Geographical Information Resolution and its Application to the Question Answering Systems

Daniel Ferrés Domènech
dferres@lsi.upc.edu

Director
Horacio Rodríguez Hontoria

Memòria del DEA i Projecte de Tesi
Programa de Doctorat en Intel·ligència Artificial
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

12 de Gener de 2007

Contents

Chapter 1

Introduction

Information Retrieval and Question Answering techniques could be defined as algorithms that help the user to satisfy an information need. Question Answering (QA) is the task of, given a question expressed in Natural Language (NL), retrieving its correct answer (a single item, a text snippet,...) from closed collections or the Web. This task could be considered a step beyond Information Retrieval (IR) which consists in searching information in documents, documents themselves, or metadata which describe documents. IR systems retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates & Ribeiro-Neto, 1999). Usually the QA task consists in solve factoid questions. Factoid questions are questions that seek short fact-based answers like entities, organizations, persons, dates,... (e.g. *What is the capital of France?, Who is the President of the United States?, Which is the color of the sky?*).

Question Answering and Information Retrieval need a set of Natural Language Processing (NLP) algorithms to perform a comprehension of a user textual request and the textual documents involved in the search. NLP techniques process electronic texts and analyze them in order to provide lexical, syntactic, semantic, and/or discourse information about the text.

Major Internet search engines (e.g. Google, Yahoo, Lycos, etc.) are mainly dealing with Open Domain Information Retrieval and Question Answering. Due to the difficulty to treat specific domain queries with an open domain technology, in the last years has emerged a growing community of NLP researchers that explore the focus on Restricted-Domains such as genomics (in several workshops and international evaluations (Hersh et al., 2006)), geography (in Geographical Information Retrieval (GIR) workshops and evaluations (Gey et al., 2005)), laws (e.g. Legal track in TREC 2006 (Baron et al., 2006)), among others are currently being explored.

Building Restricted-Domain NLP applications implies the need of more precision and the use specific knowledge of the domain (e.g. lexicons, dictionaries, corpora, axioms, etc.). Data-driven methods based on exploiting the redundancy of answers in several documents are not useful in these systems.

This thesis project studies general framework architectures for Restricted-Domain Factoid QA and Restricted-Domain IR applications. Concretely we are testing our techniques in the Geographical Domain. Geography is a widely used domain in web queries. A study by (Sanderson & Kohler, 2004), over a random sample of 2,500 queries of the 2001 Excite query log showed that a 18,6% of the queries contained a geographic term and 14.8% contained a place name.

Automatic understanding of the geographical concepts appearing in an electronic text could be defined as Geographical Information Resolution (GIRE). GIRE implies that every geographical concept in the text must be recognized, classified in a fine geographical ontology and disambiguated into its geographical world referent.

1.1 Structure of the Document

The rest of this thesis project is structured as follows. The next three chapters describe the state-of-the-art about Question Answering (Chapter 2), Geographical Information Retrieval (Chapter 3), and Geographical Information Resolution (Chapter 4). Then the following chapters explain our work and the approaches that we took to face these NLP tasks: Open-Domain Question Answering for factoid questions (Chapter 5), Geographical Question Answering (Chapter 6), Geographical Information Retrieval (Chapter 7), Geographical Named Entity Subclassification (Chapter 8). Our work plan proposal to complete the thesis project is presented in Chapter 9. To conclude, Chapter 10 lists the publications that our work has produced.

Chapter 2

Question Answering - State-of-the-art

Question Answering (QA) is the task of, given a question expressed in Natural Language (NL), retrieving its correct answer (a single item, a text snippet,...) from closed collections or the Web. This task could be considered a step beyond Information Retrieval (IR) which consists in searching information in documents, documents themselves, or metadata which describe documents. IR systems retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates & Ribeiro-Neto, 1999). The user query normally consists on a set of relevant keywords and/or regular expressions. In QA the user request is a single piece of information instead of an entire document, and the input is a question expressed in Natural Language instead of a query. Document collections, digital libraries, web search engines and other sources of electronics documents are often used in both IR and QA.

2.1 History

The history of QA begins in the 1960s (Simmons, 1965). In this period appeared the early QA systems named in these lines Natural Language Interfaces (NLI) and aiming to look for the answer in databases instead of free text. Most of those systems were restricted-domain and database-oriented such as BASEBALL (Green et al., 1963) and LUNAR (Woods, 1977). BASEBALL answered questions about the US baseball league over a period of one year. LUNAR answered questions about the geological analysis of rocks returned by the Apollo moon missions using Augmented Transition Networks (ATNs) and procedural semantics[1].

Some of the early AI systems were restricted-domain QA systems with a core database handwritten by experts in the domain like SHRLDU (Winograd, 1972), ELIZA (Weizenbaum, 1966), and GUS(Bobrow et al., 1977). SHRDLU simulated the operation of a robot in a toy world (the ”blocks world”), and it offered the possibility to ask the robot questions about the state of the world. It was the first real demo of combination of syntax, semantics, and reasoning. ELIZA simulated a conversation with a psychologist. ELIZA was able to converse on any topic by resorting to very

[1]The assumption is that input sentences correspond to programs that perform some desired action on the database, and that words in the input correspond to program steps.

simple rules that detected important words in the person's input. GUS was a dialog system for airline reservation.

The 1970s and 1980s appeared some ambitious projects in text comprehension and QA such as: LIFER, QUALM, CHAT-80, LILOG and the Unix Consultant (UC). LIFER (Hendrix, 1977) allowed asking questions about U.S. Navy ships. This system used a semantic grammar with domain information built within. QUALM (Lehnert, 1978) was an application of scripts and plans for story comprehension in a very restrictive domains (e.g., restaurant scripts). CHAT 80 (Warren & Pereira, 1982) was a NLP query system about world geography developed in Prolog and using Definite Clause Grammars (DCGs). Another interesting project was LILOG (Herzog & Rollinger, 1991), a text-understanding system that operated on the domain of tourism information in a German city. Finally, the Unix Consultant (UC) (Wilensky et al., 2000) was a system that answered questions from the domain of Unix. The system had a comprehensive hand-crafted knowledge base of its domain.

In the late 1990s the annual Text Retrieval Conference (TREC) started to include a Question Answering track which has been running until the present (Voorhees, 1999). Systems participating in this competition were expected to answer questions in English on any topic by searching a corpus of text. This competition fostered research and development in open-domain text-based QA. The best system of the 2004 competition achieved 77% correct fact-based questions (Voorhees, 2004). On the other hand, annual QA evaluations for European and Asian languages were established in the CLEF and NTCIR workshops in the early 2000s.

