or more patterns written as regular expressions. The questions that do not match any defined pattern are labeled with *OTHER*. If a question matches more than one pattern, it is assigned the label of the longest matching pattern (i.e., we consider longest patterns to be less generic than shorter ones).

The Question Analyzer has the purpose of identifying patterns that are used as constraints in the AE phase. In order to carry out this task, the set of different n-grams in which each input question can be segmented are extracted, after the removal of the initial quetsion stop-words. For instance consider the question: "*Where is the Sea World aquatic park?*", then the following n-grams are generated:

```
[Sea] [World] [aquatic] [park]
```

```
[Sea World] [aquatic] [park]
[Sea] [World aquatic] [park]
[Sea] [World] [aquatic park]
[Sea World] [aquatic park]
[Sea] [World aquatic park]
[Sea World aquatic] [park]
[Sea World aquatic park]
```

The weight for each segmentation is calculated in the following way:

$$\prod_{x \in S_q} \frac{\log 1 + N_D - \log f(x)}{\log N_D} \tag{6.1}$$

where S_q is the set of n-grams extracted from query q, $f(x)$ is the frequency of n-gram x in the collection D, and N_D is the total number of documents in the collection D.

The n-grams that compose the segmentation with the highest weight are the *contextual* constraints, which represent the information that has to be included in the retrieved passage in order to have a chance of success in extracting the correct answer.

6.1.2 The Passage Retrieval Module

The sentences containing the relevant terms are retrieved using the Lucene IR system with the default $tf \cdot idf$ weighting scheme. The query sent to the IR system includes the constraints extracted by the Question Analysis module, passed as phrase search terms. The objective of constraints is to avoid to retrieve sentences with n-grams that are not relevant to the question.

For instance, suppose the question is "*What is the capital of Croatia?*" and the extracted constraint is "capital of Croatia". Suppose that the following two sentences are contained in the document collection: "*...Tudjman, the president of Croatia, met Eltsin during his visit to Moscow, the capital of Russia...*", and "*...they discussed the situation in Zagreb, the capital of Croatia...*". Considering just the keywords would result in the same weight for both sentences, however, taking into account the constraint, only the second passage is retrieved.

The results are a list of sentences that are used to form the passages in the Sentence Aggregation module. Passages are ranked using a weighting model based on the density of question n-grams. The passages are formed by attaching to each sentence in the ranked list one or more contiguous sentences of the original document in the following way: let a document d be a sequence of n sentences $d = (s_1, \ldots, s_n)$. If a sentence s_i is retrieved by the search engine, a passage of size $m = 2k + 1$ is formed by the

concatenation of sentences $s_{(i-k)} \dots s(i+k)$. If $(i-k) < 1$, then the passage is given by the concatenation of sentences $s_1 \dots s_{(k-i+1)}$. If $(i+k) > n$, then the passage is obtained by the concatenation of sentences $s_{(i-k-n)} \dots s_n$. For instance, let us consider the following text extracted from the Glasgow Herald 95 collection (`GH950102-000011`):

> "Andrei Kuznetsov, a Russian internationalist with Italian side Les Copains, died in a road crash at the weekend. He was 28. A car being driven by Ukraine-born Kuznetsov hit a guard rail alongside a central Italian highway, police said. No other vehicle was involved. Kuznetsov's wife was slightly injured in the accident but his two children escaped unhurt."

This text contains 5 sentences. Let us suppose that the question is "*How old was Andrei Kuznetsov when he died?*"; the search engine would return the first sentence as the best one (it contains "Andrei", "Kuznetsov" and "died"). If we set the Passage Retrieval (PR) module to return passages composed by 3 sentences, it would return "*Andrei Kuznetsov, a Russian internationalist with Italian side Les Copains, died in a road crash at the weekend. He was 28. A car being driven by Ukraine-born Kuznetsov hit a guard rail alongside a central Italian highway, police said.*". If we set the PR module to return passages composed by 5 sentences or more, it would return the whole text. This example also shows a case in which the answer is not contained in the same sentence, demonstrating the usefulness of splitting the text into passages.

Gómez et al. (2007) demonstrated that almost 90% in answer coverage can be obtained with passages consisting of 3 contiguous sentences and taking into account only the first 20 passages for each question. This means that the answer can be found in the first 20 passages returned by the PR module in 90% of the cases where an answer exists, if passages are composed by 3 sentences

In order to calculate the weight of n-grams of every passage, the greatest n-gram in the passage or the associated expanded index is identified and it is assigned a weight equal to the sum of all its term weights. The weight of every term is determined by means of formula 6.2:

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \quad . \tag{6.2}$$

Where n_k is the number of sentences in which the term appears, and N is the number of sentences in the document collection. We make the assumption that stopwords occur in every sentence (i.e., $n_k = N$ for stopwords). Therefore, if the term appears once in the passage collection, its weight will be equal to 1 (the greatest weight).

6.1.3 WordNet-based Indexing

In the indexing phase (Sentence Retrieval module) two indices are created: the first one (*text*) contains all the terms of the sentence; the second one (expanded index, or *wn* index) contains all the synonyms of the disambiguated words; in the case of nouns and verbs, it contains also their hypernyms. For nouns, the holonyms (if available) are also added to the index. For instance, let us consider the following sentence from document `GH951115-000080-03`:

> Splitting the left from the Labour Party would weaken the battle for progressive policies inside the Labour Party.

The underlined words are those that have been disambiguated in the collection. For these words we can found their synonyms and related concepts in WordNet, as listed in Table 6.2.

Table 6.2: Expansion of terms of the example sentence. NA : not available (the relationship is not defined for the Part-Of-Speech of the related word).

lemma	ass. sense	synonyms	hypernyms	holonyms
split	4	separate part	move	NA
left	1	–	position place	–
Labour Party	2	labor party	political party party	–
weaken	1	–	change alter	NA
battle	1	conflict fight engagement	military action action	war warfare
progressive	2	reformist	NA	NA
policy	2	–	argumentation logical argument line of reasoning line	–

Therefore, the *wn* index will contain the following terms: *separate, part, move, position, place, labor party, political party, party, change, alter, conflict, fight, engagement,*

war, warfare, military action, action, reformist, argumentation, logical argument, line of reasoning, line.

During the search phase, the *text* and *wn* indices are both searched for question terms. The top 20 sentences are returned for each question. Passages are built from these sentences, by appending them the previous and next sentences in the collection. For instance, if the above example were a retrieved sentence, the resulting passage would be composed by the following sentences:

- `GH951115-000080-2` : "The real question is how these policies are best defeated and how the great mass of Labour voters can be won to see the need for a socialist alternative."
- `GH951115-000080-3` : "Splitting the left from the Labour Party would weaken the battle for progressive policies inside the Labour Party."
- `GH951115-000080-4` : "It would also make it easier for Tony Blair to cut the crucial links that remain with the trade-union movement."

Figure 6.2 shows the first 5 sentences returned for the question "*What is the political party of Tony Blair?*" using only the *text* index; in Figure 6.3 we show the first 5 sentences returned using also the *wn* index; it can be noted that the sentences retrieved with the expanded WordNet index are shorter than those retrieved with the basic method.

The Labour Party , under Tony Blair , is poised to achieve political power for the first time in 16 years	GH950821-000120-5
No Headline Present Political peace : A truce is agreed on the Westminster front as party leaders John Major , Paddy Ashdown , and Tony Blair celebrate VJ-Day	GH950821-000164-0
Blair puts the family centre stage LABOUR leader Tony Blair , in a further move to occupy the political centre ground , yesterday staked his claim for Labour to be the ' ' party of the family	GH950330-000184-0
Blair should beware I AM mystified by your political editor ' s comment about Tony Blair ' ' dispatching the Bennite Marxist left ' ' and rescuing the Labour Party from the ' ' false historical perspective ' ' of a ' ' Marxist intellectual analysis ' ' (March 23	GH950327-000057-0
Blair wrestles with age-old links TRADE union leaders of all political persuasions have confronted Labour leader Tony Blair with a demand that the unions retain a 50 % voice in the party they created	GH951002-000214-1

Figure 6.2: Top 5 sentences retrieved with the standard Lucene search engine.

The method was adapted to the geographical domain by adding to the *wn* index all the containing entities of every location included in the text.

6.1.4 Answer Extraction

The input of this module is constituted by the n passages returned by the PR module and the constraints (including the expected type of the answer) obtained through the

The Labour Party , under Tony Blair , is poised to achieve political power for the first time in 16 years	GH950821-000120-5
The Labour Party has been set a simple test by Tony Blair	GH950310-000026-16
No Headline Present Political peace : A truce is agreed on the Westminster front as party leaders John Major , Paddy Ashdown , and Tony Blair celebrate VJ-Day	GH950821-000164-0
' The investigation is understood to have the full support of Labour Party leader Tony Blair	GH950227-000161-5
On the eve of the Labour party conference , there were also sharp words for Tony Blair	GH951002-000228-9

Figure 6.3: Top 5 sentences retrieved with the WordNet extended index.

Question Analysis module. A *TextCrawler* is instantiated for each of the n passages with a set of patterns for the expected answer type and a pre-processed version of the passage text. The pre-processing consists in separating all the punctuation characters from the words and in stripping off the annotations (related concepts extracted from WordNet) included in the passage. It is important to keep the punctuation symbols because we observed that they usually offer important clues for the individuation of the answer (this is true especially for *definition* questions): for instance, it is more frequent to observe a passage containing "*The president of Italy, Giorgio Napolitano*" than one containing "*The president of Italy is Giorgio Napolitano*"; moreover, movie and book titles are often put between apices.

The positions of the passages in which occur the constraints are marked before passing them to the TextCrawlers. The TextCrawler begins its work by searching all the passage's substrings matching the expected answer pattern. Then a weight is assigned to each found substring s, inversely proportional to the distance of s from the constraints, if s does not include any of the constraint words.

The *Filter* module uses a knowledge base of allowed and forbidden patterns. Candidate answers which do not match with an allowed pattern or that do match with a forbidden pattern are eliminated. For instance, if the expected answer type is a geographical name (class LOCATION), the candidate answer is searched for in the Wikipedia-World database in order to check that it could correspond to a geographical name. When the Filter module rejects a candidate, the TextCrawler provide it with the next best-weighted candidate, if there is one.

Finally, when all TextCrawlers have finished their analysis of the text, the *Answer Selection* module selects the answer to be returned by the system. The final answer is selected with a strategy named "weighted voting": each vote is multiplied by the weight assigned to the candidate by the TextCrawler and for the passage weight as returned by the PR module. If no passage is retrieved for the question or no valid candidates are selected, then the system returns a NIL answer.

6.2 Experiments

We selected a set of 77 questions from the CLEF-QA 2005 and 2006 cross-lingual English-Spanish test sets. The questions are listed in Appendix C. 53 questions out of 77 (68.8%) contained an answer in the GeoCLEF document collection. The answers were checked manually in the collection, since the original CLEF-QA questions were intended to be searched for in a Spanish document collection. In Table 6.3 are shown the results obtained over this test sets with two configuration: "no WSD", meaning that the index is the index built with the system that do not use WordNet for the index expansion, while the "CLIR-WSD" index is the index expanded, where disambiguation has been carried out with the supervised method by Agirre and Lopez de Lacalle (2007) (see Section 2.2.1 for details over R, X and U measures).

Table 6.3: QA Results with SemQUASAR, using the standard index and the WordNet expanded index.

run	R	X	U	Accuracy
no WSD	9	3	0	16.98%
CLIR-WSD	7	2	0	13.21%

The results have been evaluated using the CLEF setup, detailed in Section 2.2.1. From these results it can be observed that the basic system was able to answer correctly to two question more than the WordNet-based system. The next experiment consisted in introducing errors in the disambiguated collection and checking whether accuracy changed or not with respect to the use of the CLIR-WSD expanded index. The results are showed in Table 6.4.

Table 6.4: QA Results with SemQUASAR, varying the error level in Toponym Disambiguation.

run	R	X	U	Accuracy
CLIR-WSD	7	2	0	13.21%
10% error	7	0	1	13.21%
20% error	7	0	0	13.21%
30% error	7	0	0	13.21%
40% error	7	0	0	13.21%
50% error	7	0	0	13.21%
60% error	7	0	0	13.21%

These results show that the performance in QA does not change, whatever the level of TD errors are introduced in the collection. In order to check whether this behaviour is dependent on the Answer Extraction method or not, and what is the contribution of TD on the passage retrieval module, we calculated the Mean Reciprocal Rank of the answer in the retrieved passages. In this way, $MRR = 1$ means that the right answer is contained in the passage retrieved at the first position, $MRR = 1/2$ at the second retrieved passage and so on.

Table 6.5: MRR calculated with different TD accuracy levels.

question	err.0%	err.10%	err.20%	err.30%	err.40%	err.50%	err.60%
7	0	0	0	0	0	0	0
8	0.04	0	0	0	0	0	0
9	1.00	0.04	1.00	1.00	0	0	0
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12	0.50	1.00	0.50	0.50	1.00	1.00	1.00
13	0.00	1.00	0.14	0.14	0	0	0
14	1.00	0.00	0.00	0.00	0	0	0
15	0.04	0.17	0.17	0.17	0.17	0.17	0.50
16	1.00	0.50	0.00	0.00	0.25	0.33	0.25
17	1.00	1.00	1.00	1.00	0.50	1.00	0.50
18	0.50	0.04	0.04	0.04	0.04	0.04	0.04
27	0.00	0.25	0.33	0.33	0.17	0.13	0.13
28	0.03	0.03	0.04	0.04	0.04	0.04	0.04
29	0.50	0.17	0.10	0.10	0.04	0.04	0.09
30	0.17	0.33	0.25	0.25	0.25	0.20	0.25
31	0.00	0	0	0	0	0	0
32	0.20	1.00	1.00	1.00	1.00	1.00	1.00
36	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40	0.00	0	0	0	0	0	0
41	1.00	1.00	0.50	0.50	1.00	1.00	1.00
45	0.17	0.08	0.10	0.10	0.09	0.10	0.08
46	0.00	1.00	1.00	1.00	1.00	1.00	1.00
47	0.05	0.50	0.50	0.50	0.50	0.50	0.50
48	1.00	1.00	0.50	0.50	0.33	1.00	0.33
50	0.00	0.00	0.06	0.06	0.05	0	0

Continued on Next Page...

question	err.0%	err.10%	err.20%	err.30%	err.40%	err.50%	err.60%
51	0.00	0	0	0	0	0	0
53	1.00	1.00	1.00	1.00	1.00	1.00	1.00
54	0.50	1.00	1.00	1.00	0.50	1.00	1.00
57	1.00	0.50	0.50	0.50	0.50	0.50	0.50
58	0.00	0.33	0.33	0.33	0.25	0.25	0.25
60	0.11	0.11	0.11	0.11	0.11	0.11	0.11
62	1.00	0.50	0.50	0.50	1.00	0.50	1.00
63	1.00	0.07	0.08	0.08	0.08	0.08	0.08
64	0.00	1.00	1.00	1.00	1.00	1.00	1.00
65	1.00	1.00	1.00	1.00	1.00	1.00	1.00
67	1.00	0.00	0.17	0.17	0	0	0
68	0.50	1.00	1.00	1.00	1.00	1.00	1.00
71	0.14	0.00	0.00	0.00	0.00	0.00	0.00
72	0.09	0.20	0.20	0.20	0.20	0.20	0.20
73	1.00	1.00	1.00	1.00	1.00	1.00	1.00
74	0.00	0.00	0.00	0.00	0.00	0.00	0.00
76	0.00	0.00	0.00	0.00	0.00	0.00	0.00

In Figure 6.4 it can be noted how average MRR decreases when TD errors are introduced. The decrease is statistically relevant only for the 40% error level, although the difference is due mainly to the result on question 48: "*Which country is Alexandria in?*". In the 40% error level run, a disambiguation error assigned "Low Countries" as an holonym for Sofia, Bulgaria: the effect was to raise the weight of the passage containing "Sofia" with respect to the question term "country". However, this kind of errors do not affect the final output of the complete QA system, since the Answer Extraction module is not able to find a match for "Alexandria" in the better ranked passage.

Question 48 highlights also an issue with the evaluation of the answer: both "United States" and "Egypt" would be correct answers in this case, although the original information need expressed by means of the question probably was related to the Egyptian referent. This kind of questions constitute the ideal scenario for Diversity Search, where the user becomes aware of meanings that he did not know at the moment of formulating the question.

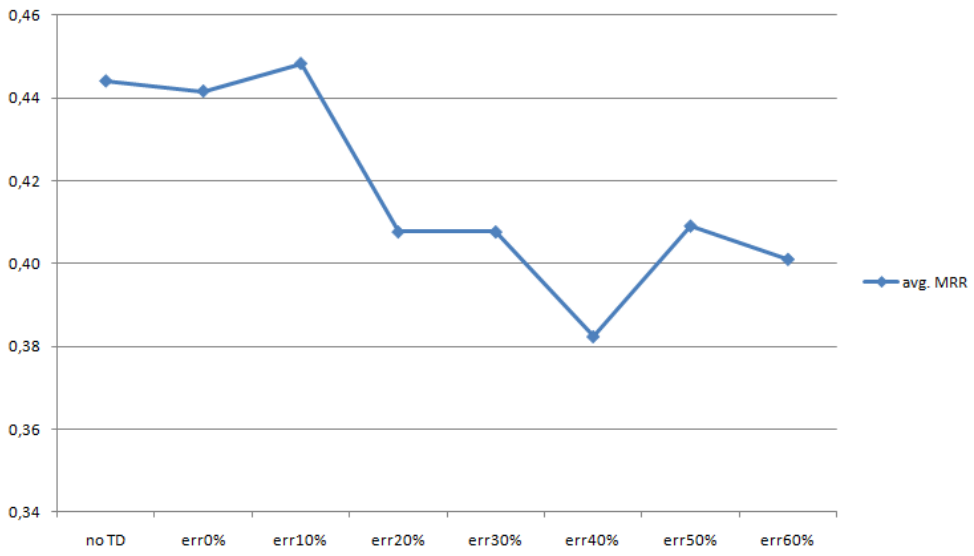


Figure 6.4: Average MRR for passage retrieval on geographical questions, with different error levels.

6.3 Analysis

The carried out experiments do not show any significant effect of Toponym Disambiguation in the Question Answering task, even with a test set composed uniquely of geographically-related questions. Moldovan et al. (2003) observed that QA systems can be affected by a great quantity of errors occurring in different modules of the system itself. In particular, wrong question classification is usually so devastating that it is not possible to answer correctly to the question even if all the other modules carry out their work without errors. Therefore, the errors that can be produced by Toponym Disambiguation have only a minor importance with respect to this kind of errors. On the other hand, even if no errors occur in the various modules of a QA system, redundancy allows to compensate the errors that may result from the incorrect disambiguation of toponyms. In other words, retrieving a passage with an error is usually not affecting the results if the system already retrieved 29 more passages that contain the right answer.

6.4 Final Remarks

In this chapter we carried out some experiments with the SemQUASAR system, which has been adapted to work on the CLIR-WSD collection. The experiments consisted in

submitting to the system a set composed of geographically-related questions extracted from the CLEF QA test set. We observed no difference in accuracy results using toponym disambiguation or not, as no difference in accuracy were observed using the collections where artificial errors were introduced. We analysed the results only from a Passage Retrieval perspective, to understand the contribution of TD to the performance of the PR module. This evaluation was carried out taking into account the MRR measure. Results indicate that average MRR decreases when TD errors are introduced, with the decrease being statistically relevant only for the 40% error level.

Chapter 7

Geographical Web Search: Geooreka!

The results obtained with GeoCLEF topics suggest that the use of term-based queries may not be the optimal method to express a geographically constrained information need. Actually, there are queries in which the terms used do not allow to clearly define a footprint. For instance, fuzzy concepts that are commonly used in geography, like "Northern" and "Southern", which could be easily introduced in databases using mathematical operations on coordinates, are often interpreted subjectively by humans. Let us consider the topic `GC-022` "Restored buildings in *Southern Scotland*": no existing gazetteer has an entry for this toponym. What does the user mean for "Southern Scotland"? Should results include places in Fife, for instance, or not? Looking at the map in Figure 7.1, one may say that the Fife region is in the Southern half of Scotland, but probably a Scotsman would not agree on this criterion. Vernacular names that define a fuzzy area are another case of toponyms that are used in queries (Schockaert and De Cock (2007); Twaroch and Jones (2010)), especially for local searches. In this case, the problem is that a name is commonly used by a group of people that knows very well some area, but it is not significant outside this group. For instance, almost everyone in Genoa (Italy) is able to say what "Ponente" (West) is: "the coastal suburbs and towns located west of the city centre". However, people living outside the region of Genoa do not know this terminology and there is no resource that maps the word into the set of places it is referring to. Therefore, two approaches can be followed to solve this issue: the first one is to build or enrich gazetteers with vernacular place names; the second one is to change the way users interact with GIR systems, such that they do not depend exclusively on place names in order to define the query footprint. I followed

this second approach in the effort of developing a web search engine (*Geooreka*[1]) that allows users to express their information needs in a graphical way, taking advantage from the Yahoo! Maps API. For instance, for the above example query, users would just select the appropriate area in the map, write the theme that they want to find information about ("Restored buildings") and the engine would do the rest. Vaid et al. (2005) showed that combining textual with spatial indexing would allow to improve geographically constrained searches in the web; in the case of Geooreka, geography is deduced from text (toponyms), since it was not feasible (due to time and physical resource issues) to geo-tag and spatially analyse every web document.

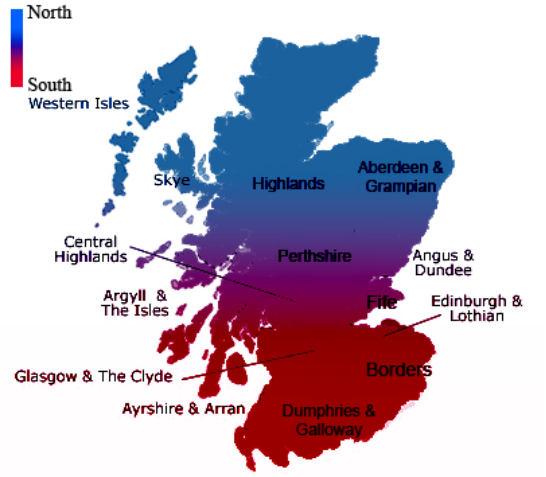


Figure 7.1: Map of Scotland with North-South gradient.

7.1 The Geooreka! Search Engine

Geooreka! (Buscaldi and Rosso (2009b)) works in the following way: the user selects an area (the query *footprint*) and write an information topic (the *theme* of the query) in a textbox. Then, all toponyms that are relevant for the map zoom level are extracted (*Toponym Selection*) from the PostGIS-enabled GeoDB database: for instance, if the map zoom level is set at "country", only country names and capital names are selected. Then, web counts and mutual information are used in order to determine which combinations theme-toponym are most relevant with respect to the information need expressed by the user (*Selection of Relevant Queries*). In order to speed-up the process, web counts are calculated using the static Google 1T Web database[2], whereas

[1] `http://www.geooreka.eu`

[2] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13`

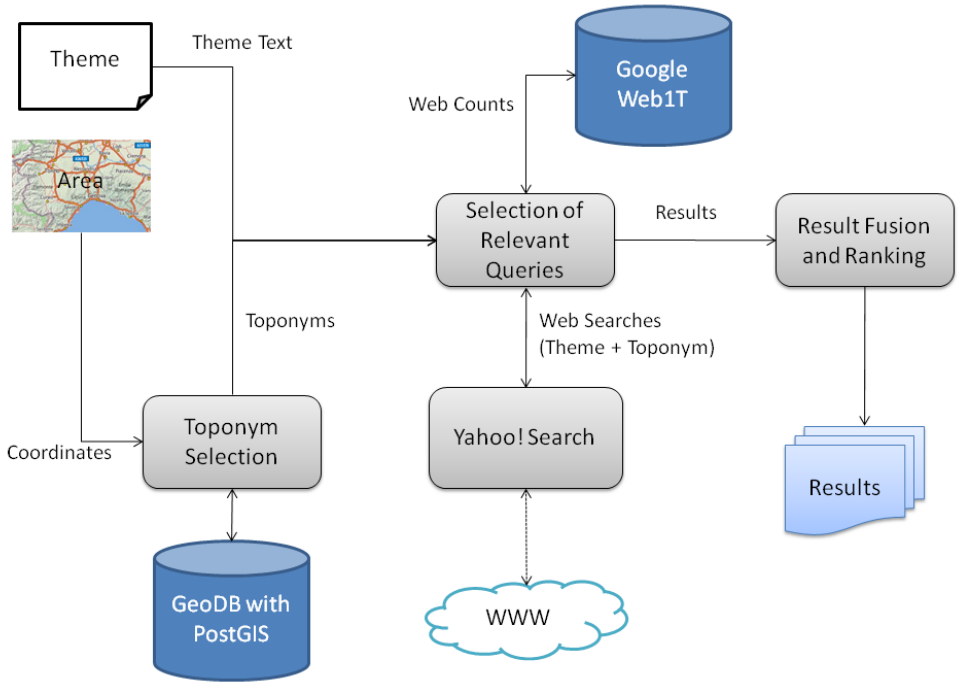


Figure 7.2: Overall architecture of the Geooreka! system.

Yahoo! Search is used to retrieve the results of the queries composed by the combination of a theme and a toponym. The final step (*Result Fusion and Ranking*) consists in the fusion of the results obtained from the best combinations and their ranking.

7.1.1 Map-based Toponym Selection

The first step in order to process the query is to select the toponyms that are relevant to the area and zoom level selected by the user. Geonames was selected as toponym repository and its data loaded into a PostgreSQL server. The choice of PostgreSQL was due to the availability of PostGIS[1], an extension to PostgreSQL that allows it to be used as a backend spatial database for Geographic Information Systems. PostGIS supports many types of geometries, such as points, polygons and lines. However, due to the fact that GNS provides just one point per place (e.g. it does not contain shapes for regions), all data in the database is associated to a *POINT* geometry. Toponyms are stored in a single table named *locations*, whose columns are detailed in Table 7.1.

Table 7.1: Details of the columns of the *locations* table.

column name	type	description
`title`	varchar	the name of the toponym
`coordinates`	PostGIS POINT	position of the toponym
`country`	varchar	name of the country the toponym belongs to
`subregion`	varchar	the name of the administrative region
`style`	varchar	the class of the toponym (using GNS features)

The selection of the toponyms in the query footprint is carried out by means of the bounding box operator (BOX3D) of PostGIS: for instance, suppose that we need to find all the places contained in a box defined by the coordinates: $(44.440N, 8.780E)$ and $(44.342N, 8.986E)$. Therefore, we have to submit to the database the following query:

```
SELECT title, AsText(coordinates), country, subregion, style
FROM locations WHERE
coordinates && SetSRID('BOX3D(8.780 44.440, 8.986 44.342)'::box3d, 4326);
```

The code '4326' indicates that we are using the WGS84 standard for the representation of geographical coordinates. The use of PostGIS allows to obtain the results efficiently, avoiding the slowness problems reported by Chen et al. (2006).

An subset of the resulting tuples of this query can be observed in Table 7.2. From

[1] `http://postgis.refractions.net/`

Table 7.2: Excerpt of the tuples returned by the Geooreka! PostGIS database after the execution of the query relative to the area delimited by $8.780E44.440N$, $8.986E44.342N$.

title	coordinates	country	subregion	style
Genova	POINT(8.95 44.4166667)	IT	Liguria	ppla
Genoa	POINT(8.95 44.4166667)	IT	Liguria	ppla
Cornigliano	POINT(8.8833333 44.4166667)	IT	Liguria	pplx
Monte Croce	POINT(8.8666667 44.4166667)	IT	Liguria	hill

the tuples in Table 7.2 we can see that GNS contains variants in different language for the toponyms (in this case *Genova*), and some of the feature codes of Geonames: *ppla* which is used to indicate that the toponym is an administrative capital, *pplx* which indicates a subdivision of a city, and *hill* that indicates a minor relief.

Feature codes are important, because, depending on the zoom level, only certain types of places are selected. In Table 7.3 are showed the filters applied at each zoom level. The greater the zoom level, the farther the viewpoint from the Earth is, and the fewer are the selected toponyms.

Table 7.3: Filters applied to toponym selection depending on zoom level.

zoom level	zone desc	applied filter
$16, 17$	world	do not use toponyms
$14, 15$	continents	continent names
13	sub-continent	states
$12, 11$	state	states, regions and capitals
10	region	as *state*, with provinces
$8, 9$	sub-region	as *region*, with all cities and physical features
$5, 6, 7$	cities	as *sub-region*, includes *pplx* features
< 5	street	all features

The selected toponyms are passed to the next module, which assembles the web queries as strings of the form +"*theme*" + "*toponym*" and verifies which ones are relevant. The quotation marks are used to carry out phrase searches instead than keyword searches. The + symbol is a standard Yahoo! operator that forces the presence of the word or phrase in the web page.

7.1.2 Selection of Relevant Queries

The key issue in the selection of the relevant queries is to obtain a relevance model that is able to select pairs theme-toponym that are most promising to satisfy the user's information need.

We assume, on the basis of the theory of probability, that the two composing parts of the queries, theme T and toponym G, are independent if their conditional probabilities are independent, i.e., $p(T|G) = p(T)$ and $p(G|T) = p(G)$, or, equivalently, their joint probability is the product of their probabilities:

$$\hat{p}(T \cap G) = p(G)p(T) \tag{7.1}$$

Where $\hat{p}(T \cap G)$ is the *expected* probability of co-occurrence of T and G in the same web page. The probabilities are calculated as the number of pages in which the term (or phrase) representing the theme or toponym appears, divided by $2,147,436,244$ which is the maximum term frequency contained in the Google Web 1T database.

Considering this model for the independence of theme and toponym, we can measure the divergence of the expected probability $\hat{p}(T \cap G)$ from the observed probability $p(T \cap G)$: the more the divergence, the more informative is the result of the query.

The Kullback-Leibler measure Kullback and Leibler (1951) is commonly used in order to determine the divergence of two probability distributions. For a discrete random variable:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{7.2}$$

where P represents the actual distribution of data and Q the expected distribution. In our approximation, we do not have a distribution but we are interested to determine the divergence point-by-point. Therefore we do not sum for all the queries. Substituting in Formula 7.2 our probabilities we obtain:

$$D_{KL}(p(T \cap G)||\hat{p}(T \cap G)) = p(T \cap G) \log \frac{p(T \cap G)}{\hat{p}(T \cap G)} \tag{7.3}$$

that is, substituting $\hat{p}$ according to Formula 7.1:

$$D_{KL}(p(T \cap G)||\hat{p}(T \cap G)) = p(T \cap G) \log \frac{p(T \cap G)}{p(T)p(G)} \tag{7.4}$$

This formula is exactly one of the formulations of the *Mutual Information* (MI) of T and G, usually denoted as $(I(T;G))$.

For instance, the frequency of "pesto" (a basil sauce, typical of the area of Genova) in the web is $29,700,000$, the frequency of "Genova" is $420,817$. This results in $p(\text{“}pesto\text{”}) = 29,700,000/2,147,436,244 = 0.014$ and $p(\text{“}Genova\text{”}) = 420,817/2,147,436,244 = 0.0002$. Therefore, the expected probability for "pesto" and "Genova" occurring in the same page is $p(\text{“}pesto\text{”} \cap \text{“}Genova\text{”}) = 0.0002 * 0.014 = 0.0000028$, which corresponds to an expected page count of $6,013$ pages. Looking for the actual web counts, we obtain $103,000$ pages for the query "+pesto +Genova", well above the expected: this clearly indicates that the thematic and geographical parts of the query are strongly correlated and this query is particularly relevant to the user's information needs. The MI of "pesto" and "Genova" turns out to be 0.0011. As a comparison, the MI obtained for "pesto" and "Torino" (a city that has no connection with the famous pesto sauce) is only 0.00002.

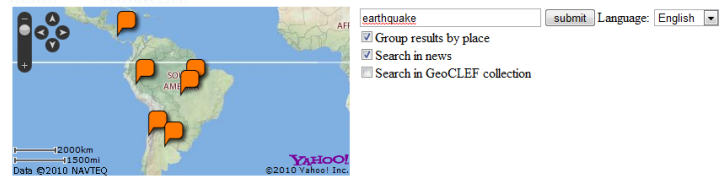
Users may decide to get the results grouped by locations, sorted by the MI of the location with respect to the query, or to obtain a unique list of results. In the first case, the result fusion step is skipped. More options include the possibility to search in news, or in the GeoCLEF collection (see Figure 7.3). In Figure 7.4 we see an example of results grouped by locations, with the query "earthquake", news search mode, and a footprint covering South America (results retrieved on May 25th, 2010). The day before an earthquake of magnitude 6.5 occurred in the Amazonian state of Acre in Brazil's North Region. Results reflect this event by presenting Brazil as the first result. This example show how Geooreka! can be used to detect occurring events in specific regions.

7.1.3 Result Fusion

The fusion of the results is done by carrying out a voting among the 20 most relevant (according to their MI) searches. The voting scheme is a modification the Borda count, a scheme introduced in 1770 for the election of members of the French Academy of Sciences and currently used in many electoral systems and in the economics field Levin and Nalebuff (1995). In the classical (discrete) Borda count, each experts assign a mark to the candidates. The mark is given by the number of candidates that the experts considers worse than it. The winner of the election is the candidate whose sum of marks is greater (see Figure 7.5 for an example).

In our approach, each search is an expert and the candidates are the search entries (snippets). The differences with respect to the standard Borda count are that: marks are given by 1 plus the number of candidates worse than the voted candidate, normalised over the length of the list of returned snippets (normalisation is required due to the

Figure 7.3: Geooreka! input page.



Query text: **earthquake** freq: 2770
Searching in News
Map Center: Lat -23.322080011378436 , Lon -63.28125
Box Coordinates:
Left Lat 11.092165893501996 , Left Lon -116.015625
Right Lat -50.79204706440685 , Right Lon -10.546875
Points in the box: (zoom Level: 15)

1. Brazil , lat:-10 lon:-55 , webCounts: 1447 ,
2. Argentina , lat:-34 lon:-64 , webCounts: 708 ,
3. Peru , lat:-10 lon:-76 , webCounts: 535 ,
4. South America , lat:-14.6048471550539 lon:-57.65625 , webCounts: 452 ,
5. Chile , lat:-30 lon:-71 , webCounts: 351 ,
6. Costa Rica , lat:10 lon:-84 , webCounts: 157 ,

Top hits, grouped by place relevance:
Results for: *earthquake* and *Brazil* MI: 2.9526762878708037E-7more results from Yahoo!

1. **Strong earthquake in Brazil**
   KUALA LUMPUR: A strong earthquake with a magnitude of 6.4 on the Richter scale occurred in western Brazil at 12.18am today.
   http://www.nst.com.my/nst/articles/Strongearthquakeinbrazil/Article/
2. **Strong Earthquake In Brazil :: Bernama.com**
   KUALA LUMPUR,May 25 (BERNAMA) -- A strong earthquake with a magnitude of 6.4 on the Richter scale occurred in western Brazil at 12.18am Tuesday.
   http://www.bernama.com/bernama/v5/news_lite.php?id=500874
3. **USGS: 6.5-magnitude earthquake in Acre, Brazil**
   SAN FRANCISCO (MarketWatch) -- A 6.5-magnitude earthquake struck the Acre region in Brazil at 11:18 a.m. local time, the U.S. Geological Survey said Monday. The epicenter was about 79 miles from

Figure 7.4: Geooreka! result page for the query "Earthquake", geographically constrained to the South America region using the map-based interface.