Currently, an increasing number of QA systems include the World Wide Web as one more corpus of text. Some of the most well-know systems that use the Internet as a corpus are START, MURAX, AskJeeves, AnswerBus, PowerAnswer and Aranea. START[2], was the first Web-based QA system (Katz et al., 2002) for English, it has been on-line and continuously operating since December, 1993. MURAX (Kupiec, 1993) was also an earlier QA system that combined NLP with the use of an online encyclopedia with hand-coded annotations to sources. Ask Jeeves [3] , which licensed its technology from START, became the first large scale question-answering system, until it moved to a more traditional search engine business. AnswerBus [4] is an Open-Domain Question Answering (ODQA) system (Zheng, 2002). It accepts questions in several languages (including English, German, French, Italian, Spanish and Portuguese) and extracts answers from the Web. PowerAnswer [5] is, currently, one of the best factoid and definitional QA systems. This system has been developed by the LanguageComputer Corporation [6] (LCC). It achieved performances of 77% and 71% of correct fact-based questions in TREC 2004 (Moldovan et al., 2004) and TREC 2005 (Harabagiu et al., 2005) respectively. LCC has developed advanced algorithms including question semantics. Aranea[7] is the first fully downloadable open-source Web-based factoid question answering system (Lin & Katz, 2003). It is an improved version of the Microsoft's askMSR (Michele Banko and Eric Brill and Susan Dumais and Jimmy Lin, 2002) system that uses a combination of data redundancy

[2] **START**. `http://start.csail.mit.edu/`
[3] **Actually AskJeeves is Ask**. `http://www.ask.com`
[4] **AnswerBus**. `http://www.answerbus.com`
[5] **PowerAnswer demo**. `http://www.languagecomputer.com/demos/question_answering`
[6] **LCC**. `http://www.languagecomputer.com/`
[7] **Aranea**. `http://www.umiacs.umd.edu/~jimmylin/downloads`

and database techniques.

Current QA systems, use a combination of Natural Language Processing, Information Extraction, and Information Retrieval Techniques. In the last years NLP tools such as part-of-speech taggers, Named Entity taggers, syntactic parsers and semantic taggers (e.g. using WordNet) have become widely used in several approaches for QA.

2.2 Classification of QA Systems

Question Answering systems can be classified from different points of view. This section presents different QA categorization types.

2.2.1 Classification by Questioner Type

Discussed in (Carbonell et al., 2000) and (Burger et al., 2000), a set of 4 types of questioners could determine the type of QA system by means of the questioner type (see examples in Table 2.1):

- **Level 1. Casual Questioner**. The Casual Question is the type of QA questioner who asks simple factual questions, which could be answered in a single short phrase. For Example: Where is New York City? What is the currency unit of India? Who was the first man in the moon? etc.

- **Level 2. Template Questioner**. This type of user requires a QA system which could be able to create ”standard templates” with certain types of information to be found and filled in. It it possible that the answer will require retrieve multiple documents and combine portions of answers into a single response. The questions are basically factual but is required more information than a single phrase. Some questions can require information about some real entity, e.g. “What do we know about X?. And others are list questions, e.g. ”What are all of the countries that border Brazil?” or ”Who are all of the Major League Baseball Players who have had 3000 or more hits during their major league careers? The QA System might also need to resolve finding in different documents overlapping lists of products that may include variations in the ways in which the products are identified.

- **Level 3. Questioner as a Reporter.** The QA questioner would be someone who is still focused factually, but now needs to pull together information from a variety of sources. At this level, the QA system will need to move beyond text sources and involve multiple media and multiple foreign languages. As an example a reporter covering an event who needs to do a series of highly related questions to a general information system to pull together all the information.

- **Level 4. Professional Information Analyst.** This profile requires analytic tools capable of providing answers to complex, multi-faceted questions involving judgement terms that analysts might wish to pose to multiple, very large, very heterogeneous data sources that may physically reside in multiple agencies and may include: Structured and unstructured language

data of all media types, multiple languages, multiple styles, formats, etc., ” Image data to include document images, still photographic images, and video; and Abstract/technical data. The Professional Information Analyst profile might include: reporters for national newspapers, police agents, historians, stock market analysts, scientists and researchers, intelligence analysts, among others.

Table 2.1: Questions and their focus at different questioner type levels (Burger et al., 2000)

Level	Question	Focus
Level 1: Casual Questioner	Q: Why did Elian Gonzales leave the U.S.?	Focus: the departure of Elian Gonzales.
Level 2: Template Questioner	Q: What was the position of the U.S. Government regarding the immigration of Elian Gonzales in the U.S.?	Focus: set of templates that are generated to extract information about (1) INS statements and actions regarding the immigration of Elian Gonzales; (2) the actions and statements of the Attorney General with respect to the immigration of Elian Gonzales; (3) actions and statements of other members of the administration regarding the immigration of Elian Gonzales;
Level 3: Cub reporter	Q: How did Elian Gonzales come to be considered for immigration in the U.S.? –translated into a set of simpler questions: Q1: How did Elian Gonzales enter the U.S.? Q2: What is the nationality of Elian Gonzales? Q3: How old is Elian Gonzales? Q4: What are the provisions in the Immigration Law for Cuban refugees? Q5: Does Elian Gonzales have any immediate relatives?	Focus: composed of the question foci of all the simpler questions in which the original question is translated. Focus Q1: the arrival of Elian Gonzales in the U.S. Focus Q2: the nationality of Elian Gonzales. Focus Q3: the age of Elian Gonzales. Focus Q4: immigration law. Focus Q5: immediate relatives of Elian Gonzales.
Level 4: Information Analyst	Q: What was the reaction of the Cuban community in the U.S. to the decision regarding Elian Gonzales?	Focus: every action and statement, present or future, taken by any American-Cuban, and especially by Cuban anti-Castro leaders, related to the presence and departure of Elian Gonzales from the U.S. Any action, statements or plans involving Elian’s Miami relatives or their lawyers.

2.2.2 Classification by Knowledge Used

(Moldovan et al., 1999) provided a taxonomy of QA based on the necessary knowledge to resolve the questions. They considered important the three following criteria: Knowledge Bases (KB), Reasoning, and Natural Language Processing (NLP) indexing techniques. Knowledge bases and reasoning provide the medium for building question contexts and matching them against text documents. Indexing identifies the text passages where answers may lie, and natural language processing provides a framework for answer extraction (Moldovan et al., 1999). See more details of these levels in Table 2.2.