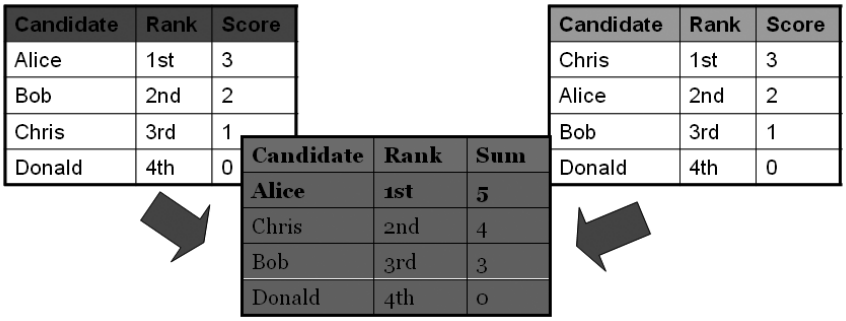


Figure 7.5: Borda count example.

fact that the lists may not have the same length); and that we assign to each expert a confidence score, consisting in the MI obtained for the search itself.

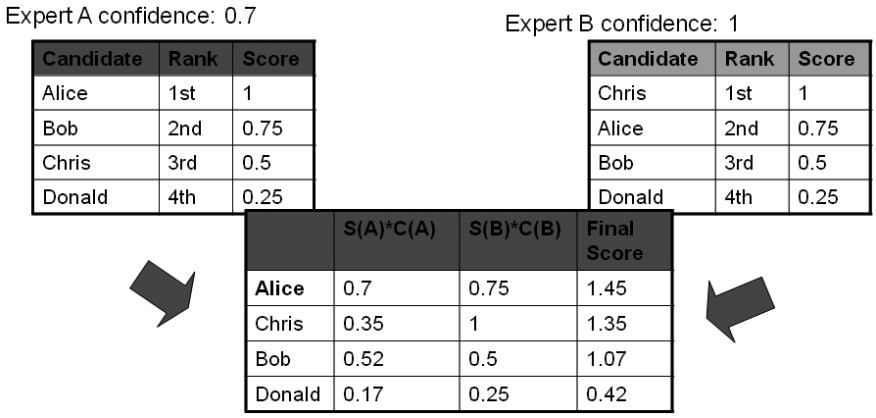


Figure 7.6: Example of our modification of Borda count. $S(x)$: score given to the candidate by expert x. $C(x)$ confidence of expert x.

In Figure 7.6 we show the differences with respect to the previous example using our weighting scheme. In this way, we assure that the relevance of the search is reflected in the ranked list of results.

7.2 Experiments

An evaluation was carried out by adapting the system to work on the GeoCLEF collection. In this way, it was possible to compare the results that could be obtained by specifying the geographic footprint by means of keywords and those that could be obtained using a map-based interface to define the geographic footprint of the query.

With this setup, topic title only was used as input for the Geooreka! thematic part, while the area corresponding to the geographic scope of the topic was manually selected. Probabilities were calculated using the number of occurrences in the GeoCLEF collection, indexed with GeoWorSE using GeoWordNet as a resource (see Section 5.1). Occurrences for toponyms were calculated by taking into account only the *geo* index. The results were calculated over the 25 topics of GeoCLEF-2005, minus the queries in which the geographic footprint was composed of disjoint areas (for instance, "Europe" and "USA" or "California" and "Australia"). Mean Reciprocal Rank (MRR) was used as a measure of accuracy, since MAP could not be calculated for Geooreka! without fusion. Table 7.4 shows the obtained results.

The results show that using result fusion the MRR drops with respect to the other systems, indicating that redundancy (obtaining the same documents for different places) in general is not useful. The reason is that repeated results, although not relevant, obtain more weight than relevant results that appear only one time. The *Geooreka*! version that does not use fusion but shows the results grouped by place obtained better MRR than the keyword-based system.

Table 7.5 shows the MRR obtained for each of the 5 most relevant toponyms identified by Geooreka with respect to the thematic part of every query. In many cases the toponym related to the most relevant result is different from the original query keyword, indicating that the system did not return merely a list of relevant documents, but carried out also a sort of geographical mining of the collection. In many cases, it was possible to obtain a relevant result for each of the most 5 relevant toponyms, and a MRR of 1 for every toponym in topic `GC-017`: "Bosnia", "Sarajevo", "Srebrenica", "Pale". These results indicate that geographical diversity may represent an interesting direction for further investigation.

Table 7.5: MRR obtained for each of the most relevant toponym on GeoCLEF 2005 topics.

topic	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
GC-002	1.000	0.000	0.500	1.000	1.000
	London	Italy	Moscow	Belgium	Germany
GC-003	1.000	1.000	0.000	1.000	0.000
	Haiti	Mexico	Guatemala	Brazil	Chile
GC-005	1.000	1.000			
	Japan	Tokyo			

Continued on Next Page...

topic	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
GC-007	1.000 UK	0.200 Ireland	1.000 Europe	1.000 Belgium	0.000 France
GC-008	1.000 France	0.333 Turkey	1.000 UK	0.250 Denmark	0.000 Europe
GC-009	1.000 India	1.000 Asia	0.200 China	1.000 Pakistan	1.000 Nepal
GC-010	0.333 Germany	1.000 Netherlands	1.000 Amsterdam		
GC-011	1.000 UK	0.500 Europe	0.000 Italy	0.000 France	1.000 Ireland
GC-012	0.000 Germany	0.000 Berlin			
GC-014	1.000 Great Britain	0.500 Irish Sea	1.000 North Sea	0.333 Denmark	
GC-015	1.000 Ruanda	1.000 Kigali			
GC-017	1.000 Bosnia	1.000 Sarajevo	1.000 Srebrenica	1.000 Pale	1.000
GC-018	0.333 Glasgow	1.000 Scotland	0.000 Park	0.250 Edinburgh	1.000 Braemer
GC-019	1.000 Spain	0.200 Germany	0.500 Italy	1.000 Europe	0.500 Ireland
GC-020	1.000 Orkney				
GC-021	1.000 North Sea	1.000 UK			
GC-022	1.000 Scotland	0.500 Edinburgh	1.000 Glasgow	1.000 West Lothian	0.000 Falkirk
GC-023	0.200 Glasgow	0.000 Scotland			
GC-024	1.000 Scotland				

Table 7.4: MRR obtained with Geooreka!, compared to MRR obtained using the GeoWordNet-based GeoWorSE system, Topic Only runs.

topic	GeoWN	Geooreka! (No Fusion)	Geooreka! (+ Borda Fusion)
GC-002	0.250	1.000	0.077
GC-003	0.013	1.000	1.000
GC-005	1.000	1.000	1.000
GC-006	0.143	0.000	0.000
GC-007	1.000	1.000	0.500
GC-008	0.143	1.000	0.500
GC-009	1.000	1.000	0.167
GC-010	1.000	0.333	0.200
GC-012	0.500	1.000	0.500
GC-013	1.000	0.000	0.200
GC-014	1.000	0.500	0.500
GC-015	1.000	1.000	1.000
GC-017	1.000	1.000	1.000
GC-018	1.000	0.333	1.000
GC-019	0.200	1.000	1.000
GC-020	0.500	1.000	0.125
GC-021	1.000	1.000	1.000
GC-022	0.333	1.000	0.500
GC-023	0.019	0.200	0.167
GC-024	0.250	1.000	0.000
GC-025	0.500	0.000	0.000
average	0.612	0.756	0.497

7.3 Toponym Disambiguation for Probability Estimation

An analysis of the results of topic GC-008 ("Milk Consumption in Europe") in Table 7.5 showed that the MI obtained for "Turkey" was abnormally high with respect to the expected value for this country. The reason is that in most documents, the name "turkey" was referring to the animal and not to the country. This kind of ambiguity represents one of the most important issue at the time of estimating the probability of occurence of places. The importance of this issue grows together with the size and the scope of the collection being searched. The web, therefore, constitutes the worst scenario with respect to this problem. For instance, in Figure 7.7 it can be seen a search for "water sports" near the city of Trento in Italy. One of the toponyms in the area is "Vela", which means "sail" in Italian (it means also "candle" in Spanish). Therefore, the number of page hits obtained for "Vela", used to estimate the probability of finding this toponym in the web, is flawed because of the different meanings that it could take. This issue has been partially overcome in Geoooreka! by adding to the query the holonym of the placenames. However, even in this way, errors are very common, especially due to geo-non geo ambiguities. For instance, the web count of "Paris" may be refined with the including entity, obtaining "Paris, France" and "Paris, Texas", among others. However, the web count of "Paris, Texas" includes the occurrences of a Wim Wenders' movie with the same name. This problem shows the importance of tagging places in the web, and in particular of disambiguating them in order to give search engines a way to improve searches.

submit Language: English

Group results by place
Search in news
Search in GeoCLEF collection

Query text: **sport acquatici** freq: 3040045
Map Center: Lat 46.07084881222032 , Lon 11.125030517578125
Box Coordinates:
Left Lat 46.094662318622824 , Left Lon 11.056365966796875
Right Lat 46.04702502686093 , Right Lon 11.193695068359375
Points in the box: (zoom Level: 6)

1. Vela , lat:46.0823698794352 lon:11.1020708084106 , webCounts: 12600055 ,
2. Trent , lat:46.0678714011874 lon:11.1210823059082 , webCounts: 957016 ,
3. Trento , lat:46.0704319843444 lon:11.1207389831543 , webCounts: 838003 ,
4. Provincia di Trento , lat:46.0696578670802 lon:11.121768951416 , webCounts: 411002 ,
5. Povo , lat:46.0669781465587 lon:11.1550283432007 , webCounts: 181000 ,
6. San Rocco , lat:46.05 lon:11.15 , webCounts: 113001 ,
7. Sopramonte , lat:46.0740642357402 lon:11.061429977417 , webCounts: 1910 ,

Top hits, grouped by place relevance:
Results for: *sport acquatici* and *Sopramonte* MI: 3.1006577930890607E-9**more results from Yahoo!**

1. **Hotel Folgaria - Prenotare alberghi a Folgaria a prezzi ...**
   Prenotare gli hotel in Folgaria a prezzi esclusivi - Le valutazioni degli hotel di chi è stato a Folgaria sono d'ausilio per la scelta - Hotel Trentino-Alto Adige, Italia
   http://www.hotel.info/it.hotels/Folgaria_57874/hotels.aspx

Figure 7.7: Results of the search "water sports" near Trento in Geooreka!.

Chapter 8

Conclusions, Contributions and Future Work

This Ph.D. thesis represents the first attempt to carry out an exhaustive research over Toponym Disambiguation from an NLP perspective and to study its relation to IR applications such as Geographical Information Retrieval, Question Answering and Web search. The research work was structured as follows:

1. Analysis of resources commonly used as Toponym repositories, such as gazetteers and geographic ontologies;

2. Development and comparison of Toponym Disambiguation methods;

3. Analysis of the effect of TD in GIR and QA;

4. Study of applications in which TD may result useful.

8.1 Contributions

The main contributions of this work are:

- The Geo-WordNet[1] expansion for the WordNet ontology, especially aimed to researchers working on toponym disambiguation and in the Geographical Information Retrieval field;

[1]Listed in the official WordNet "related projects" page: `http://wordnet.princeton.edu/wordnet/related-projects/`

- The analysis of different resources and how they fit with the needs of researchers and developers working on Toponym Disambiguation, including a case study of the application of TD to a practical problem;
- The design and the evaluation of two Toponym Disambiguation methods based on WordNet structure and maps, respectively;
- Experiments to determine under which conditions TD may be used to improve the performance in GIR and QA;
- Experiments to determine the relation between error levels in TD and results in GIR and QA;
- The study on the "L'Adige" news collection highlighted the problems that could be found while working on a local news collection, with a street level granularity;
- Implementation of a prototype search engine (Geooreka!) that exploits co-occurrences of toponyms and concepts.

8.1.1 Geo-WordNet

Geo-WordNet was obtained as an extension of WordNet 2.0 obtained by mapping the locations included in WordNet to locations in the Wikipedia-World gazetteer. This resource allowed to carry out the comparative evaluation between the two Toponym Disambiguation methods, which otherwise would have been impossible. Since the resource has been distributed online, it has been downloaded by 237 universities, institutions and private companies, indicating the level of interest for this resource. Apart from the contributions to TD research, it can be used in various NLP tasks to include geometric calculations and thus create a kind of bridge between GIS and GIR research communities.

8.1.2 Resources for TD in Real-World Applications

One of the main issues encountered during the research work related to this Ph.D. thesis was the selection of a proper resource. It has been observed that resources vary in *scope*, *coverage* and *detail*, and compared the most commonly used ones. The study carried out over TD in news, using "L'Adige" collection, showed that 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 such cases, it is necessary to develop a customized resource, integrating

information from different sources: in our case we had to complement Wikipedia and Geonames data with information retrieved using the Google maps API.

8.1.3 Conclusions drawn from the Comparison of TD Methods

The combination of GeoSemCor and Geo-WordNet allows to compare the performance of different methods: knowledge-based, map-based and data-driven. In this work, for the first time, a knowledge-based method was compared to a map-based method on the same test collection. In this comparison, the results showed that the map-based method needs more context than the knowledge-based one, and that the second one obtains better accuracy. However, GeoSemCor is biased toward the first (most common) sense and is derived from SemCor, which was developed for the evaluation of WSD methods, not TD methods. Although it could be used for the comparison of methods that employ WordNet as a toponym resource, it cannot be used to compare methods that are based on resources with a wider coverage and detail such as Geonames or GeoPlanet. Leidner (2007) in his TR-CoNLL corpus detected a bias towards the "most salient" sense, which in the case of GeoSemCor corresponds to the most frequent sense. He considered this bias to be a factor rendering supervised TD infeasible due to overfitting.

8.1.4 Conclusions drawn from TD Experiments

The results obtained in the experiments with Toponym Disambiguation and the GeoWorSE system revealed that disambiguation is useful only in the case of short queries (as observed by Sanderson (1996) in the case of general WSD) and if a detailed toponym repository is used, reflecting the working configuration of web search engines. 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. Actually, disambiguation over WordNet has the effect of worsening the retrieval accuracy, because of the disambiguation errors introduced. Toponym Disambiguation allowed also to improve results when the ranking method was modified, using a Geographically Adjusted Ranking technique, only in the cases where Geonames was used. This result remarks the importance of the detail of the resource used with respect to TD. The experiments carried out with the introduction of artificial ambiguity showed that, using WordNet, the variation is small even if the number of errors is 60% of the total toponyms in the collection. However, it should be noted that the 60% errors is relative to the space of referents given by WordNet 1.6, the resource used in the CLIR-WSD collection. Is it possible that some of the introduced errors

had the result of correcting instances instead than introduce actual errors. Another conclusion that could be drawn at this point is that GeoCLEF somehow failed in its supposed purpose of evaluating the performance in *geographical* IR: in this work we noted that long queries, like those used in the "title and description" and "all fields" runs for the official evaluation, were not representing an issue. The geographical scope of such queries is well-defined enough to not represent a problem for generic IR system. Short queries, like those of the "title only" configuration were not evaluated, and the results obtained with this configuration were worse than those that could be obtained with longer queries. Most queries were also too broad from a geographical viewpoint in order to be affected by disambiguation errors.

It has been observed that the results in QA are not affected by Toponym Disambiguation. QA systems can be affected by a quantity of errors, such as wrong question classification, wrong analysis, incorrect candidate entity detection that are more relevant to the final result than the errors that can be produced by Toponym Disambiguation. On the other hand, even if no errors occur in the various modules of QA systems, redundancy allows to compensate the errors that may result from incorrect disambiguation of toponyms.

8.1.5 Geooreka!

This search engine has been developed on the basis of the results obtained with GeoCLEF topics, suggesting that the use of term-based queries may not be the optimal method to express a geographically constrained information need. Geooreka! represents a prototype search engine that can be used both for basic web retrieval purposes or for information mining on the web, returning toponyms that are particularly relevant to some event or item. The experiments showed that it is very difficult to correctly estimate the probabilities for the co-occurrences of place and events since place names in the web are not disambiguated. This result confirms that Toponym Disambiguation plays a key role in the development of the geospatial-semantic web, with regard to facilitating the search for geographical information.

8.2 Future Work

The use of the LGL (Local/GLobal) collection that has recently been introduced by Michael D. Lieberman (2010) could represent an interesting follow-up of the experiments on toponym ambiguity. The collection (described in Appendix D) contains documents extracted from both local newspaper and general ones and enough instances to

represent a sound test-bed. This collection was not yet available at the time of writing. Comparing with Yahoo! placemaker would also be interesting in order to see how the developed TD methods perform with respect to this commercial system.

We should also consider postal codes, since they can also be ambiguous: for instance, "16156" is a code that may refer to Genoa in Italy, or to a place in Pennsylvania in the United States. They could also provide useful context to disambiguate other ambiguous toponyms. In this work, we did not take them into account because there was no resource listing them together with their coordinates. Only recently they have been added to Geonames.

Another work could be the use of different IR models and a different configuration of the IR system. Terms still play the most important role in the search engine, and the parameters for the Geographically Adjusted Ranking were not studied extensively. These parameters can be studied in future to determine an optimal configuration that allows to better exploit the presence of toponyms (that is, geographical information) in the documents. The *geo* index could also be used as a spatial index, and some research could be carried out by combining the results of text-based search with the spatial search using result fusion techniques.

Geooreka! should be improved, especially under the aspect of user interface. In order to do this, it is necessary to implement techniques that allow to query the search engine with the same toponyms that are visible on the map, by allowing to users to select the query footprint by drawing an area on the map and not, as in the prototype, use the visualized map as the query footprint. Users should also be able to select multiple areas and not a single area. It should be carried out an evaluation in order to obtain a numerical estimation of the advantage obtained by the diversification of the results from the geographical point of view. Finally, we need also to evaluate the system from a user perspective: the fact that people would like to query the web through drawing regions on a map is not clear, and spatial literacy of users on the web is very low which means they may find it hard to interact with maps.

Currently, another extension of WordNet similar to Geo-WordNet, named Star-WordNet, is under study. This extension would label astronomical object with their astronomical coordinates, like toponyms were labelled geographical coordinates in Geo-WordNet. Ambiguity of astronomical objects like planets, stars, constellations and galaxies is not a problem, since there are policies in order to assign names that are established by supervising entities; however, StarWordNet may help in detecting some astronomical/not astronomical ambiguities (such as Saturn the planet or the family of rockets?) in specialised texts.

Bibliography

Steven Abney, Michael Collins, and Amit Singhal. Answer extraction. In *In Proceedings of ANLP 2000*, pages 296–301, 2000. 29

Rita M. Aceves, Luis Villaseñor, and Manuel Montes. Towards a Multilingual QA System Based on the Web Data Redundancy. In Piotr S. Szczepaniak, Janusz Kacprzyk, and Adam Niewiadomski, editors, *AWIC*, volume 3528 of *Lecture Notes in Computer Science*, pages 32–37. Springer, 2005. 29

Eneko Agirre and Oier Lopez de Lacalle. UBC-ALM: Combining k-NN with SVD for WSD. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 341–345. ACL, 2007. 53, 102, 113

Eneko Agirre and German Rigau. Word Sense Disambiguation using Conceptual Density. In *16th Conference on Computational Linguistics (COLING '96)*, pages 16–22, Copenhaghen, Denmark, 1996. 65

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, New York, NY, USA, 2009. ACM. doi: http://doi.acm.org/10.1145/1498759.1498766. 18

Kisuh Ahn, Beatrice Alex, Johan Bos, Tiphaine Dalmas, Jochen L. Leidner, and Matthew Smillie. Cross-lingual question answering using off-the-shelf machine translation. In Peters et al. (2005), pages 446–457. 28

James Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer International Series on Information Retrieval. Kluwer Academic Publ, 2002. 5

Einat Amitay, Nadav Harel, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280, Sheffield, UK, 2004. 60

Geoffrey Andogah. *Geographically Constrained Information Retrieval*. PhD thesis, University of Groningen, 2010. iii, 3

Geoffrey Andogah, Gosse Bouma, John Nerbonne, and Erwin Koster. Placename ambiguity resolution. In Nicoletta Calzolari et al., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/. 60

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, NY, 1999. xv, 9, 10

Ricardo Baeza-Yates, Aristides Gionis, Flavio Junqueira, Vanessa Murdock, Vassilis Plachouras, and Fabrizio Silvestri. The impact of caching on search engines. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 183–190, New York, NY, USA, 2007. ACM. doi: http://doi.acm.org/10.1145/1277741.1277775. 93

Matthias Baldauf and Rainer Simon. Getting context on the go: mobile urban exploration with ambient tag clouds. In *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 1–2, New York, NY, USA, 2010. ACM. doi: http://doi.acm.org/10.1145/1722080.1722094. 33

Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of CICLing 2002*, pages 136–145, London, UK, 2002. Springer-Verlag. 57, 69, 70

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *J. Artif. Int. Res.*, 17(1): 35–55, 2002. 18

Alberto Belussi, Omar Boucelma, Barbara Catania, Yassine Lassoued, and Paola Podestà. Towards similarity-based topological query languages. In *Current Trends in Database Technology - EDBT 2006, EDBT 2006 Workshops PhD, DataX, IIDB, IIHA, ICSNW, QLQP, PIM, PaRMA, and Reactivity on the Web, Munich, Germany, March 26-31, 2006, Revised Selected Papers*, pages 675–686. Springer, 2006. 17

Imene Bensalem and Mohamed-Khireddine Kholladi. Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, 6(6):653–659, 2010. 5, 179

Davide Buscaldi and Bernardo Magnini. Grounding toponyms in an italian local news corpus. In *Proceedings of GIR'10 Workshop on Geographical Information Retrieval*, 2010. 76, 179

Davide Buscaldi and Paolo Rosso. On the relative importance of toponyms in geoclef. In Peters et al. (2008), pages 815–822. 13

Davide Buscaldi and Paolo Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 22(3): 301–313, 2008a. 59, 72

Davide Buscaldi and Paolo Rosso. Geo-WordNet: Automatic Georeferencing of WordNet. In *Proc. 5th Int. Conf. on Language Resources and Evaluation, LREC-2008*, Marrakech, Morocco, 2008b. 45

Davide Buscaldi and Paolo Rosso. Using GeoWordNet for Geographical Information Retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.*, pages 863–866, 2009a. 13

BIBLIOGRAPHY

Davide Buscaldi and Paolo Rosso. Geooreka: Enhancing Web Searches with Geographical Information. In *Proc. Italian Symposium on Advanced Database Systems SEBD-2009*, pages 205–212, Camogli, Italy, 2009b. 120

Davide Buscaldi, Paolo Rosso, and Francesco Masulli. The upv-unige-CIAOSENSO WSD System. In *Senseval-3 workshop, ACL 2004*, pages 77–82, Barcelona, Spain, 2004. 67

Davide Buscaldi, José Manuel Gómez, Paolo Rosso, and Emilio Sanchis. N-gram vs. keyword-based passage retrieval for question answering. In Peters et al. (2007), pages 377–384. 105

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. A wordnet-based indexing technique for geographical information retrieval. In Peters et al. (2007), pages 954–957. 17

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Mller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, Maarten de Rijke, and Danilo Giampiccolo, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 939–946. Springer, Berlin, 2006c. 16, 88

Davide Buscaldi, Yassine Benajiba, Paolo Rosso, and Emilio Sanchis. Web-based anaphora resolution for the quasar question answering system. In Peters et al. (2008), pages 324–327. 105

Davide Buscaldi, José M. Perea, Paolo Rosso, Luis Alfonso Ureña, Daniel Ferrés, and Horacio Rodríguez. Geotextmess: Result fusion with fuzzy borda ranking in geographical information retrieval. In Peters et al. (2009), pages 867–874. 16

Davide Buscaldi, Paolo Rosso, José Manuel Gómez, and Emilio Sanchis. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems (JIIS)*, 34(2):113–134, 2009. doi: 10.1007/s10844-009-0082-y. 105

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM. doi: http://doi.acm.org/10.1145/290941.291025. 18

Nuno Cardoso, David Cruz, Marcirio Silveira Chaves, and Mário J. Silva. Using geographic signatures as query and document scopes in geographic ir. In Peters et al. (2008), pages 802–810. 17

Yen-Yu Chen, Torsten Suel, and Alexander Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 277–288, New York, NY, USA, 2006. ACM. doi: http://doi.acm.org/10.1145/1142473.1142505. 122

Paul Clough, Mark Sanderson, Murad Abouammoh, Sergio Navarro, and Monica Paramita. Multiple approaches to analysing query diversity. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 734–735, New York, NY, USA, 2009. ACM. doi: http://doi.acm.org/10.1145/1571941.1572102. 18

David Fernández-Amorós, Julio Gonzalo, and Felisa Verdejo. The role of conceptual relation in word sense disambiguation. In *NLDB'01*, pages 87–98, Madrid, Spain, 2001. 75

Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, and Fernando Llopis. University of alicante at geoclef 2005. In Peters et al. (2006), pages 924–927. 13

Daniel Ferrés and Horacio Rodríguez. Experiments adapting an open-domain question answering system to the geographical domain using scope-based resources. In *Proceedings of the Multilingual Question Answering Workshop of the EACL 2006*, Trento, Italy, 2006. 27

Daniel Ferrés and Horacio Rodríguez. TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, chapter 5152, pages pp. 830–833. Springer, Budapest, Hungary, 2008. 13, 146

Daniel Ferrés, Alicia Ageno, and Horacio Rodríguez. The geotalp-ir system at geoclef 2005: Experiments using a qa-based ir system, linguistic analysis, and a geographical thesaurus. In Peters et al. (2006), pages 947–955. 17

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages pp. 363–370, U. of Michigan - Ann Arbor, 2005. ACL. 13, 88

Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *LOCWEB '08: Proceedings of the first international workshop on Location and the web*, pages 49–56, New York, NY, USA, 2008. ACM. doi: http://doi.acm.org/10.1145/1367798.1367806. 3

Eric Garbin and Inderjeet Mani. Disambiguating toponyms in news. In *conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT05)*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1220575.1220621. 2, 60

Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. Geoclef: The clef 2005 cross-language geographic information retrieval track overview. In Peters et al. (2006), pages 908–919. 15, 24

Fredric C. Gey, Ray R. Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Maria Di Nunzio, and Nicola Ferro. Geoclef 2006: The clef 2006 cross-language geographic information retrieval track overview. In Peters et al. (2007), pages 852–876. xi, 24, 25, 27

Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, and Biswanath Dutta. GeoWordNet: A Resource for Geospatial Applications. In Lora Aroyo, Grigoris Antoniou,

Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *ESWC (1)*, volume 6088 of *Lecture Notes in Computer Science*, pages 121–136. Springer, 2010. 45, 179

José Manuel Gómez, Davide Buscaldi, Empar Bisbal, Paolo Rosso, and Emilio Sanchis. Quasar: The question answering system of the universidad politécnica de valencia. In Peters et al. (2006), pages 439–448. 105

José Manuel Gómez, Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. Jirs language-independent passage retrieval system: A comparative study. In *5th Int. Conf. on Natural Language Processing, ICON-2007*, Hyderabad, India, 2007. 109

Julio Gonzalo, Felisa Verdejo, Irin Chugur, and José Cigarrán. Indexing with WordNet Synsets can improve Text Retrieval. In *COLING/ACL '98 workshop on the Usage of WordNet for NLP*, pages 38–44, Montreal, Canada, 1998. 51, 87

Ronald L. Graham. An efficient algorith for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133, 1972. 91

Mark A. Greenwood. Using pertainyms to improve passage retrieval for questions requesting information about a location. In *SIGIR*, 2004. 28

Sanda Harabagiu, Dan Moldovan, and Joe Picone. Open-domain Voice-activated Question Answering. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1072228.1072397. 31

Andreas Henrich and Volker Luedecke. Characteristics of Geographic Information Needs. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 1–6, New York, NY, USA, 2007. ACM. doi: 10.1145/1316948.1316950. 12

Ed Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin yew Lin. Question Answering in Webclopedia. In *The Ninth Text REtrieval Conference*, 2000. 27, 28

David Johnson, Vishv Malhotra, and Peter Vamplew. More effective web search using bigrams and trigrams. *Webology*, 3(4), 2006. 12

Christopher 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 *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–388, New York, NY, USA, 2002. ACM. doi: http://doi.acm.org/10.1145/564376.564457. 12, 19

Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1): pp. 79–86, 1951. 124

Ray R. Larson. Cheshire at geoclef 2008: Text and fusion approaches for gir. In Peters et al. (2009), pages 830–837. 16

Ray R. Larson, Fredric C. Gey, and Vivien Petras. Berkeley at geoclef: Logistic regression and fusion for geographic information retrieval. In Peters et al. (2006), pages 963–976. 16

Joon Ho Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages pp. 267–276, New York, NY, USA, 1997. ACM. doi: http://doi.acm.org/10.1145/258525.258587. 149, 151

Jochen L. Leidner. Experiments with geo-filtering predicates for ir. In Peters et al. (2006), pages 987–996. 13

Jochen L. Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417, July 2006. doi: 10.1016/j.compenvurbsys.2005.07.003. 55

Jochen L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, School of Informatics, University of Edinburgh, 2007. iii, 3, 4, 5, 135

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *5th annual international conference on Systems documentation (SIGDOC '86)*, pages 24–26, 1986. 57, 69

Jonathan Levin and Barry Nalebuff. An Introduction to Vote-Counting Schemes. *Journal of Economic Perspectives*, 9(1): 3–26, 1995. 125

Yi Li. Probabilistic Toponym Resolution and Geographic Indexing and Querying. Master's thesis, University of Melbourne, 2007. 15

Yi Li, Alistair Moffat, Nicola Stokes, and Lawrence Cavedon. Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In *3rd Workshop on Geographic Information Retrieval (GIR 2006)*, 2006a. 60, 61

Yi Li, Nicola Stokes, Lawrence Cavedon, and Alistair Moffat. Nicta i2d2 group at geoclef 2006. In Peters et al. (2007), pages 938–945. 17

ACE English Annotation Guidelines for Entities. Linguistic Data Consortium, 2008. `http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf`. 76

Xiaoyong Liu and W. Bruce Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, 2002. 28

Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Multilingual question/answering: the DIOGENE system. In *The 10th Text REtrieval Conference*, 2001. 105

Thomas Mandl, Paula Carvalho, Giorgio Maria Di Nunzio, Fredric C. Gey, Ray R. Larson, Diana Santos, and Christa Womser-Hacker. Geoclef 2008: The clef 2008 cross-language geographic information retrieval track overview. In Peters et al. (2009), pages 808–821. 145

BIBLIOGRAPHY

Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. SpatialML: Annotation Scheme, Corpora, and Tools. In Nicoletta Calzolari et al., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/. 55

Fernando Martínez, Miguel Angel García, and Luis Alfonso Ureña. Sinai at clef 2005: Multi-8 two-years-on and multi-8 merging-only tasks. In Peters et al. (2006), pages 113–120. 13

Bruno Martins, Ivo Anastácio, and Pável Calado. A machine learning approach for resolving place references in text. In *13th International Conference on Geographic Information Science (AGILE 2010)*, 2010. 61

Jagan Sankaranarayanan Michael D. Lieberman, Hanan Samet. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE'10)*, pages 201–212, 2010. 136, 179

Rada Mihalcea. Using wikipedia for automatic word sense disambiguation. In Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*, pages 196–203. The Association for Computational Linguistics, 2007. 58

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 43

Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, New York, USA, 2003. 27, 116

David Mountain and Andrew MacFarlane. Geographic Information Retrieval in a Mobile Environment: Evaluating the Needs of Mobile Individuals. *Journal of Information Science*, 33(5):515–530, 2007. 16

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. URL `http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002`. Publisher: John Benjamins Publishing Company. 13

Günter Neumann and Bogdan Sacaleanu. Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In Peters et al. (2005), pages 411–422. 105

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. Exploiting parallel texts for word sense disambiguation: an empirical study. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 455–462, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075154. 53, 58

Appendix to the 15th TREC proceedings (TREC 2006). NIST, 2006. `http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf`. 21

Hannu Nurmi. Resolving Group Choice Paradoxes Using Probabilistic and Fuzzy Concepts. *Group Decision and Negotiation*, 10(2):177–199, 2001. 147

Andreas M. Olligschlaeger and Alexander G. Hauptmann. Multimodal Information Systems and GIS: The Informedia Digital Video Library. In *1999 ESRI User Conference*, San Diego, CA, 1999. 59, 60

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006. 146

Simon Overell. *Geographic Information Retrieval: Classification, Disambiguation and Modelling.* PhD thesis, Imperial College London, 2009. xi, 3, 5, 24, 25, 36, 82, 179

Simon E. Overell, João Magalhães, and Stefan M. Rüger. Forostar: A system for gir. In Peters et al. (2007), pages 930–937. 60

Monica Lestari Paramita, Jiayu Tang, and Mark Sanderson. Generic and Spatial Approaches to Image Search Results Diversification. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 603–610, Berlin, Heidelberg, 2009. Springer-Verlag. doi: http://dx.doi.org/10.1007/978-3-642-00958-7_56. 18

Robert C. Pasley, Paul Clough, and Mark Sanderson. Geo-Tagging for Imprecise Regions of Different Sizes. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 77–82, New York, NY, USA, 2007. ACM. 59

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 4th International Conference*, volume 2588 of *Lecture Notes in Computer Science*, pages 241–257. Springer, Berlin, 2003. 69

José M. Perea, Miguel Angel García, Manuel García, and Luis Alfonso Ureña. Filtering for Improving the Geographic Information Search. In Peters et al. (2008), pages 823–829. 145

Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors. *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*, 2005. Springer. 139, 142

Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors. *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, volume 4022 of *Lecture Notes in Computer Science*, 2006. Springer. 140, 141, 142

Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors. *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006,*

Alicante, Spain, September 20-22, 2006, Revised Selected Papers, volume 4730 of *Lecture Notes in Computer Science*, 2007. Springer. 140, 141, 142

Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors. *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, 2008. Springer. 139, 140, 142

Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors. *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, 2009. Springer. 140, 141

Emanuele Pianta and Roberto Zanoli. Exploiting SVM for Italian Named Entity Recognition. *Intelligenza Artificiale, Special issue on NLP Tools for Italian*, IV(2), 2007. In Italian. 76

Bruno Pouliquen, Marco Kimler Marco, Ralf Steinberger, Camelia Igna, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, and Clive Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC 2006*, Genova, Italy, 2006. 19

Ross Purves and Chris B. Jones. Geographic information retrieval (gir). *Computers, Environment and Urban Systems*, 30(4):375–377, July 2006. xv, 12

Erik Rauch, Michael Bukatin, and Kenneth Baker. A confidence-based framework for disambiguating geographic terms. In *HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 50–54, Edmonton, Alberta, Canada, 2003. 59, 60

Ian Roberts and Robert J. Gaizauskas. Data-intensive question answering. In *ECIR*, volume 2997 of *Lecture Notes in Computer Science*. Springer, 2004. 28