Table 2.2: QA taxonomy based on Knowledge bases, reasoning and NLP techniques (Moldovan et al., 1999)

Class	KB	Reasoning	NLP/Indexing	Examples/Comments
1	dictionaries	simple heuristics, pattern matching	complex noun, apposition, simple semantics, keyword indexing	Q33: What is the largest city in Germany? A: .. Berlin, the largest city in Germany.. Answer is: simple datum or list of items found verbatim in a sentence keyword or paragraph.
2	ontologies	low level	verb nominalization, semantics, coherence, discourse	Q198: How did Socrates die? A: .. Socrates poisoned himself.. Answer is contained in multiple sentences, scattered throughout discourse a document.
3	very large KB	medium level	advanced nlp, semantic indexing	Q: What are the arguments for and against prayer in school? Answer across several texts.
4	Domain KA and Classification, HPKB	high level		Q: Should Fed raise interest rates at their next meeting? Answer across large number of documents,domain specific knowledge acquired automatically.
5	World knowledge	very high level, special purpose		Q: What should be the US foreign policy in the Balkans now? Answer is a solution to a complex, possible developing scenario.

2.2.3 Classification by Question Types

- **Factoid Questions**. These are questions that seek short fact-based answers like entities, organizations, persons, dates,...(e.g. *What is the capital of France?*, *Who is the President of the United States?*, *Which is the color of the sky?*). Usually the answer is a noun (e.g. blue), a noun phrase (slightly blue) or a Named Entity (e.g. 1979, Paris, George Bush). But some times an adjective or an adverb could be the answer (i.e. in most of the "How" questions). Since 1999, open-domain factoid questions have been largely evaluated in several QA international contests such as TREC, CLEF, and NTCIR.

- **Definitional Questions**. Definitional questions require definitions of certain concepts or entities. These questions are usually posed with this manner; "What is ...?" or "Who is ...?" (e.g. *What is AIDS?*, *Who is Aaron Copland?*, *What is an atom?*). QA evaluations such TREC and CLEF have been evaluating this kind of questions.

- **List Questions**. List questions are requests for a set of instances of a specified type. (e.g. Which soccer players won the UEFA Champions League in 1996?, What are the brand names of the most known French wines?, Name 20 countries other than the United States that have a McDonalds restaurant. etc). In the TREC 2001 QA track (Voorhees, 2001) started the use of list questions in the TREC evaluations. In those evaluation a task exclusively for list questions was established and the questions specified a target number of instances to retrieve. For example, one question in the task was "What are 9 novels written by John Updike?"

- **Context Questions**. The context QA consists in the capability of a QA system to answer series of questions posed in the context of previous questions or answers. It implies the user interaction with the system and the system's understanding of the context. In the TREC-10 (2001) (Voorhees, 2001) started a task for context QA. The task consisted on a set of question series in that the interpretation of the questions could depend on the meaning or answer of an earlier question in the series.

- **Interactive Questions**. Interactive Question Answering (IQA) consists in the active participation of the user in the QA process with the aim of getting better results than without its participation. Human-Machine interaction is done by dialog systems that allow to do questions in the context of previous interactions. The dialog with the systems allows the user to: i) clarify the information that has been detected ambiguous or fuzzy by the system, ii) correct the errors that appeared in different phases of the system's architecture, iii) help the QA system giving information that couldn't be handled by the system, iv) negotiate with the system the desired features in the answer retrieving.

  The initial IQA systems began in restricted domains and database querying, like SHRLDU (Winograd, 1972) and GUS (Bobrow et al., 1977). A recent relevant project is the HITIQA (Strzalkowski et al., 2005). This system helps the analyst questions in context and answers them according to the global information of the task. In the last years some international evaluations of IQA systems have been taken place in TREC-9 (2000) and in CLEF 2004 and CLEF 2005, these last ones focused in cross-lingual IQA.

2.2.4 Classification by Domain

- **Open-Domain Question Answering (ODQA)**. These systems deal with general questions about many themes. Normally they use huge corpus and/or the World Wide Web to extract the answer. In several QA contests huge news corpora such as AQUAINT or TIDES including news from LA Times, APW, NYT among others are used for evaluation purposes. ODQA systems tend to use NLP techniques in combination with IR engines. Some of the top-performance systems in different contests such as TREC or CLEF use a wide range of NLP tools such as POS taggers, NE taggers, multi-word detectors, syntactic parsers, semantic sense annotators, etc. Sometimes large hierarchical conceptual ontologies as WordNet and World Knowledge Bases are used in combination of reasoning mechanisms based on first logics.

- **Restricted-Domain Question Answering (RDQA)**. These systems deal with questions about a specific domain (e.g. geography, medicine, etc.). They often use domain-specific knowledge bases and corpus. Usually, for RDQA, the answers are searched in relatively small domain specific collections, so methods based on exploiting the redundancy of answers in several documents are not useful. Furthermore, a highly accurate Passage Retrieval module is required because frequently the answer occurs in a very small set of passages. RDQAs are frequently task-based. So, the repertory of question patterns is limited allowing a good accuracy in Question Processing with limited effort. User requirements regarding the quality of the answer tend to be higher in RDQA. As (Chung et al., 2004) pointed out, "no answer"

is preferred to a wrong answer. In RDQA not only NEs but also domain specific terminology plays a central role. This fact usually implies that domain specific lexicons and gazetteers have to be used. In some cases, as in Geographical Domain, many documents included in the collections are far to be standard NL texts but contain tables, lists, ill-formed sentences, etc. sometimes following a more or less defined structure. Thus, extraction systems based, as our, on the linguistic structure of the sentences have to be relaxed in some way to deal with this kind of texts. More information about RDQA systems can be found in the ACL 2004 Workshop on QA in Restricted Domains[8] and the AAAI 2005 Worshop on Question Answering in Restricted Domains (Molla & Vicedo, 2005) .

2.2.5 Classification by Information Access

Two basic types of Question Answering systems can be distinguished depending of the structure of the knowledge that they use to answer.

- **Database-oriented**: systems that access to structured information contained in a database in order to answer the questions. The main challenge of these systems is to transform a natural language question into a database query (Monz, 2003). Normally, database QA systems are focused in narrow domains. BASEBALL (Green et al., 1963) and LUNAR (Woods, 1977) are two early systems. The fact that this systems are focused in good results, but expand to other domains is a hard task, expertise is required.

- **Text-based**: Most systems use unstructured information such as plain texts: newspapers, manuals, encyclopedias, etc. to find the answer. Textual question answering systems match the question with text units, e.g., phrases or sentences, in the document collection, and within those units, identify the element the question is asking for. The task of identifying elements of the appropriate type is closely related to the research area of Information Extraction and Named Entity Recognition and Classification. Moreover, for text-based QA system data redundancy plays and important role for answer extraction (i.e. more data implies higher chance that appear occurrences in text where this information is expressed in a way similar to the question). On the other hand, huge amounts of data increases the computational costs of finding an answer.

2.3 Architecture of QA Systems

The common architecture of most of the existing QA systems is generally divided into 3 phases: Question Classification, Passage Retrieval (sometimes divided into Document Retrieval and true Passage Retrieval) and Answer Extraction (sometimes divided into Candidates Extraction and Answer Selection). In most systems, these phases are executed sequentially, but some systems such as PowerAnswer (Harabagiu et al., 2005) perform several iterations in order to get the correct answer. In every system NLP and IR techniques are applied, some times with manually built rules or databases or learned approaches using ML techniques.

[8] `http://acl.ldc.upenn.edu/acl2004/qarestricteddomain/`

2.3.1 Natural Language Analysis

Natural Language Analysis (NLA) consists in obtaining detailed information about the words, relations between words and sentences. NLP tools are widely used in QA tasks. A common classification of linguistic processors could have 4 layers: lexical analysis, syntactic analysis, semantic analysis and contextual analysis. The lexical layer includes basic processors, such as: tokenizers (including specific rules for dealing with dates, quantities, time expressions, formulas, jargon, etc.) sentence segmenters, morphological analyzers, morphological disambiguators (Part-of-Speech taggers), Named Entity Recognizers and Classifiers, Multi-word detectors, Semantic taggers (usually WordNet with synsets)... In a second layer, syntactic processors such as: Syntactic Parsers, Shallow Parsers.... The third layer, deals with semantic analysis (i.e. represents the meaning of a sentence) and includes all kinds of semantic representations (logic forms, dependency trees, etc.). Finally, a top-layer, the discourse analysis includes contextual processors such as: Coreference Resolution, Word Sense Disambiguation, lexical chainers...

Lexical Analysis

POS tagging algorithms, NERC algorithms and Lexical databases are the most important tools for the lexical analysis. These tools are described above.

Part-of-Speech Tagging. Part-of-Speech tagging task consists in attaching the lexical category of each lexical unit of a sentence. Lexical categories can be open word classes or closed ones depending if they can acquire constantly new members. Open word classes are nouns, verbs, adjectives, adverbs and interjections. POS tags can include morphological information such as: gender, number and person, verbal modes. The most common POS tag-set for English is Penn Tree-Bank (PTB) tag-set [9]. The EAGLES [10] group tag-set is used for other languages such as Spanish [11] and Catalan [12].

POS tagging is strongly related with corpus linguistics, because ML approaches to POS tagging use huge corpora for training. English corpora often used for training POS taggers are: a) Brown Corpus (1,000,000 words of running English prose text), b) British National Corpus (BNC) with a 100-million-words English, c) Penn Tree-bank Corpus (Wall Street Journal (WSJ)) (1,200,000 words).

ML approaches to POS tagging often are statistical, with some exceptions as rule-based systems (EngCG) (Voutilainen, 1997), or hybrid systems (as Brill's tagger (Brill, 1992) that used an error-correction transformation approach, Padro's Relax (Padró, 1996) that combines statistical and symbolic rules and a optimization by relaxation framework.

[9] **Penn Tree-Bank (PTB) annotation guidelines.** `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`

[10] **EAGLES group**. `http://www.ilc.cnr.it/EAGLES96/home.html`

[11] **EAGLES Spanish tag-set**. `http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html`

[12] **EAGLES Catalan tag-set**. `http://garraf.epsevg.upc.es/freeling/doc/userman/parole-ca.html`

State-of-the art POS tagging techniques achieve high effectivity. The HMM statistical-based tagger Trigrams 'n' Tags (TnT) (Brants, 2000) performs 96.7% of accuracy in English when trained with the WSJ corpus. (Collins, 2002) used algorithms based on the perceptron algorithm and Viterbi decoding, performing an accuracy of 97.11% On the other hand, (Toutanova et al., 2003) reported an accuracy of 97.24% over the Penn Tree-bank WSJ corpus using Bidirectional Dependency Networks. Finally, a Support Vector Machines approach, SVMTagger (Giménez & Màrquez, 2004) outperformed TnT with a 97.2% of accuracy in the WSJ corpus. SVMTagger achieves also good results for Spanish: 96.89% of accuracy.

Named Entity Recognition and Classification. Named Entity Recognition and Classification (NERC) is the task of recognizing and properly classifying named Noun Phrases in a set of predefined categories. NERC is a central issue in many basic NLP tasks such as co-reference resolution, document linking or topic detection, and also has currently become present in most of the Text Mining applications. NERC can be seen as a two-step process: Named Entity Recognition (NER) and Named Entity Classification (NEC). Named Entity Recognition consists on locating a sequence of one or more contiguous words that can be considered candidate to be a NE and deciding if it is an actual one. Named Entity Classification implies assigning a class from a closed dataset to the NE. Most Named Entity Classification (NEC) systems reduce this set to the basic 7 MUC classes: LOCATION, PERSON, etc. (see Table 2.3), while finer grained classification has been faced recently in extended NEC (Sekine et al., 2002).

Different NERC systems have been evaluated in several NERC tasks in different international Information Extraction conferences and workshops. In 1996 the Multilingual Entity Task (Merchant & Okurowski, 1996) (Sundheim, 1995a) in the Message Understanding Conference (MUC-6) (Sundheim, 1995b), was the first evaluation on NERC. MUC campaigns were run by the American National Institute of Standards and Technology (NIST) with the aim of stimulating research in NERC providing a set of entity classes to extract, training and test data, and a evaluation plan. In the MUC evaluations the following expressions to detect were considered: Named Entities (persons, locations and organizations), temporal expressions (time and date) and numeric expressions (percentage and money) (see Table 2.3). An example is given in Figure 2.1.

Element	Entity Class	Expected Names
ENAMEX	ORGANIZATION	named corporate, governmental, or other organizational entity
	PERSON	named person or family
	LOCATION	name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)
TIMEX	DATE	complete or partial date expression
	TIME	complete or partial expression of time of day
NUMEX	MONEY	monetary expression
	PERCENT	percentage

Table 2.3: Tag elements and entity classes at the NE task of MUC conferences (Chinchor & Robinson, 1997).