Kirk Roberts, Cosmin Adrian Bejan, and Sanda Harabagiu. Toponym disambiguation using events. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, 2010. 179

Vincent B. Robinson. Individual and multipersonal fuzzy spatial relations acquired using human-machine interaction. *Fuzzy Sets and Systems*, 113(1):133 – 145, 2000. doi: DOI:10.1016/S0165-0114(99)00017-2. URL http://www.sciencedirect.com/science/article/B6V05-43G453N-C/2/e0369af09e6faac7214357736d3ba30b. 17

Paolo Rosso, Francesco Masulli, Davide Buscaldi, Ferran Pla, and Antonio Molina. Automatic noun sense disambiguation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 4th International Conference*, volume 2588 of *Lecture Notes in Computer Science*, pages 273–276. Springer, Berlin, 2003. 67

Gerard Salton and Michael Lesk. Computer evaluation of indexing and text processing. *J. ACM*, 15(1):8–36, 1968. 11

Mark Sanderson. Word sense disambiguation and information retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 142–151, New York, NY, USA, 1994. Springer-Verlag New York, Inc. 87

Mark Sanderson. *Word Sense Disambiguation and Information Retrieval*. PhD thesis, University of Glasgow, Glasgow, Scotland, UK, 1996. 6, 51, 135

Mark Sanderson. Retrieving with good sense. *Information Retrieval*, 2(1):49–69, 2000. 87

Mark Sanderson and Yu Han. Search Words and Geography. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 13–14, New York, NY, USA, 2007. ACM. 12

Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *Proceedings of Workshop on Geographic Information Retrieval (GIR04)*, 2004. 3, 12

Mark Sanderson, Jiayu Tang, Thomas Arni, and Paul Clough. What else is there? search diversity examined. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soulé-Dupuy, editors, *ECIR*, volume 5478 of *Lecture Notes in Computer Science*, pages 562–569. Springer, 2009. 4, 18

Diana Santos and Nuno Cardoso. GikiP: evaluating geographical answers from wikipedia. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 59–60, New York, NY, USA, 2008. ACM. doi: http://doi.acm.org/10.1145/1460007.1460024. 32

Diana Santos, Nuno Cardoso, and Luís Miguel Cabral. How geographic was GikiCLEF?: a GIR-critical review. In *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 1–2, New York, NY, USA, 2010. ACM. doi: http://doi.acm.org/10.1145/1722080.1722110. 33

Steven Schockaert and Martine De Cock. Neighborhood Restrictions in Geographic IR. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: http://doi.acm.org/10.1145/1277741.1277772. 119

David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, volume 2163 of *Lecture Notes in Computer Science*, pages 127–137. Springer, Berlin, 2001. 2, 5, 59, 71

David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In *HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 45–49, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1119394.1119401. 60, 61

Nicola Stokes, Yi Li, Alistair Moffat, and Jiawen Rong. An empirical study of the effects of nlp components on geographic ir performance. *International Journal of Geographical Information Science*, 22(3):247–264, 2008. 13, 16, 87, 88

BIBLIOGRAPHY

Christopher Stokoe, Michael P. Oakes, and John Tait. Word Sense Disambiguation in Information Retrieval revisited. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166, New York, NY, USA, 2003. ACM. doi: 10.1145/860435.860466. 87

Strabo. *The Geography*, volume I of *Loeb Classical Library*. Harvard University Press, 1917. `http://penelope.uchicago.edu/Thayer/E/Roman/Texts/Strabo/home.html`. 1

Jiayu Tang and Mark Sanderson. Spatial Diversity, Do Users Appreciate It? In *GIR10 Workshop*, 2010. 18

Jordi Turmo, Pere R. Comas, Sophie Rosset, Olivier Galibert, Nicolas Moreau, Djamel Mostefa, Paolo Rosso, and Davide Buscaldi. Overview of QAST 2009. In *CLEF 2009 Working notes*, 2009. 31

Florian A. Twaroch and Christopher B. Jones. A web platform for the evaluation of vernacular place names in automatically constructed gazetteers. In *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 1–2, New York, NY, USA, 2010. ACM. doi: http://doi.acm.org/10.1145/1722080.1722098. 119

Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual Indexing for Geographical Search on the Web. In Claudia Bauzer Medeiros, Max J. Egenhofer, and Elisa Bertino, editors, *SSTD*, volume 3633 of *Lecture Notes in Computer Science*, pages 218–235. Springer, 2005. 120

J.L. Vicedo. A semantic approach to question answering systems. In *Proceedings of Text Retrieval Conference (TREC-9)*, pages 440–445. NIST, 2000. 105

Ellen M. Voorhees. The TREC-8 Question Answering Track Report. In *Proceedings of the 8th Text Retrieval Conference (TREC)*, pages 77–82, 1999. 23

Ian H. Witten, Timothy C. Bell, and Craig G. Neville. Indexing and Compressing Full-Text Databases for CD-ROM. *J. Information Science*, 17:265–271, 1992. 10

Ludwig Wittgenstein. *Tractatus logico-philosophicus*. Routledge and Kegan Paul, London, England, 1961. The German text of Ludwig Wittgenstein's Logisch-philosophische Abhandlung translated by D.F. Pears and B.F. McGuinness and with an introduction by Bertrand Russell. 1

Allison Woodruff and Christian Plaunt. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, 45(9):645–655, 1994. 59

George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949. 78

Appendix A

Data Fusion for GIR

In this chapter are included some data fusion experiments that I carried out in order to combine the output of different GIR systems. Data fusion is the combination of retrieval results obtained by means of different strategies into one single output result set. The experiments were carried out within the TextMess project, in cooperation with the Universitat Politécnica de Catalunya (UPC) and the University of Jaén. The GIR systems combined were GeoTALP of the UPC, SINAI-GIR of the University of Jaén and our system, GeoWorSE. A system based on the fusion of results of the UPV and Jaén systems participated in the last edition of GeoCLEF (2008), obtaining the second best result (Mandl et al. (2008)).

A.1 The SINAI-GIR System

The SINAI-GIR system (Perea et al. (2007)) is composed of the following subsystems: the *Collection Preprocessing subsystem*, the *Query Analyzer*, the *Information Retrieval subsystem* and the *Validator*. Each query is preprocessed and analyzed by the *Query Analyzer*, identifying its geo-entities and spatial relations and making use of the Geonames gazetteer. This module also applies *query reformulation*, generating several independent queries which will be indexed and searched by means of the IR subsystem. The collection is pre-processed by the *Collection Preprocessing* module and finally the documents retrieved by the IR subsystem are filtered and re-ranked by means of the *Validator* subsystem.

The features of each subsystem are:

- *Collection Preprocessing Subsystem.* During the collection preprocessing, two indexes are generated (*locations* and *keywords* indexes). The Porter *stemmer*,

the Brill POS tagger and the LingPipe Named Entity Recognizer (NER) are used in this phase. English stop-words are also discarded.

- *Query Analyzer.* It is responsible for the preprocessing of English queries as well as the generation of different query reformulations.
- *Information Retrieval Subsystem.* Lemur[1] is used as IR engine.
- *Validator.* The aim of this subsystem is to filter the lists of documents recovered by the IR subsystem, establishing which of them are valid, depending on the locations and the *geo-relations* detected in the query. Another important function is to establish the final ranking of documents, based on manual rules and predefined weights.

A.2 The TALP GeoIR system

The TALP GeoIR system (Ferrés and Rodríguez (2008)) has five phases performed sequentially: collection processing and indexing, linguistic and geographical analysis of the topics, textual IR with Terrier[2], Geographical Retrieval with Geographical Knowledge Bases (GKBs), and geographical document re-ranking.

The collection is processed and indexed in two different indexes: a geographical index with geographical information extracted from the documents and enriched with the aid of GKBs and a textual index with the lemmatized content of the documents.

The linguistic analysis uses the following Natural Language Processing tools: *TnT*, a statistical POS tagger, the WordNet 2.0 lemmatizer, and a in-house *Maximum Entropy-based NERC* system, trained with the CoNLL-2003 shared task English data set. The geographical analysis is based on a Geographical Thesaurus that uses the classes of the *ADL Feature Type Thesaurus* and includes four gazetteers: *GEOnet Names Server (GNS)*, *Geographic Names Information System (GNIS)*, *GeoWorldMap*, and a subset of *World Gazetter*[3].

The retrieval system is a textual IR system based on Terrier Ounis et al. (2006). Terrier configuration includes a TF-IDF schema, lemmatized query topics, Porter Stemmer, and Relevance Feedback using 10 top documents and 40 top terms.

The Geographical Retrieval uses geographical terms and/or geographical feature types appearing in the topics to retrieve documents from the geographical index. The

[1] `http://www.lemurproject.org`
[2] `http://ir.dcs.gla.ac.uk/terrier/`
[3] `http://world-gazetteer.com/`

geographical search allows to retrieve documents with geographical terms that are included in the sub-ontological path of the query terms (e.g. documents containing *Alaska* are retrieved from a query *United States*).

Finally, a geographical re-ranking is performed using the set of documents retrieved by Terrier. From this set of documents those that have been also retrieved in the Geographical Retrieval set are re-ranked giving them more weight than the other ones.

The system is composed of five modules that work sequentially:

1. a Linguistic and Geographical analysis module;
2. a thematic Document Retrieval module based on Terrier;
3. a Geographical Retrieval module that uses Geographical Knowledge Bases (GKBs);
4. a Document Filtering module.

The analysis module extracts relevant keywords from the topics, including geographical names, with the help of gazetteers.

The Document Retrieval module uses Terrier over a lemmatized index of the document collections and retrieves bthe relevant documents using the whole content of the tags previously lemmatized. The weighting scheme used for terrier is *tf-idf*.

The geographical retrieval module retrieves all the documents that have a token that matches totally or partially (a sub-path) the geographical keyword. As an example, the keyword *America@@Northern America@@United States* will retrieve all places in the U.S.

The Document Filtering module creates the output document list of the system, by joining the documents retrieved by Terrier with the ones retrieved by the Geographical Document Retrieval module. If the set of selected documents is less than 1,000, the top-scored documents of Terrier are selected with a lower priority than the previous ones. When the system uses only Terrier for retrieval, it returns the first $1,000$ top-scored documents by Terrier.

A.3 Data Fusion using Fuzzy Borda

In the classical (discrete) Borda count, each expert gives a mark to each alternative. The mark is given by the number of alternatives worse than it. The fuzzy variant, introduced by Nurmi (2001), allows the experts to show numerically how much alternatives are preferred over others, expressing their preference intensities from 0 to 1.

Let $R^1, R^2, \ldots, R^m$ be the fuzzy preference relations of m experts over n alternatives $x_1, x_2, \ldots, x_n$. Each expert k expresses its preferences by means of a matrix of preference intensities:

$$R^k = \begin{pmatrix} r_{11}^k & r_{12}^k & \ldots & r_{1n}^k \\ r_{21}^k & r_{22}^k & \ldots & r_{2n}^k \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1}^k & r_{n2}^k & \ldots & r_{nn}^k \end{pmatrix} \tag{A.1}$$

where each $r_{ij}^k = \mu_{R^k}(x_i, x_j)$, with $\mu_{R^k} : X \times X \to [0, 1]$ is the membership function of R^k. The number $r_{ij}^k \in [0, 1]$ is considered as the degree of confidence with which the expert k prefers x_i over x_j. The final value assigned by the expert k to each alternative x_i is the sum by row of the entries greater than 0.5 in the preference matrix, or, formally:

$$r_k(x_i) = \sum_{j=1, r_{ij}^k > 0.5}^{n} r_{ij}^k \tag{A.2}$$

The threshold 0.5 ensures that the relation R^k is an ordinary preference relation.

The fuzzy Borda count for an alternative x_i is obtained as the sum of the values assigned by each expert to that alternative:

$$\mathbf{r}(x_i) = \sum_{k=1}^{m} r_k(x_i) \tag{A.3}$$

For instance, consider two experts with the following preferences matrices:

$$R^1 = \begin{pmatrix} 0 & 0.8 & 0.9 \\ 0.2 & 0 & 0.6 \\ 0.1 & 0 & 0 \end{pmatrix}, R^2 = \begin{pmatrix} 0 & 0.4 & 0.3 \\ 0.6 & 0 & 0.6 \\ 0.7 & 0.4 & 0 \end{pmatrix}$$

This would correspond to the discrete preference matrices:

$$R^1 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, R^2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

In the discrete case, the winner would be x_2, the second option: $\mathbf{r}(x_1) = 2$, $\mathbf{r}(x_2) = 3$ and $\mathbf{r}(x_3) = 1$. But in the fuzzy case, the winner would be x_1: $\mathbf{r}(x_1) = 1.7$, $\mathbf{r}(x_2) = 1.2$ and $\mathbf{r}(x_3) = 0.7$, because the first expert was more confident about his ranking.

In our approach each system is an expert: therefore, for m systems, there are m preference matrices for each topic (query). The size of these matrices is variable: the reason is that the retrieved document list is not the same for all the systems. The

size of a preference matrix is $N_t \times N_t$, where N_t is the number of unique documents retrieved by the systems (i.e. the number of documents that appear at least in one of the lists returned by the systems) for topic t.

Each system may rank the documents using weights that are not in the same range of the other ones. Therefore, the output weights $w_1, w_2, \ldots, w_n$ of each expert k are transformed to fuzzy confidence values by means of the following transformation:

$$r_{ij}^k = \frac{w_i}{w_i + w_j} \tag{A.4}$$

This transformation ensures that the preference values are in the range $[0, 1]$. In order to adapt the fuzzy Borda count to the merging of the results of IR systems, we have to determine the preference values in all the cases where one of the systems does not retrieve a document that has been retrieved by another one. Therefore, matrices are extended in a way of covering the union of all the documents retrieved by every system. The preference values of the documents that occur in another list but not in the list retrieved by system k are set to 0.5, corresponding to the idea that the expert is presented with an option on which it cannot express a preference.

A.4 Experiments and Results

In Tables A.1 and A.2 we show the detail of each run in terms of the component systems and the topic fields used. "Official" runs (i.e., the ones submitted to GeoCLEF) are labeled with `TMESS02-08` and `TMESS07A`.

In order to evaluate the contribution of each system to the final result, we calculated the overlap rate O of the documents retrieved by the systems: $O = \frac{|D_1 \cap \ldots \cap D_m|}{|D_1 \cup \ldots \cup D_m|}$, where m is the number of systems that have been combined together and $D_i, 0 < i \leq m$ is the set of documents retrieved by the i-th system. The obtained value measures how different are the sets of documents retrieved by each system.

The R-overlap and N-overlap coefficients, based on the Dice similarity measure, were introduced by Lee (1997) to calculate the degree of overlap of relevant and non-relevant documents in the results of different systems. R-overlap is defined as $R_{overlap} = \frac{m \cdot |R_1 \cap \ldots \cap R_m|}{|R_1| + \ldots + |R_m|}$, where $R_i, 0 < i \leq m$ is the set of relevant documents retrieved by the system i. N-overlap is calculated in the same way, where each R_i has been substituted by N_i, the set of the *non*-relevant documents retrieved by the system i. $R_{overlap}$ is 1 if all systems return the same set of relevant documents, 0 if they return different sets of relevant documents. $N_{overlap}$ is 1 if the systems retrieve an identical set of non-relevant documents, and 0 if the non-relevant documents are different for each system.

Table A.1: Description of the runs of each system.

run ID	description
NLEL	
NLEL0802	base system (only text index, no wordnet, no map filtering)
NLEL0803	2007 system (no map filtering)
NLEL0804	base system, title and description only
NLEL0505	2008 system, all indices and map filtering enabled
NLEL01	complete 2008 system, title and description
SINAI	
SINAI1	base system, title and description only
SINAI2	base system, all fields
SINAI4	filtering system, title and description only
SINAI5	filtering system (rule-based)
TALP	
TALP01	system without GeoKB, title and description only

Table A.2: Details of the composition of all the evaluated runs.

run ID	fields	NLEL run ID	SINAI run ID	TALP run ID
Officially evaluated runs				
TMESS02	TDN	NLEL0802	SINAI2	
TMESS03	TDN	NLEL0802	SINAI5	
TMESS05	TDN	NLEL0803	SINAI2	
TMESS06	TDN	NLEL0803	SINAI5	
TMESS07A	TD	NLEL0804	SINAI1	
TMESS08	TDN	NLEL0505	SINAI5	
Non-official runs				
TMESS10	TD		SINAI1	TALP01
TMESS11	TD	NLEL01	SINAI1	
TMESS12	TD	NLEL01		TALP01
TMESS13	TD	NLEL0804		TALP01
TMESS14	TD	NLEL0804	SINAI1	TALP01
TMESS15	TD	NLEL01	SINAI1	TALP01

Lee (1997) observed that different runs are usually identified by a low $N_{overlap}$ value, independently from the $R_{overlap}$ value.

In Table A.3 we show the Mean Average Precision (MAP) obtained for each run and its composing runs, together with the average MAP calculated over the composing runs.

Table A.3: Results obtained for the various system combinations with the basic fuzzy Borda method.

run ID	$MAP_{combined}$	MAP_{NLEL}	MAP_{SINAI}	MAP_{TALP}	avg. MAP
TMESS02	0.228	0.201	0.226		0.213
TMESS03	0.216	0.201	0.212		0.206
TMESS05	0.236	0.216	0.226		0.221
TMESS06	0.231	0.216	0.212		0.214
TMESS07A	0.290	0.256	0.284		0.270
TMESS08	0.221	0.203	0.212		0.207
TMESS10	0.291		0.284	0.280	0.282
TMESS11	0.298	0.254		0.280	0.267
TMESS12	0.286	0.254	0.284		0.269
TMESS13	0.271	0.256		0.280	0.268
TMESS14	0.287	0.256	0.284	0.280	0.273
TMESS15	0.291	0.254	0.284	0.280	0.273

The results in Table A.4 show that the fuzzy Borda merging method always allows to improve the average of the results of the components, and only in one case it cannot improve the best component result (TMESS13). The results also show that the results with $MAP \geq 0.271$ were obtained for combinations with $R_overlap \geq 0.75$, indicating that the Chorus Effect plays an important part in the fuzzy Borda method. In order to better understand this result, we calculated the results that would have been obtained by calculating the fusion over different configurations of each group's system. These results are shown in Table A.5.