<ENAMEX TYPE="ORGANIZATION">Grupo Televisa</ENAMEX> and <ENAMEX TYPE="ORGANIZATION">Globo</ENAMEX> plan to offer national and local programming in Spanish and Portuguese. Initially, the venture's partners said they planned to invest <NUMEX TYPE="MONEY">$500 million</NUMEX>.
But a similar explosion <TIMEX TYPE="DATE">last year</TIMEX> delayed the plans of several American media companies to offer a package of satellite television services in <ENAMEX TYPE="LOCATION">Asia</ENAMEX>.

Figure 2.1: Example of NERC from MUC-7 Conference.

Different Information Extraction contests organized Named Entity Extraction tasks:

In 1996, MET (Multilingual Entity Task) for instance organized a task similar to the MUC one for Spanish, Chinese and Japanese (Merchant & Okurowski, 1996). MET task consisted allowed 10 entity types: Person, Organization, Location, Date, Time, Duration, Percent, Money, Measure, and Number.

The conference on *Computational Natural Language Learning* (CoNLL) [13] organized a shared Task[14] [15] in 2002 and 2003 that consisted on NERC for four languages: English, German, Dutch and Spanish (Tjong Kim Sang, 2002) (Tjong Kim Sang & De Meulder, 2003a). This task consisted in an evaluation of different state-of-the-art algorithms for NERC. The classes used were: person, organization, location and others. From 1998 to 1999, the IREX (Information Retrieval and Entity Extraction) project [16] organized a Named Entity task (Sekine & Eriguchi, 2000). The IREX evaluation used the basic 7 MUC classes plus the class "artifact" (e.g. "Odissey" as a book title or "Windows" as a software product name). In 1999, the DARPA-TIDES project carried a Named Entity Recognition task for various textual information sources. This task was called the Information Extraction-Entity Recognition Evaluation (IE-ER). The task 1999 IE-ER evaluation was designed to evaluate Named Entity Recognition in Newswire texts and Automatically recognized transcripts of news for English and Chinese. Some of the major updates include: ENAMEX, TIMEX: A DURATION tag has been added, NUMEX: with addition of MEASURE (standard numeric measurement phrases such as age, area, distance, energy, speed, temperature, volume, and weight plus syntactically-defined measurement phrases) and CARDINAL (a numerical count or quantity of some object in the form of whole numbers, decimals, or fractions)

The NE task in MUC was inherited by the ACE project [17] in the U.S.A., where 2 new categories are added, GPE (Geographical and Political Entities, such as France or New York) and facility, such as Empire State Building . In the ACE project were used 5 coarse classes (ENAMEX, TIMEX, NUMEX, MEASURE, CARDINAL) which could be expanded to 11 classes (Person, Organization, Location, GPE, Facility, Date, Time, Duration, Percent, Money, Measure, and Number).

[13] **CoNLL**. CoNLL is the yearly conference of SIGNLL, the Special Interest Group of the Association for Computational Linguistics on Machine Learning of Language; `http://www.aclweb.org/signll`

[14] **CoNLL Shared Task 2002.** `http://www.cnts.ua.ac.be/conll2002/ner/`

[15] **CoNLL Shared Task 2003.** `http://www.cnts.ua.ac.be/conll2003/ner/`

[16] **IREX project.** `http://nlp.cs.nyu.edu/irex/index-e.html`

[17] **ACE project.** `http://www.itl.nist.gov/iad/894.01/tests/ace/`

Most state-of-the-art NEC systems use coarse-grained MUC-style datasets for performing the classification task reducing it to distinguish among LOCATION, PERSON, ORGANIZATION and so. There is, however, currently, a growing interest on going beyond using finer-grained classification sets. (Sekine et al., 2002), for instance, proposes an extended NE hierarchy of 150 types, while (Manov et al., 2003) use 97 classes for the location sub-ontology. Current approaches to NERC are manual rules, supervised or unsupervised ML, and hybrid approaches (see Tables 2.4 and Table 2.5 for a detailed view of top performance systems)

Algorithm	System	Test Corpus	F-measure
AdaBoost	ABIONET (Carreras et al., 2003a)	CONLL-2003	85.00±0.8
Decision Trees	NYU (Sekine, 1998)	MET-2 (japanese)	88.62%
Conditional Random Fields	(McCallum & Li, 2003)	CONLL-2003	84.04±0.9
Hidden Markov Models	IdentiFinder (Bikel et al., 1999)	MUC-6 (en)	94.9%
	Nymble (Bikel et al., 1997)	MUC-6	93%
	SIFT (Miller et al., 1998)	MUC-7	90.4%
	(Whitelaw & Patrick, 2003)	CONLL-2003	79.78±1.0
	(Zhou & Su, 2001)	MUC-6	96.60%
	(Zhou & Su, 2001)	MUC-7	94.10%
Maximum Entropy Models	MENERGI (Chieu & Ng, 2002)	MUC-7	87.24%
	MENE (Borthwick et al., 1998)	MUC-7	84.22%
	(Chieu & Ng, 2003)	CONLL-2003	88.31±0.7
	(Bender et al., 2003)	CONLL-2003	83.92±1.0
	(Curran & Clark, 2003)	CONLL-2003	84.89±0.9
Memory-Based Learning	(Hendrickx & van den Bosch, 2003)	CONLL-2003	78.20±1.0
Robust Risk Minimization	(Zhang & Lee, 2003)	CONLL-2003	85.50±0.9
Support Vector Machines	GATE-SVM (Li et al., 2005)	CONLL-2002	86.30%
System Combination	(Surdeanu et al., 2005)	CONLL-2003	90.17%
	(Florian et al., 2003)	CONLL-2003	88.76±0.7
	(Klein et al., 2003)	CONLL-2003	86.07±0.8
	(Mayfield et al., 2003)	CONLL-2003	84.67±1.0
Voted Perceptrons	(Carreras et al., 2003b)	CONLL-2003	84.30±0.9

Table 2.4: Machine Learning Approaches to Named Entity Recognition and Classification for English.

The main features used for NERC in the CoNLL 2002 and CoNLL 2003 shared tasks are lexical features, part-of-speech tags, previously predicted NE tags, affix information (n-grams), orthographic information, gazetteers, chunk tags, orthographic patterns and global case information (Tjong Kim Sang, 2002) (Tjong Kim Sang & De Meulder, 2003a). In the CoNLL-2003 evaluation (Tjong Kim Sang & De Meulder, 2003a) few approaches used trigger words, bag of words, global document information, and quoting.