The fuzzy Borda method, as shown in Table A.5, when applied to different configurations of the same system, results also in an improvement of accuracy with respect to the results of the component runs. O, $R_{overlap}$ and $N_{overlap}$ values for same-group fusions are well above the O values obtained in the case of different systems (more than 0.73, while the values observed in Table A.4 are in the range $0.31 - 0.47$). However, the obtained results show that the method is not able to combine in an optimal way

Table A.4: O, $R_{overlap}$, $N_{overlap}$ coefficients, difference from the best system (*diff. best*) and difference from the average of the systems (*diff. avg.*) for all runs.

run ID	$MAP_{combined}$	diff. best	diff. avg.	O	$R_{overlap}$	$N_{overlap}$
TMESS02	0.228	0.002	0.014	0.346	0.692	0.496
TMESS03	0.216	0.004	0.009	0.317	0.693	0.465
TMESS05	0.236	0.010	0.015	0.358	0.692	0.508
TMESS06	0.231	**0.015**	0.017	0.334	0.693	0.484
TMESS07A	0.290	0.006	**0.020**	0.356	0.775	0.563
TMESS08	0.221	0.009	0.014	0.326	0.690	0.475
TMESS10	0.291	0.007	0.009	0.485	**0.854**	0.625
TMESS11	**0.298**	0.018	0.031	0.453	0.759	0.621
TMESS12	0.286	0.002	0.017	0.356	0.822	**0.356**
TMESS13	0.271	−0.009	0.003	0.475	0.796	0.626
TMESS14	0.287	0.003	0.013	0.284	0.751	0.429
TMESS15	0.291	0.007	0.019	0.277	0.790	0.429

Table A.5: Results obtained with the fusion of systems from the same participant. M_1: MAP of the system in the first configuration, M_2: MAP of the system in the second configuration.

run ID	$MAP_{combined}$	M_1	M_2	O	$R_{overlap}$	$N_{overlap}$
SINAI1+SINAI4	0.288	0.284	0.275	0.792	0.904	0.852
NLEL0804+NLEL01	0.265	0.254	0.256	0.736	0.850	0.828
TALP01+TALP02	0.285	0.280	0.272	0.792	0.904	0.852

the systems that return different sets of relevant document (i.e., when we are in presence of the Skimming Effect). This is due to the fact that a relevant document that is retrieved by system A and not by system B has a 0.5 weight in the preference matrix of B, making that its ranking will be worse than any non-relevant document retrieved by B and ranked better than the worst document.

Appendix B

GeoCLEF Topics

B.1 GeoCLEF 2005

```
<topics>
<top>
<num> GC001 </num>
<title> Shark Attacks off Australia and California </title>
<desc> Documents will report any information relating to shark
attacks on humans. </desc>
<narr> Identify instances where a human was attacked by a shark,
including where the attack took place and the circumstances
surrounding the attack. Only documents concerning specific attacks
are relevant; unconfirmed shark attacks or suspected bites are not
relevant. </narr>
</top>

<top>
<num> GC002 </num>
<title> Vegetable Exporters of Europe </title>
<desc> What countries are exporters of fresh, dried or frozen
vegetables? </desc>
<narr> Any report that identifies a country or territory that
exports fresh, dried or frozen vegetables, or indicates the country
of origin of imported vegetables is relevant. Reports regarding
canned vegetables, vegetable juices or otherwise processed
vegetables are not relevant. </narr>
</top>

<top>
<num> GC003 </num>
<title> AI in Latin America </title>
<desc> Amnesty International reports on human rights in Latin
America. </desc>
<narr> Relevant documents should inform readers about Amnesty
International reports regarding human rights in Latin America, or on reactions
```

```
to these reports. </narr>
</top>

<top>
<num> GC004 </num>
<title> Actions against the fur industry in Europe and the U.S.A. </title>
<desc> Find information on protests or violent acts against the fur
industry.
</desc>
<narr> Relevant documents describe measures taken by animal right
activists against fur farming and/or fur commerce, e.g. shops selling items in
fur. Articles reporting actions taken against people wearing furs are also of
importance. </narr>
</top>

<top>
<num> GC005 </num>
<title> Japanese Rice Imports </title>
<desc> Find documents discussing reasons for and consequences of the
first imported rice in Japan. </desc>
<narr> In 1994, Japan decided to open the national rice market for
the first time to other countries. Relevant documents will comment on this
question. The discussion can include the names of the countries from which the
rice is imported, the types of rice, and the controversy that this decision
prompted in Japan. </narr>
</top>

<top>
<num> GC006 </num>
<title> Oil Accidents and Birds in Europe </title>
<desc> Find documents describing damage or injury to birds caused by
accidental oil spills or pollution. </desc>
<narr> All documents which mention birds suffering because of oil accidents
are relevant. Accounts of damage caused as a result of bilge discharges or oil
dumping are not relevant. </narr>
</top>

<top>
<num> GC007 </num>
<title> Trade Unions in Europe </title>
<desc> What are the differences in the role and importance of trade
unions between European countries? </desc>
<narr> Relevant documents must compare the role, status or importance
of trade unions between two or more European countries. Pertinent
information will include level of organisation, wage negotiation mechanisms, and
the general climate of the labour market. </narr>
</top>

<top>
<num> GC008 </num>
<title> Milk Consumption in Europe </title>
<desc> Provide statistics or information concerning milk consumption
```

```
in European countries. </desc>
<narr> Relevant documents must provide statistics or other information about
milk consumption in Europe, or in single European nations. Reports on milk
derivatives are not relevant. </narr>
</top>

<top>
<num> GC009 </num>
<title> Child Labor in Asia </title>
<desc> Find documents that discuss child labor in Asia and proposals to
eliminate it or to improve working conditions for children. </desc>
<narr> Documents discussing child labor in particular countries in
Asia, descriptions of working conditions for children, and proposals of
measures to eliminate child labor are all relevant. </narr>
</top>

<top>
<num> GC010 </num>
<title> Flooding in Holland and Germany </title>
<desc> Find statistics on flood disasters in Holland and Germany in
1995.
</desc>
<narr> Relevant documents will quantify the effects of the damage
caused by flooding that took place in Germany and the Netherlands in 1995 in
terms of numbers of people and animals evacuated and/or of economic losses.
</narr>
</top>

<top>
<num> GC011 </num>
<title> Roman cities in the UK and Germany </title>
<desc> Roman cities in the UK and Germany. </desc>
<narr> A relevant document will identify one or more cities in the United
Kingdom or Germany which were also cities in Roman times. </narr>
</top>

<top>
<num> GC012 </num>
<title> Cathedrals in Europe </title>
<desc> Find stories about particular cathedrals in Europe, including the
United Kingdom and Russia. </desc>
<narr> In order to be relevant, a story must be about or describe a
particular cathedral in a particular country or place within a country in
Europe, the UK or Russia.  Not relevant are stories which are generally
about tourist tours of cathedrals or about the funeral of a particular
person in a cathedral. </narr>
</top>

<top>
<num> GC013 </num>
<title> Visits of the American president to Germany </title>
<desc> Find articles about visits of President Clinton to Germany.
```

```
</desc>
<narr>
Relevant documents should describe the stay of President Clinton in Germany
not purely the status of American-German relations. </narr>
</top>

<top>
<num> GC014 </num>
<title> Environmentally hazardous Incidents in the North Sea </title>
<desc> Find documents about environmental accidents and hazards in
the North Sea region. </desc>
<narr>
Relevant documents will describe accidents and environmentally hazardous
actions in or around the North Sea. Documents about oil production
can be included if they describe environmental impacts. </narr>
</top>

<top>
<num> GC015 </num>
<title> Consequences of the genocide in Rwanda </title>
<desc> Find documents about genocide in Rwanda and its impacts. </desc>
<narr>
Relevant documents will describe the country's situation after the
genocide and the political, economic and other efforts involved in attempting
to stabilize the country. </narr>
</top>

<top>
<num> GC016 </num>
<title> Oil prospecting and ecological problems in Siberia
and the Caspian Sea </title>
<desc> Find documents about Oil or petroleum development and related
ecological problems in Siberia and the Caspian Sea regions. </desc>
<narr>
Relevant documents will discuss the exploration for, and exploitation of
petroleum (oil) resources in the Russian region of Siberia and in or near
the Caspian Sea. Relevant documents will also discuss ecological issues or
problems, including disasters or accidents in these regions. </narr>
</top>

<top>
<num> GC017 </num>
<title> American Troops in Sarajevo, Bosnia-Herzegovina </title>
<desc> Find documents about American troop deployment in Bosnia-Herzegovina,
especially Sarajevo. </desc>
<narr>
Relevant documents will discuss deployment of American (USA) troops as
part of the UN peacekeeping force in the former Yugoslavian regions of
Bosnia-Herzegovina, and in particular in the city of Sarajevo. </narr>
</top>

<top>
```

<num> GC018 </num>
<title> Walking holidays in Scotland </title>
<desc> Find documents that describe locations for walking holidays in
Scotland. </desc>
<narr> A relevant document will describe a place or places within Scotland where
a walking holiday could take place. </narr>
</top>

<top>
<num> GC019 </num>
<title> Golf tournaments in Europe </title>
<desc> Find information about golf tournaments held in European locations. </desc>
<narr> A relevant document will describe the planning, running and/or results of
a golf tournament held at a location in Europe. </narr>
</top>

<top>
<num> GC020 </num>
<title> Wind power in the Scottish Islands </title>
<desc> Find documents on electrical power generation using wind power
in the islands of Scotland. </desc>
<narr> A relevant document will describe wind power-based electricity generation
schemes providing electricity for the islands of Scotland. </narr>
</top>

<top>
<num> GC021 </num>
<title> Sea rescue in North Sea </title>
<desc> Find items about rescues in the North Sea. </desc>
<narr> A relevant document will report a sea rescue undertaken in North Sea. </narr>
</top>

<top>
<num> GC022 </num>
<title> Restored buildings in Southern Scotland </title>
<desc> Find articles on the restoration of historic buildings in
the southern part of Scotland. </desc>
<narr> A relevant document will describe a restoration of historical buildings
in the southern Scotland. </narr>
</top>

<top>
<num> GC023 </num>
<title> Murders and violence in South-West Scotland </title>
<desc> Find articles on violent acts including murders in the South West
part of Scotland. </desc>
<narr> A relevant document will give details of either specific acts of violence
or death related to murder or information about the general state of violence in
South West Scotland. This includes information about violence in places such as
Ayr, Campeltown, Douglas and Glasgow. </narr>
</top>

```
<top>
<num> GC024 </num>
<title> Factors influencing tourist industry in Scottish Highlands </title>
<desc> Find articles on the tourism industry in the Highlands of Scotland
and the factors affecting it. </desc>
<narr> A relevant document will provide information on factors which have
affected or influenced tourism in the Scottish Highlands. For example, the
construction of roads or railways, initiatives to increase tourism, the planning
and construction of new attractions and influences from the environment (e.g.
poor weather). </narr>
</top>


<top>
<num> GC025 </num>
<title> Environmental concerns in and around the Scottish Trossachs </title>
<desc> Find articles about environmental issues and concerns in
the Trossachs region of Scotland. </desc>
<narr> A relevant document will describe environmental concerns (e.g. pollution,
damage to the environment from tourism) in and around the area in Scotland known
as the Trossachs. Strictly speaking, the Trossachs is the narrow wooded glen
between Loch Katrine and Loch Achray, but the name is now used to describe a
much larger area between Argyll and Perthshire, stretching north from the
Campsies and west from Callander to the eastern shore of Loch Lomond. </narr>
</top>

</topics>
```

B.2 GeoCLEF 2006

```
<GeoCLEF-2006-topics-English>
<top>
<num>GC026</num>
<title>Wine regions around rivers in Europe</title>
<desc>Documents about wine regions along the banks of European rivers</desc>
<narr>Relevant documents describe a wine region along a major river in
European countries. To be relevant the document must name the region and the river.</narr>
</top>
<top>
<num>GC027</num>
<title>Cities within 100km of Frankfurt</title>
<desc>Documents about cities within 100 kilometers of the city of Frankfurt in
Western Germany</desc>
<narr>Relevant documents discuss cities within 100 kilometers of Frankfurt am
Main Germany, latitude 50.11222, longitude 8.68194.  To be relevant the document
must describe the city or an event in that city. Stories about Frankfurt itself
are not relevant</narr>
</top>
<top>
```

```
<num>GC028</num>
<title>Snowstorms in North America</title>
<desc>Documents about snowstorms occurring in the north part of the American
continent</desc>
<narr>Relevant documents state cases of snowstorms and their effects in North
America.  Countries are Canada, United States of America and Mexico.  Documents
about other kinds of storms are not relevant (e.g. rainstorm, thunderstorm,
electric storm, windstorm)</narr>
</top>
<top>
<num>GC029</num>
<title>Diamond trade in Angola and South Africa</title>
<desc>Documents regarding diamond trade in Angola and South Africa</desc>
<narr>Relevant documents are about diamond trading in these two countries and
its consequences (e.g. smuggling, economic and political instability)</narr>
</top>
<top>
<num>GC030</num>
<title>Car bombings near Madrid</title>
<desc>Documents about car bombings occurring near Madrid</desc>
<narr>Relevant documents treat cases of car bombings occurring in the capital of
Spain and its outskirts</narr>
</top>
<top>
<num>GC031</num>
<title>Combats and embargo in the northern part of Iraq</title>
<desc>Documents telling about combats or embargo in the northern part of
Iraq</desc>
<narr>Relevant documents are about combats and effects of the 90s embargo in the
northern part of Iraq. Documents about these facts happening in other parts of
Iraq are not relevant</narr>
</top>
<top>
<num>GC032</num>
<title>Independence movement in Quebec</title>
<desc>Documents about actions in Quebec for the independence of this Canadian
province</desc>
<narr>Relevant documents treat matters related to Quebec independence movement
(e.g. referendums) which take place in Quebec</narr>
</top>
<top>
<num>GC033</num>
<title> International sports competitions in the Ruhr area</title>
<desc> World Championships and international tournaments in
the Ruhr area</desc>
<narr> Relevant documents state the type or name of the competition,
the city and possibly results. Irrelevant are documents where only part of the
competition takes place in the Ruhr area of Germany, e.g. Tour de France,
Champions League or UEFA-Cup games.</narr>
</top>
<top>
<num> GC034 </num>
```

<title> Malaria in the tropics </title>
<desc> Malaria outbreaks in tropical regions and preventive
vaccination </desc>
<narr> Relevant documents state cases of malaria in tropical regions
and possible preventive measures like chances to vaccinate against the
disease. Outbreaks must be of epidemic scope. Tropics are defined as the region
between the Tropic of Capricorn, latitude 23.5 degrees South and the Tropic of
Cancer, latitude 23.5 degrees North. Not relevant are documents about a single
person's infection. </narr>
</top>
<top>
<num> GC035 </num>
<title> Credits to the former Eastern Bloc </title>
<desc> Financial aid in form of credits by the International
Monetary Fund or the World Bank to countries formerly belonging to
the Eastern Bloc aka the Warsaw Pact, except the republics of the former
USSR</desc>
<narr> Relevant documents cite agreements on credits, conditions or
consequences of these loans. The Eastern Bloc is defined as countries
under strong Soviet influence (so synonymous with Warsaw Pact) throughout
the whole Cold War. Excluded are former USSR republics. Thus the countries
are Bulgaria, Hungary, Czech Republic, Slovakia, Poland and Romania. Thus not
all communist or socialist countries are considered relevant.</narr>
</top>
<top>
<num> GC036 </num>
<title> Automotive industry around the Sea of Japan </title>
<desc> Coastal cities on the Sea of Japan with automotive industry or
factories </desc>
<narr> Relevant documents report on automotive industry or factories in
cities on the shore of the Sea of Japan (also named East Sea (of Korea))
including economic or social events happening there like planned joint-ventures
or strikes. In addition to Japan, the countries of North Korea, South Korea and
Russia are also on the Sea of Japan.</narr>
</top>
<top>
<num> GC037 </num>
<title> Archeology in the Middle East </title>
<desc> Excavations and archeological finds in the Middle East
</desc>
<narr> Relevant documents report recent finds in some town, city, region or
country of the Middle East, i.e. in Iran, Iraq, Turkey, Egypt, Lebanon, Saudi
Arabia, Jordan, Yemen, Qatar, Kuwait, Bahrain, Israel, Oman, Syria, United Arab
Emirates, Cyprus, West Bank, or the Gaza Strip</narr>
</top>
<top>
<num> GC038 </num>
<title> Solar or lunar eclipse in Southeast Asia </title>
<desc> Total or partial solar or lunar eclipses in Southeast Asia
</desc>
<narr> Relevant documents state the type of eclipse and the region or country
of occurrence, possibly also stories about people travelling to see it.

```
Countries of Southeast Asia are Brunei, Cambodia, East Timor, Indonesia, Laos,
Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam.
</narr>
</top>
<top>
<num> GC039 </num>
<title> Russian troops in the southern Caucasus </title>
<desc> Russian soldiers, armies or military bases in the Caucasus region
south of the Caucasus Mountains </desc>
<narr> Relevant documents report on Russian troops based at, moved to or
removed from the region. Also agreements on one of these actions or combats
are relevant. Relevant countries are: Azerbaijan, Armenia, Georgia, Ossetia,
Nagorno-Karabakh. Irrelevant are documents citing actions between troops of
nationality different from Russian (with Russian mediation between the two.)
</narr>
</top>
<top>
<num> GC040 </num>
<title> Cities near active volcanoes </title>
<desc> Cities, towns or villages threatened by the eruption of a volcano
</desc>
<narr> Relevant documents cite the name of the cities, towns, villages that
are near an active volcano which recently had an eruption or could erupt soon.
Irrelevant are reports which do not state the danger (i.e. for example necessary
preventive evacuations) or the consequences for specific cities , but just
tell that a particular volcano (in some country) is going to erupt, has erupted
or that a region has active volcanoes. </narr>
</top>

<top>
<num>GC041</num>
<title>Shipwrecks in the Atlantic Ocean</title>
<desc>Documents about shipwrecks in the Atlantic Ocean</desc>
<narr>Relevant documents should document shipwreckings in any part of the
Atlantic Ocean or its coasts.</narr>
</top>

<top>
<num>GC042</num>
<title>Regional elections in Northern Germany</title>
<desc>Documents about regional elections in Northern Germany</desc>
<narr>Relevant documents are those reporting the campaign or results for the
state parliaments of any of the regions of Northern Germany. The states of
northern Germany are commonly Bremen, Hamburg, Lower Saxony, Mecklenburg-Western
Pomerania and Schleswig-Holstein. Only regional elections are relevant;
municipal, national and European elections are not.</narr>
</top>

<top>
<num>GC043</num>
<title>Scientific research in New England Universities</title>
<desc>Documents about scientific research in New England universities</desc>
```