Named Entity Evaluation tasks use the evaluation measures of Precision, Recall and F1. Precision is the percentage of NEs found that are predicted correctly:

$$\text{precision} = \frac{\#NEs\_predicted\_and\_correct}{\#NEs} \tag{2.1}$$

Model	Algorithm	Test Corpus	F-measure
Hand-Crafted Rules	SRA (Krupka, 1995)	MUC-6	96.42%
	FACILE (Black et al., 1998)	MUC-7	82.25%
	ALEMBIC (Aberdeen et al., 1995)	MUC-6	91.20%
	LaSIE (Gaizauskas et al., 1995)	MUC-6	94.41%
	LaSIE-II (Humphreys et al., 1998)	MUC-7	90.41%
	NetOWL (G. R. Krupka, 1998)	MUC-7	91.60%
	Fastus (Appelt et al., 1995)	MUC-6	94.00%
	PLUM (Weischedel, 1995)	MUC-6	93.65%
Hybrid Systems	LTG (Mikheev et al., 1998)	MUC-7	93.39%
	(Yu et al., 1998)	MUC-7	77.74 %
	MENE+Proteus (Borthwick et al., 1998)	MUC-7	88.80 %

Table 2.5: Hybrid and Hand-Crafted Rules systems for NERC (English).

The recall measures the proportion of NE present in the corpus that are found by the system:

$$\text{recall} = \frac{\#NEs\_found}{\#NEs} \tag{2.2}$$

The F measure controls the relative importance of recall and precision. The general formula of the F measure is:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2.3}$$

The β parameter can be used to tuned the relative importance of the recall and precision. NERC evaluations often use β set to 1. In this case the F-Measure is called F1.

$$F_{(\beta=1)} = \frac{2PR}{P + R} \tag{2.4}$$

Lexical Knowledge Bases Lexical Knowledge Bases (LKB) are lexical resources (lexicons, thesauri, ontologies, dictionaries, corpus) that formalize the meanings of the words in one or more natural languages. LKBs are build with Machine-aided manual construction, with automatic acquisition from pre-existing Lexical Resources, or by a mixed approach. Lexical Knowledge Bases can be structured resources such as Machine Readable Dictionaries (MRDs) and Thesauri or unstructured resources such as corpora. The most widely used LKBs for QA are CYC, SUMO, and WordNet (including EuroWordNet and MCR).

CYC is a large database of common-sense knowledge (Lenat, 1995) that includes 100,000 concepts and 1,000,000 axioms. The Suggested Upper Merged Ontology (SUMO) (Niles & Pease, 2001) is an upper ontology created at Teknowledge Corporation and proposed as starting point for the IEEE Standard Upper Ontology Working group. SUMO provides definitions for general purpose terms and is the result of merging different free upper ontologies. It incorporates over 50 publicly available sources (e.g. Sowa's upper ontology, Allen's temporal axioms, Guarino's formal mereotopology, etc.). SUMO has approximately 1,000 concepts, 4000 assertions, and 600 rules

and includes Domain Specific Ontologies (e.g. Air force planning, Finance and investment, Terrain features, etc.).

WordNet is an electronic lexical database inspired by current psycholinguistic and computational theories of human lexical memory (Miller et al., 1990). The WordNet database was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller. This database includes English nouns, verbs, adjectives, and adverbs organized into synonym sets (called synsets), each representing one underlying lexicalized concept. Different relations link the synonym sets: synonymy, antonymy, hypernymy-hyponymy, meronymy, implication, causation. WordNet has been used widely in Text Mining applications such as: Word Sense Disambiguation, Information Retrieval, Question Answering, etc. WordNet has more than 123,000 words organized in 99,000 synsets and more than 116,000 relations between synsets

EuroWordnet is a multilingual database with wordnets for several European languages (English, Spanish, Dutch and Italian) similar to the Princeton WordNet 1.5 version (Vossen, 1997)

The Multilingual Central Repository (MCR) is a large and rich lexical knowledge base developed under the Meaning project (Atserias et al., 2004). MCR integrates five local wordnets (Basque, Catalan, English, Italian and Spanish) and the four English WordNet versions (1.5, 1.6. 1.7 and 1.7.1), EWN Top Concept ontology, MultiWordNet Domains (MWND), Suggested Upper Merged Ontology (SUMO), and large collections of semantic preferences, acquired both from SemCor[18] and from BNC, also instances, including Named Entities.

Syntactic Analysis

Syntactic analysis consists in apply a full or a shallow Natural Language parser to determine syntactic information about a text. A Natural Language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words are units (”phrases”) and which words are the subject or object of a verb.

In recent years, there have been a few attempts at creating hand-tagged corpora annotated syntactically. One idea behind creating these corpora was to make it possible for the community at large, to train supervised ML classifiers that can be used to automatically tag unseen text with syntactic information. These corpora are called Tree-banks (i.e. syntactically annotated corpora). The annotations can be dependency structure and/or syntactic relationships (constituents). The most known tree-bank is the Penn Tree-bank (PTB) (Marcus et al., 1994), which encodes constituents structures for English. PTB has 3 million words in 40.000 sentences from British Corpus and Wall-Street Journal. TIGER (Brants et al., 2002) is a tree-bank for German that contains 35.000 annotated sentences. 3LB is also a known tree-bank for Spanish, Catalan and Basque (Palomar et al., 2004).

Finally, two interesting corpora for syntactic purposes are FrameNet (created by University of California at Berkeley) and PropBank (UPenn/Colorado).

- **Full parsing**. Full parsing (i.e. deep parsing) is the process of analyzing a sentence to determine its grammatical structure given a formal grammar. Probabilistic parsers use knowledge of language gained from annotated sentences to try to produce the most likely analysis of

[18] **SemCor**. A synset-level tagged version of a subset of Penn Tree-Bank

new sentences. Lexicalized statistical parsers include lexical information in the probabilistic models.

RASP [19] (Briscoe et al., 2006) is an statistical parser that uses the most probable PoS tags to generate a a parse forest representation containing all possible subanalyses with associated probabilities. From this representation it is able to construct the n-best syntactic trees. Collins (Collins, 1999) proposed a head-driven approach to statistical parsing that outperformed the previous models. This parser achieved a 88.1% and 88.3% in labelled recall and labelled precision on the Section 23 of the WSJ Penn Tree-bank. *Spear* [20] (Ferrés et al., 2005) is a modified version of the Collins Parser. Mihai Surdeanu modified the Collin's parser and trained it on Tree-Bank version 2.0, and on an additional QuestionBank developed from the TREC 8-12 questions. *Stanford Parser* is a probabilistic natural language parser that includes both highly optimized PCFG and dependency parsers, and a lexicalized PCFG parser (Klein & Manning, 2003b) (Klein & Manning, 2003a). The *Link Grammar Parser* is a syntactic parser of English, based on link grammar, an original theory of English syntax. Given a sentence, the system assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words. The parser also produces a "constituent" representation of a sentence (showing noun phrases, verb phrases, etc.). *Bikel's parser* is parsing engine that accommodates many different types of generative, statistical parsing models (including an emulation of Mike Collins' parsing model) with equally good performance.

- **Shallow Parsing**. Shallow parsers often detect syntactic heads (verbal heads, nominal heads, adverbial heads, and adjectival and nominal modifiers) and build a dependency graph among them. These parsers normally are rule-based, and some times with statistical learnt rules (Voutilainen & Padró, 1997). *Tacat*, for instance, is a partial parser that recognizes shallow nominal, prepositional and verbal phrases (Atserias et al., 1998).

- **Chunking** Chunking process is the task of detection, delimitation, and classification of the simple syntactic fragments (syntactic phrases) of a sentence that could include a nuclear element with a non-recursive modifier (Abney, 1996). In CONLL-2000 (Tjong Kim Sang & Buchholz, 2000) a chunking task was defined using the PTB categories to propose 11 chunk types (NP: nominal phrase, VP: verbal phrase, ADVP: adverbial phrase, etc.).

- **Dependency Parsing**. Dependency parsing is based on dynamic programming techniques, constraint satisfaction, and deterministic parsing algorithms combined with Machine Learning techniques. *MINIPAR* [21] (Lin, 1998) is a largely used broad-coverage parser for English based on a constituency grammar that outputs dependency trees. It was evaluated with the manually parsed SUSANNE corpus showing that 89% of dependency relationships outputs are correct.

- **Semantic Role Labelling**. A semantic role in the relationship that a syntactic constituent has with a predicate. Typical semantic arguments include Agent, Patient, Instrument, etc.

[19] **RASP**. `http://www.informatics.susx.ac.uk/research/nlp/rasp/index.html`
[20] **Spear.** `http://www.lsi.upc.edu/~surdeanu/spear.html`
[21] **MINIPAR**. `http://www.cs.ualberta.ca/~lindek/minipar.htm`

and also adjunctive arguments indicating Locative, Temporal, Manner, Cause, etc. aspects. Recognizing and labeling semantic arguments is a key task for answering ”Who”, ”When”, ”What”, ”Where”, ”Why”, etc. questions in Information Extraction, Question Answering, Summarization, and, in general, in all NLP tasks in which some kind of semantic interpretation is needed. *ASSERT*, for instance, is an automatic statistical semantic role tagger, that is trained to tag: PropBank arguments, thematic roles, and opinions in plain text (Pradhan et al., 2004).

Semantic Analysis

Semantic analysis algorithms try to represent the meaning of isolated sentences. Often the semantic representation of the sentence text is done in logic forms. The analysis is performed using syntactic trees and sometimes conceptual ontologies such as WordNet, SUMO, or CyC. (Ferrés et al., 2004) represents sentences with a structure called *Environment*. This structure contains binary and unary predicates extracted from the sentence and extracted from a taxonomic ontology about 100 semantic classes and 25 relations. Binary predicates encode the semantic relations between the different components identified in the question text. (Bos, 2006) used Combinatory Categorial Grammar (CCG) to generate syntactic analysis of questions and potential answer snippets, and Discourse Representation Theory (DRT) is used as formalism to match the meaning of questions and answers. WordNet and NomLex are used also. A CCG parser output is used to construct a semantic representation of the question and the passages using a Discourse Representation Structure (DRS).

(Moldovan et al., 2006) uses sophisticated NLP components to transform questions and answer passages into logic representations. The LCC’s language logic prover, COGEX (Moldovan et al., 2003), does a final re-ranking of candidate answers based on the degree of semantic entailment between the candidate answer passage and the question.

Discourse Analysis

Some NLP task go beyond sentence limits for facing discourse-level interpretation. Although this contextual analysis is not common in current QA approaches some tasks as Coreference Resolution, Topic Detection, Topic-Based Segmentation, etc. are performing this level.

2.3.2 Question Classification

The Question Classification task consist in: given a question q, assign one or more class labels c_i from a class set C to the question. Question Classification for QA could be seen as a multi-class single-label or multi-label classification problem. Depending on the QA typology, question’s ambiguity among classes could be allowed and a multi-label tagging could be accepted. For example, the question *Who designed the Eiffel Tower?* in some typologies could be seen as a *Who-person* question and/or a *definee* question.

Question Classification (QC) is a crucial issue in QA because a the question class leads the Answer Extraction system to extract the correct expected answer. Consequently, question categories

strongly depend on the Named Entity set of the extraction component employed to tag the documents of the collection. Depending on the system, several entity sets were employed (typically the MUC set).

Early works in theoretical QA proposed question categorization schemes. (Lehnert, 1978) question categorization grouped together questions under 13 conceptual categories. Arthur Graesser's Taxonomy of Inquiries (Graesser et al., 1992) has foundations both in theory and in empirical research. It uses Lehnert's 13 categories to which have been added 4 new categories. Graesser showed that its taxonomy is able to accommodate all inquiries that occur in a discourse.

Open-domain QA systems started with few categories, normally related with the expected noun classes (Named Entities) to be returned as the answer and strongly based on the interrogative pronouns used in the question. The approaches to the QC task are based on manually built rules or Machine Learning techniques that use sets of lexical, semantic or syntactic features to perform the task. Manual rules based on patterns to detect questions of the same answer type (Breck et al., 1999), (Prager et al., 2000) used on particular words and on part-of-speech tags. (e.g for example if the pattern <*how (large—small—big)*> is matched, the type MEASURE is returned.) (Pasca & Harabagiu, 2001a) were the first QA system to use patterns with syntactic parsing features and semantic information from WordNet. The WebClopedia (Hovy et al., 2000) project annotated a QA typology based in the user's intention. They analyzed a set of 17,384 questions and answers to create the typology. The QA Typology contains 94 nodes, of which 47 are leaf nodes and includes classes such as Why-Famous (for *Who was Christopher Columbus?*), Abbreviation-Expansion (for *What does NASA stand for?*).