<narr>Valid documents should report specific scientific research or
breakthroughs occurring in universities of New England. Both current and past
research are relevant. Research regarded as bogus or fraudulent is also
relevant. New England states are: Connecticut, Rhode Island, Massachusetts,
Vermont, New Hampshire, Maine. </narr>
</top>

<top>
<num>GC044</num>
<title>Arms sales in former Yugoslavia</title>
<desc>Documents about arms sales in former Yugoslavia</desc>
<narr>Relevant documents should report on arms sales that took place in the
successor countries of the former Yugoslavia. These sales can be legal or not,
and to any kind of entity in these states, not only the government itself.
Relevant countries are: Slovenia, Macedonia, Croatia, Serbia and Montenegro, and
Bosnia and Herzegovina.
</narr>
</top>

<top>
<num>GC045</num>
<title>Tourism in Northeast Brazil</title>
<desc>Documents about tourism in Northeastern Brazil</desc>
<narr>Of interest are documents reporting on tourism in Northeastern Brazil,
including places of interest, the tourism industry and/or the reasons for taking
or not a holiday there. The states of northeast Brazil are Alagoas, Bahia,
Cear, Maranho, Paraba, Pernambuco, Piau, Rio Grande do Norte and
Sergipe.</narr>
</top>

<top>
<num>GC046</num>
<title>Forest fires in Northern Portugal</title>
<desc>Documents about forest fires in Northern Portugal</desc>
<narr>Documents should report the ocurrence, fight against, or aftermath of
forest fires in Northern Portugal. The regions covered are Minho, Douro
Litoral, Trs-os-Montes and Alto Douro, corresponding to the districts of Viana
do Castelo, Braga, Porto (or Oporto), Vila Real and Bragana.
</narr>
</top>

<top>
<num>GC047</num>
<title>Champions League games near the Mediterranean </title>
<desc>Documents about Champion League games played in European cities bordering
the Mediterranean </desc>
<narr>Relevant documents should include at least a short description of a
European Champions League game played in a European city bordering the
Mediterranean Sea or any of its minor seas. European countries along the
Mediterranean Sea are Spain, France, Monaco, Italy, the island state of Malta,
Slovenia, Croatia, Bosnia and Herzegovina, Serbia and Montenegro, Albania,
Greece, Turkey, and the island of Cyprus.</narr>

```
</top>

<top>
<num>GC048</num>
<title>Fishing in Newfoundland and Greenland</title>
<desc>Documents about fisheries around Newfoundland and Greenland</desc>
<narr>Relevant documents should document fisheries and economical, ecological or
legal problems associated with it, around Greenland and the Canadian island of
Newfoundland. </narr>
</top>

<top>
<num>GC049</num>
<title>ETA in France</title>
<desc>Documents about ETA activities in France</desc>
<narr>Relevant documents should document the activities of the Basque terrorist
group ETA in France, of a paramilitary, financial, political nature or others. </narr>
</top>

<top>
<num>GC050</num>
<title>Cities along the Danube and the Rhine</title>
<desc>Documents describe cities in the shadow of the Danube or the Rhine</desc>
<narr>Relevant documents should contain at least a short description of cities
through which the rivers Danube and Rhine pass, providing evidence for it. The
Danube flows through nine countries (Germany, Austria, Slovakia, Hungary,
Croatia, Serbia, Bulgaria, Romania, and Ukraine).  Countries along the Rhine are
Liechtenstein, Austria, Germany, France, the Netherlands and Switzerland. </narr>
</top>
</GeoCLEF-2006-topics-English>
```

B.3 GeoCLEF 2007

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<top lang="en">
<num>10.2452/51-GC</num>
<title>Oil and gas extraction found between the UK and the Continent</title>
<desc>To be relevant documents describing oil or gas production between the UK
and the European continent will be relevant</desc>
<narr>Oil and gas fields in the North Sea will be relevant.</narr>
</top>
<top lang="en">
<num>10.2452/52-GC</num>
<title>Crime near St Andrews</title>
<desc>To be relevant, documents must be about crimes occurring close to or in
St. Andrews.</desc>
<narr>Any event that refers to criminal dealings of some sort is relevant, from
thefts to corruption.</narr>
</top>
```

```
<top lang="en">
<num>10.2452/53-GC</num>
<title>Scientific research at east coast Scottish Universities</title>
<desc>For documents to be relevant, they must describe scientific research
conducted by a Scottish University located on the east coast of Scotland</desc>
<narr>Universities in Aberdeen, Dundee, St Andrews and Edinburgh wil be
considered relevant locations.</narr>
</top>
<top lang="en">
<num>10.2452/54-GC</num>
<title>Damage from acid rain in northern Europe</title>
<desc>Documents describing the damage caused by acid rain in the countries of
northern Europe</desc>
<narr>Relevant countries include Denmark, Estonia, Finland, Iceland, Republic of
Ireland, Latvia, Lithuania, Norway, Sweden, United Kingdom and northeastern
parts of Russia</narr>
</top>
<top lang="en">
<num>10.2452/55-GC</num>
<title>Deaths caused by avalanches occurring in Europe, but not in the
Alps</title>
<desc>To be relevant a document must describe the death of a person caused by an
avalanche that occurred away from the Alps but in Europe.</desc>
<narr>for example mountains in Scotland, Norway, Iceland</narr>
</top>
<top lang="en">
<num>10.2452/56-GC</num>
<title>Lakes with monsters</title>
<desc>To be relevant, the document must describe a lake where a monster is
supposed to exist.</desc>
<narr>The document must state the alledged existence of a monster in a
particular lake and must name the lake. Activities which try to prove the
existence of the monster and reports of witnesses who have seen the monster are
relevant. Documents which mention only the name of a particular monster are not
relevant.</narr>
</top>
<top lang="en">
<num>10.2452/57-GC</num>
<title>Whisky making in the Scottlsh Islands</title>
<desc>To be relevant, a document must describe a whisky made, or a whisky
distillery located, on a Scottish island.</desc>
<narr>Relevant islands are Islay, Skye, Orkney, Arran, Jura, Mull.&#13;
Relevant whiskys are Arran Single Malt; Highland Park Single Malt; Scapa; Isle
of Jura; Talisker; Tobermory;  Ledaig; Ardbeg; Bowmore; Bruichladdich;
Bunnahabhain; Caol Ila; Kilchoman; Lagavulin; Laphroaig</narr>
</top>
<top lang="en">
<num>10.2452/58-GC</num>
<title>Travel problems at major airports near to London</title>
<desc>To be relevant, documents must describe travel problems at one of the
major airports close to London.</desc>
<narr>Major airports to be listed include Heathrow, Gatwick, Luton, Stanstead
```

```
and London City airport.</narr>
</top>
<top lang="en">
<num>10.2452/59-GC</num>
<title>Meetings of the Andean Community of Nations (CAN)</title>
<desc>Find documents mentioning cities in on the meetings of the Andean
Community of Nations (CAN) took place</desc>
<narr>relevant documents mention cities in which meetings of the members of the
Andean Community of Nations (CAN - member states Bolivia, Columbia, Ecuador, Peru).</narr>
</top>
<top lang="en">
<num>10.2452/60-GC</num>
<title>Casualties in fights in Nagorno-Karabakh</title>
<desc>Documents reporting on casualties in the war in Nagorno-Karabakh</desc>
<narr>Relevant documents report of casualties during the war or in fights in the
Armenian enclave Nagorno-Karabakh</narr>
</top>
<top lang="en">
<num>10.2452/61-GC</num>
<title>Airplane crashes close to Russian cities</title>
<desc>Find documents mentioning airplane crashes close to Russian cities</desc>
<narr>Relevant documents report on airplane crashes in Russia. The location is
to be specified by the name of a city mentioned in the document.</narr>
</top>
<top lang="en">
<num>10.2452/62-GC</num>
<title>OSCE meetings in Eastern Europe</title>
<desc>Find documents in which Eastern European conference venues of the
Organization for Security and Co-operation in Europe (OSCE) are mentioned</desc>
<narr>Relevant documents report on OSCE meetings in Eastern Europe. Eastern
Europe includes Bulgaria, Poland, the Czech Republic, Slovakia, Hungary,
Romania, Ukraine, Belarus, Lithuania, Estonia, Latvia and the European part of
Russia.</narr>
</top>
<top lang="en">
<num>10.2452/63-GC</num>
<title>Water quality along coastlines of the Mediterranean Sea</title>
<desc>Find documents on the water quality at the coast of the Mediterranean
Sea</desc>
<narr>Relevant documents report on the water quality along the coast and
coastlines of the Mediterranean Sea. The coasts must be specified by their
names.</narr>
</top>
<top lang="en">
<num>10.2452/64-GC</num>
<title>Sport events in the french speaking part of Switzerland</title>
<desc>Find documents on sport events in the french speaking part of
Switzerland</desc>
<narr>Relevant documents report sport events in the french speaking part of
Switzerland. Events in cities like Lausanne, Geneva, Neuchtel and Fribourg are
relevant.</narr>
</top>
```

B. GEOCLEF TOPICS

```
<top lang="en">
<num>10.2452/65-GC</num>
<title>Free elections in Africa</title>
<desc>Documents mention free elections held in countries in Africa</desc>
<narr>Future elections or promises of free elections are not relevant</narr>
</top>
<top lang="en">
<num>10.2452/66-GC</num>
<title>Economy at the Bosphorus</title>
<desc>Documents on economic trends at the Bosphorus strait</desc>
<narr>Relevant documents report on economic trends and development in the
Bosphorus region close to Istanbul</narr>
</top>
<top lang="en">
<num>10.2452/67-GC</num>
<title>F1 circuits where Ayrton Senna competed in 1994</title>
<desc>Find documents that mention circuits where  the Brazilian driver Ayrton
Senna participated in 1994. The name and location of the circuit is
required</desc>
<narr>Documents should indicate that Ayrton Senna participated in a race in a
particular stadion, and the location of the race track.</narr>
</top>
<top lang="en">
<num>10.2452/68-GC</num>
<title>Rivers with floods</title>
<desc>Find documents that mention rivers that flooded. The name of the river is
required.</desc>
<narr>Documents that mention floods but fail to name the rivers are not
relevant.</narr>
</top>
<top lang="en">
<num>10.2452/69-GC</num>
<title>Death on the Himalaya</title>
<desc>Documents should mention deaths due to climbing mountains in the Himalaya
range.</desc>
<narr>Only death casualties of mountaineering athletes in the Himalayan
mountains, such as Mount Everest or Annapurna, are interesting. Other deaths,
caused by e.g. political unrest in the region, are irrelevant.</narr>
</top>
<top lang="en">
<num>10.2452/70-GC</num>
<title>Tourist attractions in Northern Italy</title>
<desc>Find documents that identify tourist attractions in the North of
Italy.</desc>
<narr>Documents should mention places of tourism in the North of Italy, either
specifying particular tourist attractions (and where they are located) or
mentioning that the place (town, beach, opera, etc.) attracts many
tourists.</narr>
</top>
<top lang="en">
<num>10.2452/71-GC</num>
<title>Social problems in greater Lisbon</title>
```

<desc>Find information about social problems afflicting places in greater Lisbon.</desc>
<narr>Documents are relevant if they mention any social problem, such as drug consumption, crime, poverty, slums, unemployment or lack of integration of minorities, either for the region as a whole or in specific areas inside it. Greater Lisbon includes the Amadora, Cascais, Lisboa, Loures, Mafra, Odivelas, Oeiras , Sintra and Vila Franca de Xira districts.</narr>
</top>
<top lang="en">
<num>10.2452/72-GC</num>
<title>Beaches with sharks</title>
<desc>Relevant documents should name beaches or coastlines where there is danger of shark attacks. Both particular attacks and the mention of danger are relevant, provided the place is mentioned.</desc>
<narr>Provided that a geographical location is given, it is sufficient that fear or danger of sharks is mentioned. No actual accidents need to be reported.</narr>
</top>
<top lang="en">
<num>10.2452/73-GC</num>
<title>Events at St. Paul's Cathedral</title>
<desc>Any event that happened at St. Paul's cathedral is relevant, from concerts, masses, ceremonies or even accidents or thefts.</desc>
<narr>Just the description of the church or its mention as a tourist attraction is not relevant. There are three relevant St. Paul's cathedrals for this topic: those of So Paulo, Rome and London.</narr>
</top>
<top lang="en">
<num>10.2452/74-GC</num>
<title>Ship traffic around the Portuguese islands</title>
<desc>Documents should mention ships or sea traffic connecting Madeira and the Azores to other places, and also connecting the several isles of each archipelago. All subjects, from wrecked ships, treasure finding, fishing, touristic tours to military actions, are relevant, except for historical narratives.</desc>
<narr>Documents have to mention that there is ship traffic connecting the isles to the continent (portuguese mainland), or between the several islands, or showing international traffic. Isles of Azores are: So Miguel, Santa Maria, Formigas, Terceira, Graciosa, So Jorge, Pico, Faial, Flores and Corvo. The Madeira islands are: Mardeira, Porto Santo, Desertas islets and Selvagens islets.</narr>
</top>
<top lang="en">
<num>10.2452/75-GC</num>
<title>Violation of human rights in Burma</title>
<desc>Documents are relevant if they mention actual violation of human rights in Myanmar, previously named Burma.</desc>
<narr>This includes all reported violations of human rights in Burma, no matter when (not only by the present government). Declarations (accusations or denials) about the matter only, are not relevant.</narr>
</top>
</topics>

B.4 GeoCLEF 2008

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topics>
<topic lang="en">
<identifier>10.2452/76-GC</identifier>
<title>Riots in South American prisons</title>
<description>Documents mentioning riots in prisons in South
America</description>
<narrative>Relevant documents mention riots or uprising on the South American
continent. Countries in South America include Argentina, Bolivia, Brazil, Chile,
Suriname, Ecuador, Colombia, Guyana, Peru, Paraguay, Uruguay and Venezuela.
French Guiana is a French province in South America.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/77-GC</identifier>
<title>Nobel prize winners from Northern European countries</title>
<description>Documents mentioning Noble prize winners born in a Northern
European country.</description>
<narrative>Relevant documents contain information about the field of research
and the country of origin of the prize winner. Northern European countries are:
Denmark, Finland, Iceland, Norway, Sweden, Estonia, Latvia, Belgium, the
Netherlands, Luxembourg, Ireland, Lithuania, and the UK. The north of Germany
and Poland as well as the north-east of Russia also belong to Northern
Europe.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/78-GC</identifier>
<title>Sport events in the Sahara</title>
<description>Documents mentioning sport events occurring in (or passing through)
the Sahara.</description>
<narrative>Relevant documents must make reference to athletic events and to the
place where they take place. The Sahara covers huge parts of Algeria, Chad,
Egypt, Libya, Mali, Mauritania, Morocco, Niger, Western Sahara, Sudan, Senegal
and Tunisia.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/79-GC</identifier>
<title>Invasion of Eastern Timor's capital by Indonesia</title>
<description>Documents mentioning the invasion of Dili by Indonesian
troops</description>
<narrative>Relevant documents deal with the occupation of East Timor by
Indonesia and mention incidents between Indonesian soldiers and the inhabitants
of Dili.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/80-GC</identifier>
<title>Politicians in exile in Germany</title>
<description>Documents mentioning exiled politicians in Germany</description>
<narrative>Relevant documents report about politicians who live in exile in
Germany and mention the nationality and political convictions of these
politicians.</narrative>
```

```
</topic>
<topic lang="en">
<identifier>10.2452/81-GC</identifier>
<title>G7 summits in Mediterranean countries</title>
<description>Documents mentioning G7 summit meetings in Mediterranean
countries</description>
<narrative>Relevant documents must mention summit meetings of the G7 in the
mediterranean countries: Spain, Gibraltar, France, Monaco, Italy, Malta,
Slovenia, Croatia, Bosnia and Herzegovina, Montenegro, Albania, Greece, Cyprus,
Turkey, Syria, Lebanon, Israel, Palestine, Egypt, Libya, Tunisia, Algeria and
Morocco.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/82-GC</identifier>
<title>Agriculture in the Iberian Peninsula</title>
<description>Relevant documents relate to the state of agriculture in the
Iberian Peninsula</description>
<narrative>Relevant docments contain information about the state of agriculture
in the Iberian peninsula. Crops, protests and statistics are relevant. The
countries in the Iberian peninsula are Portugal, Spain and Andorra.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/83-GC</identifier>
<title>Demonstrations against terrorism in Northern Africa</title>
<description>Documents mentioning demonstrations against terrorism in Northern
Africa</description>
<narrative>Relevant documents must mention demonstrations against terrorism in
the North of Africa. The documents must mention the number of demonstrators and
the reasons for the demonstration. North Africa includes the Magreb region
(countries: Algeria, Tunisia, and Morocco, as well as the Western Sahara region)
and Egypt, Sudan, Libya and Mauritania.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/84-GC</identifier>
<title>Bombings in Northern Ireland</title>
<description>Documents mentioning bomb attacks in Northern Ireland</description>
<narrative>Relevant documents should contain information about bomb attacks in
Northern Ireland and should mention people responsible for and consequences of
the attacks.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/85-GC</identifier>
<title>Nuclear tests in the South Pacific</title>
<description>Documents mentioning the execution of nuclear tests in South
Pacific</description>
<narrative>Relevant documents should contain information about nuclear tests
which were carried out in the South Pacific. Intentions as well as plans for
future nuclear tests in this region are not considered as relevant.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/86-GC</identifier>
<title>Most visited sights in the capital of France and its vicinity</title>
```

<description>Documents mentioning the most visited sights in Paris and
surroundings</description>
<narrative>Relevant documents should provide information about the most visited
sights of Paris and close to Paris and either give this information explicitly
or contain data which allows conclusions about which places were most
visited.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/87-GC</identifier>
<title>Unemployment in the OECD countries</title>
<description>Documents mentioning issues related with the unemployment in the
countries of the Organisation for Economic Co-operation and Development (OECD)</description>
<narrative>Relevant documents should contain information about the unemployment
(rate of unemployment, important reasons and consequences) in the industrial
states of the OECD. The following states belong to the OECD: Australia, Belgium,
Denmark, Germany, Finland, France, Greece, Ireland, Iceland, Italy, Japan,
Canada, Luxembourg, Mexico, New Zealand, the Netherlands, Norway, Austria,
Poland, Portugal, Sweden, Switzerland, Slovakia, Spain, South Korea, Czech
Republic, Turkey, Hungary, the United Kingdom and the United States of America
(USA).</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/88-GC</identifier>
<title>Portuguese immigrant communities in the world</title>
<description>Documents mentioning immigrant Portuguese communities in other
countries</description>
<narrative>Relevant documents contain information about Portguese communities
who live as immigrants in other countries.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/89-GC</identifier>
<title>Trade fairs in Lower Saxony</title>
<description>Documents reporting about industrial or cultural fairs in Lower
Saxony</description>
<narrative>Relevant documents should contain information about trade or
industrial fairs which take place in the German federal state of Lower Saxony,
i.e. name, type and place of the fair. The capital of Lower Saxony is Hanover.
Other cities include Braunschweig, Osnabrck, Oldenburg and
Gttingen.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/90-GC</identifier>
<title>Environmental pollution in European waters</title>
<description>Documents mentioning environmental pollution in European rivers,
lakes and oceans.</description>
<narrative>Relevant documents should mention the kind and level of the pollution
and furthermore contain information about the type of the water and locate the
affected area and potential consequences.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/91-GC</identifier>
<title>Forest fires on Spanish islands</title>

<description>Documents mentioning forest fires on Spanish islands</description>
<narrative>Relevant documents should contain information about the location,
causes and consequences of the forest fires. Spanish Islands are: the Balearic
Islands (Majorca, Minorca, Ibiza, Formentera), the Canary Islands (Tenerife,
Gran Canaria, El Hierro, Lanzarote, La Palma, La Gomera, Fuerteventura) and some
islands located just off the Moroccan coast (Islas Chafarinas, Alhucemas,
Alborn, Perejil, Islas Columbretes and Penn de Vlez de la
Gomera).</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/92-GC</identifier>
<title>Islamic fundamentalists in Western Europe</title>
<description>Documents mentioning Islamic fundamentalists living in Western
Europe</description>
<narrative>Relevant Documents contain information about countries of origin and
current whereabouts and political and religious motives of the fundamentalists.
Western Europe consists of Western Europe consists of Belgium, Ireland, Great
Britain, Spain, Italy, Portugal, Andorra, Germany, France, Liechtenstein,
Luxembourg, Monaco, the Netherlands, Austria and Switzerland.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/93-GC</identifier>
<title>Attacks in Japanese subways</title>
<description>Documents mentioning attacks in Japanese subways</description>
<narrative>Relevant documents contain information about attackers, reasons,
number of victims, places and consequences of the attacks in subways in
Japan.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/94-GC</identifier>
<title>Demonstrations in German cities</title>
<description>Documents mentioning demonstrations in German cities</description>
<narrative>Relevant documents contain information about participants, and number
of participants, reasons, type (peaceful or riots) and consequences of
demonstrations in German cities.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/95-GC</identifier>
<title>American troops in the Persian Gulf</title>
<description>Documents mentioning American troops in the Persian
Gulf</description>
<narrative>Relevant documents contain information about functions/tasks of the
American troops and where exactly they are based. Countries with a coastline
with the Persian Gulf are: Iran, Iraq, Oman, United Arab Emirates, Saudi-Arabia,
Qatar, Bahrain and Kuwait.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/96-GC</identifier>
<title>Economic boom in Southeast Asia</title>
<description>Documents mentioning economic boom in countries in Southeast
Asia</description>
<narrative>Relevant documents contain information about (international)

```
companies in this region and the impact of the economic boom on the population.
Countries of Southeast Asia are: Brunei, Indonesia, Malaysia, Cambodia, Laos,
Myanmar (Burma), East Timor, the Phillipines, Singapore, Thailand and
Vietnam.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/97-GC</identifier>
<title>Foreign aid in Sub-Saharan Africa</title>
<description>Documents mentioning foreign aid in Sub-Saharan
Africa</description>
<narrative>Relevant documents contain information about the kind of foreign aid
and describe which countries or organizations help in which regions of
Sub-Saharan Africa. Countries of the Sub-Saharan Africa are: state of Central
Africa (Burundi, Rwanda, Democratic Republic of Congo, Republic of Congo,
Central African Republic), East Africa (Ethiopia, Eritrea, Kenya, Somalia,
Sudan, Tanzania, Uganda, Djibouti), Southern Africa (Angola, Botswana, Lesotho,
Malawi, Mozambique, Namibia, South Africa, Madagascar, Zambia, Zimbabwe,
Swaziland), Western Africa (Benin, Burkina Faso, Chad, Cte d'Ivoire, Gabon,
Gambia, Ghana, Equatorial Guinea, Guinea-Bissau, Cameroon, Liberia, Mali,
Mauritania, Niger, Nigeria, Senegal, Sierra Leone, Togo) and the African isles
(Cape Verde, Comoros, Mauritius, Seychelles, So Tom and Prncipe and
Madagascar).</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/98-GC</identifier>
<title>Tibetan people in the Indian subcontinent</title>
<description>Documents mentioning Tibetan people who live in countries of the
Indian subcontinent.</description>
<narrative>Relevant Documents contain information about Tibetan people living in
exile in countries of the Indian Subcontinent and mention reasons for the exile
or living conditions of the Tibetians. Countries of the Indian subcontinent are:
India, Pakistan, Bangladesh, Bhutan, Nepal and Sri Lanka.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/99-GC</identifier>
<title>Floods in European cities</title>
<description>Documents mentioning resons for and consequences of floods in
European cities</description>
<narrative>Relevant documents contain information about reasons and consequences
(damages, deaths, victims) of the floods and name the European city where the
flood occurred.</narrative>
</topic>
<topic lang="en">
<identifier>10.2452/100-GC</identifier>
<title>Natural disasters in the Western USA</title>
<description>Douments need to describe natural disasters in the Western
USA</description>
<narrative>Relevant documents report on natural disasters like earthquakes or
flooding which took place in Western states of the United States. To the Western
states belong California, Washington and Oregon.</narrative>
</topic>
</topics>
```

Appendix C

Geographic Questions from CLEF-QA

```
<?xml version="1.0" encoding="UTF-8"?>
<input>
<q id="0001">Who is the Prime Minister of Macedonia?</q>
<q id="0002">When did the Sony Center open at the Kemperplatz in
Berlin?</q>
<q id="0003">Which EU conference adopted Agenda 2000 in Berlin?</q>
<q id="0004">In which railway station is the "Museum fr
Gegenwart-Berlin"?</q>
<q id="0005">Where was Supachai Panitchpakdi born?</q>
<q id="0006">Which Russian president attended the G7 meeting in
Naples?</q>
<q id="0007">When was the whale reserve in Antarctica created?</q>
<q id="0008">On which dates did the G7 meet in Naples?</q>
<q id="0009">Which country is Hazor in?</q>
<q id="0010">Which province is Atapuerca in?</q>
<q id="0011">Which city is the Al Aqsa Mosque in?</q>
<q id="0012">What country does North Korea border on?</q>
<q id="0013">Which country is Euskirchen in?</q>
<q id="0014">Which country is the city of Aachen in?</q>
<q id="0015">Where is Bonn?</q>
<q id="0016">Which country is Tokyo in?</q>
<q id="0017">Which country is Pyongyang in?</q>
<q id="0018">Where did the British excavations to build the Channel
Tunnel begin?</q>
<q id="0019">Where was one of Lennon's military shirts sold at an
auction?</q>
<q id="0020">What space agency has premises at Robledo de Chavela?</q>
<q id="0021">Members of which platform were camped out in the "Paseo
de la Castellana" in Madrid?</q>
<q id="0022">Which Spanish organization sent humanitarian aid to
Rwanda?</q>
<q id="0023">Which country was accused of torture by AI's report
```

```
presented to the United Nations Committee against Torture?</q>
<q id="0024">Who called the renewable energies experts to a meeting
in Almera?</q>
<q id="0025">How many specimens of "Minke" whale are left in the
world?</q>
<q id="0026">How far is Atapuerca from Burgos?</q>
<q id="0027">How many Russian soldiers were in Latvia?</q>
<q id="0028">How long does it take to travel between London and
Paris through the Channel Tunnel?</q>
<q id="0029">What country was against the creation of a whale
reserve in Antarctica?</q>
<q id="0030">What country has hunted whales in the Antarctic Ocean?</q>
<q id="0031">What countries does the Channel Tunnel connect?</q>
<q id="0032">Which country organized "Operation Turquoise"?</q>
<q id="0033">In which town on the island of Hokkaido was there
an earthquake in 1993?</q>
<q id="0034">Which submarine collided with a ship in the English
Channel on February 16, 1995?</q>
<q id="0035">On which island did the European Union Council meet
during the summer of 1994?</q>
<q id="0036">In what country did Tutsis and Hutus fight in the
middle of the Nineties?</q>
<q id="0037">Which organization camped out at the Castellana
before the winter of 1994?</q>
<q id="0038">What took place in Naples from July 8 to July 10
1994?</q>
<q id="0039">What city was Ayrton Senna from?</q>
<q id="0040">What country is the Interlagos track in?</q>
<q id="0041">In what country was the European Football Championship
held in 1996?</q>
<q id="0042">How many divorces were filed in Finland from 1990-1993?</q>
<q id="0043">Where does the world's tallest man live?</q>
<q id="0044">How many people live in Estonia?</q>
<q id="0045">Of which country was East Timor a colony before it was
occupied by Indonesia in 1975?</q>
<q id="0046">How high is the Nevado del Huila?</q>
<q id="0047">Which volcano erupted in June 1991?</q>
<q id="0048">Which country is Alexandria in?</q>
<q id="0049">Where is the Siwa oasis located?</q>
<q id="0050">Which hurricane hit the island of Cozumel?</q>
<q id="0051">Who is the Patriarch of Alexandria?</q>
<q id="0052">Who is the Mayor of Lisbon?</q>
<q id="0053">Which country did Iraq invade in 1990?</q>
<q id="0054">What is the name of the woman who first climbed the
Mt. Everest without an oxygen mask?</q>
<q id="0055">Which country was pope John Paul II born in?</q>
<q id="0056">How high is Kanchenjunga?</q>
<q id="0057">Where did the Olympic Winter Games take place in 1994?</q>
<q id="0058">In what American state is Everglades National Park?</q>
<q id="0059">In which city did the runner Ben Johnson test positive
for Stanozol during the Olympic Games?</q>
<q id="0060">In which year was the Football World Cup celebrated in
```

```
the United States?</q>
<q id="0061">On which date did the United States invade Haiti?</q>
<q id="0062">In which city is the Johnson Space Center?</q>
<q id="0063">In which city is the Sea World aquatic park?</q>
<q id="0064">In which city is the opera house La Fenice?</q>
<q id="0065">In which street does the British Prime Minister live?</q>
<q id="0066">Which Andalusian city wanted to host the 2004 Olympic Games?</q>
<q id="0067">In which country is Nagoya airport?</q>
<q id="0068">In which city was the 63rd Oscars ceremony held?</q>
<q id="0069">Where is Interpol's headquarters?</q>
<q id="0070">How many inhabitants are there in Longyearbyen?</q>
<q id="0071">In which city did the inaugural match of the 1994 USA Football
World Cup take place?</q>
<q id="0072">What port did the aircraft carrier Eisenhower leave when it
went to Haiti?</q>
<q id="0073">Which country did Roosevelt lead during the Second World War?</q>
<q id="0074">Name a country that became independent in 1918.</q>
<q id="0075">How many separations were there in Norway in 1992?</q>
<q id="0076">When was the referendum on divorce in Ireland?</q>
<q id="0077">Who was the favourite personage at the Wax Museum in
London in 1995?</q>
</input>
```

Appendix D

Impact on Current Research

Here we discuss some works that have been published by other researchers on the basis of or in relation with the work presented in this Ph.D. thesis.

The Conceptual-Density toponym disambiguation method described in Section 4.2 has served as a starting point for the works of Roberts et al. (2010) and Bensalem and Kholladi (2010). In the first work, an "ontology transition probability" is calculated in order to find the most likely paths through the ontology to disambiguate toponym candidates. They combined the ontological information with event detection to disambiguate toponyms in a collection tagged with SpatialML (see Section 3.4.4). They obtained a recall of 94.83% using the whole document for context, confirming our results on context sizes. Bensalem and Kholladi (2010) introduced a "geographical density" measure based on the overlap of hierarchical paths and frequency, similarly to our CD methods. They compared on GeoSemCor, obtaining a F-measure of 0.878. GeoSemCor was used also in Overell (2009) for the evaluation of his SVM-based disambiguator, which obtained an accuracy of 0.671.

Michael D. Lieberman (2010) showed the importance of local contexts as highlighted in Buscaldi and Magnini (2010), building a corpus (LGL corpus) containing documents extracted from both local and general newspapers and attempting to resolve toponym ambiguities on it. They obtained 0.730 in F-measure using local lexicons and 0.548 disregarding the local information, indicating that local lexicons serve as a high precision source of evidence for geotagging, especially when the source of documents is heterogeneous, such as in the case of the web.

Geo-WordNet was recently joined by another almost homonymous project, *GeoWordNet* (without the $-$) by Giunchiglia et al. (2010). In their work, they expanded WordNet with synsets automatically extracted from Geonames, actually converting Geonames

into a hierarchical resource which inherits the underlying structure from WordNet. At the time of writing, this resource was not yet available.

Declaration

I herewith declare that this work has been produced without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This Ph.D. thesis has not previously been presented in identical or similar form to any other examination board.

The thesis work was conducted under the supervision of Dr. Paolo Rosso at the Universidad Politécnica of Valencia.

The project of this Ph.D. thesis was accepted at the Doctoral Consortium in SIGIR 2009[1] and received a travel grant co-funded by the ACM and Microsoft Research.

The Ph.D. thesis work has been carried out according to the European Ph.D. mention requirements, which include a three months stage in a foreign institution. The three months stage was completed at the Human Language Technologies group of FBK-IRST in Trento (Italy) from May 11^{th} to August 11^{th} 2009 under the supervision of Dr. Bernardo Magnini.

Formal Acknowledgments

The following projects provided funding for the completion of this work:

- TEXT-MESS 2.0 (sub-project TEXT-ENTERPRISE 2.0: Text comprehension techniques applied to the needs of the Enterprise 2.0), CICYT TIN2009-13391-C04-03
- Red Temática TIMM: Tratamiento de Información Multilingüe y Multimodal, CICYT TIN 2005-25825-E

[1] Buscaldi, D. 2009. *Toponym ambiguity in Geographical Information Retrieval.* In Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in information Retrieval (Boston, MA, USA, July 19 - 23, 2009). SIGIR '09. ACM, New York, NY, 847-847.

- TEXT-MESS: Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano (subproject UPV: MiDEs), CICYT TIN2006-15265-C06
- Answer Extraction for Definition Questions in Arabic, AECID-PCI B/017961/08
- Sistema de Búsqueda de Respuestas Inteligente basado en Agentes (AraEsp), AECI-PCI A01031707
- Systéme de Rècuperation de Rèponses AraEsp, AECI-PCI A/7067/06
- ICT for EU-India Cross-Cultural Dissemination: EU-India Economic Cross Cultural Programme ALA/95/23/2003/077-054
- R2D2: Recuperación de Respuestas en Documentos Digitalizados, CICYT TIC2003-07158-C04-03
- CIAO SENSO: Combining Corpus-Based and Knowledge-Based Methods for Word Sense Disambiguation, MCYT HI 2002-0140

I would like to thank the mentors of the 2009 SIGIR Doctoral Consortium for their valuable comments and suggestions.

October 2010, Valencia, Spain