Although manually hand-crafted rules allow a rapid development of a simple QC. A low coverage and a lack of adaptability are the main problems of this approach. On the other hand Machine Learning approaches to open-domain QC have reached successful results in the last years. This methods require a large amount of data to build good classifiers automatically. (Radev et al., 2002) applied decision rule induction using Ripper with 17 question types (person, place, date, number, definition, organization, description, abbreviation, knowfor, rate, length, money, reason, duration, purpose, nominal, and other) and using for learning 1200 questions from TREC-8, TREC-9 and TREC-10. The Results of using Ripper to identify question types with primitive lexical features were 30% of error in the testing using the TREC-10 and the other collections to train. On the other hand, (Li & Roth, 2002) which used SNoW (Winnow algorithm) to learn two simple classifiers (a coarse classifier and a fine one). They used two-layer taxonomy which represents a semantic classification of typical TREC questions. The hierarchy contains 6 coarse classes (ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION and NUMERIC VALUE) and 50 finer classes. This was a successful approach, achieving a 98.80% of precision for coarse classes with question features and 95% for the fine classes over the 500 TREC-10 questions. (Zhang & Lee, 2003) experimented with five machine learning algorithms: Nearest Neighbors (NN), Nave Bayes (NB), Decision Tree (DT), Sparse Network of Winnows (SNoW), and Support Vector Machines (SVM) using two kinds of features: bag-of-words and bag-of-ngrams. They used the same two-layer taxonomy and training-testing datasets of (Li & Roth, 2002). Their experiments results showed that with only surface text features SVM outperforms the other four methods for this task. They also discussed about the importance of the syntactic structures of questions because SVMs with a kernel tree can improve the results of a single-layer SNoW using the same syntactic features. (Suzuki

et al., 2003) used a hierarchical SVM to experiment with feature sets that include words, named entities and semantic information. They measured a question type hierarchy at different depths and achieved accuracy rate ranging from 95% at depth 1 to 75% at depth 4. (Solorio et al., 2004) proposed an algorithm for a Language-independent QC based on Support Vector Machines (SVMs). Using 7 classes (person, place, date, measure, organization, object, other) They used lexical features and Internet features avoiding semantic and syntactic information. They used the DISEQuA corpus (Magnini et al., 2004) , which consists in 450 questions formulated in four languages: Dutch, English, Italian and Spanish, with 10-fold cross-validation for English, Spanish and Italian, obtaining a results of 81.77% for English, 88.70% for Italian and 81.45% for Spanish. (Li & Roth, 2004) repeated their experiments in QC using their framework test set with the use of semantic information sources for this task. Their experiments results show that semantic information can improve the performance of the QA task. Classification accuracies over 1,000 TREC-2002 questions reach 92.5% for 6 coarse classes and 89.3% percent for 50 fine grained classes. On the other hand (Shen et al., 2006) obtained a classification accuracy of 80.8 % in fine grained classes with the previous experiments. They used a Language Modelling approach with a Kneser-Ney smoothing for bigram features.

Most systems perform in parallel QC and extraction of information from the question (as question keywords, expected answer type, etc.). Keywords Selection is one of the most important steps in the Question Processing phase. Lexical terms (keywords) from the question normally used as a query to an IR/PR system lead to the relevant documents. These keywords could possibly expanded with lexical/semantic variations.

2.3.3 Passage Retrieval

Given a textual corpus and a user query (a question, a set of keywords, ...), Passage Retrieval (PR) could be defined as the task of retrieving a set of passages from the textual corpus relevant to the user query. Obviously, a passage is considered a portion of a whole document. A Passage could have with fixed size (words, bytes or sentences) or a dynamic size (paragraph, sentence,...). In QA the aim of Passage Retrieval is to get small fragments of text (with enough context) which probably contain the answer of the question. Passage Retrieval is a task similar to Information Retrieval which only passages are retrieved instead of documents. A typical PR system has normally two phases, an indexing phase and a searching phase. The first one, called Indexing, consists in processing all the collection and extract its essential information. Then, in a following step of the same process, the information is stored in a structure that allows an easy recovery of the primordial data by querying for some features.

The core of each PR system has an Information Retrieval algorithms. IR techniques can be sub-classified in tree classes depending on its mathematical model:

- **Set Models**. These models represent documents by sets. The *Standard Boolean model* is the most popular.
- **Algebraic Models**. These algorithms represent documents and queries usually as vectors, matrices or tuples. The *Vector Space model* is the algebraic model most widely used in the IR community. In the vector space model, all the documents are mapped into a N-dimensional

space in which each term represents a dimension. Each document and query is represented as a vector in this vectorial space. Document relevance with respect to a query is computed using distance measures the document vector and the query vector. Term weighting is usually performed by TFIDF (Salton & Buckley, 1988) or Okapi's BM25 (Robertson & Walker, 1994) schemas.

- **Probabilistic Models**. This models represent similarities as probabilities. In the probabilistic models the estimated relevance of a document to a query is a function of the estimated probabilities that each of the various terms in the document occur in at least one relevant document but in no irrelevant documents. Currently, *Language models* (LM) and *Divergence From Randomness* (DFR) models are ones of the most established probabilistic models.

IR Systems

- **Smart**. Smart [22](Salton & Lesk, 1965), one of the oldest, continuously running research project in information retrieval. SMART IR system was uses a vector space model for representing documents.

- **Inquery**. The INQUERY system is a product of the Center for Intelligent Information (CIIR) at the University of Massachusetts at Amherst (Callan et al., 1992). INQUERY uses a probabilistic model of information retrieval, based on Bayesian networks .

- **MG**. MG [23] (Managing Gigabytes) (Witten et al., 1999) was used by (Hovy et al., 2000), (Massot et al., 2003), (Leidner et al., 2004). MG uses a vector space model that represents documents and queries as vectors of term frequencies.

- **Zprise**. ZPrise [24] (Downey & Tice, 1999) search developed by NIST, it is a part of the Z39.50/PRISE2.0 package. ZPrise is based on vector techniques and supports term feedback; on the other hand it does not support Boolean search and has restricted phrase search capabilities.

- **Zettair**. Zettair [25] (Billerbeck et al., 2004) The zettair search engine is a publicly available system developed by the Search Engine Group at RMIT. It is intended to be a straightforward implementation of effective and efficient query evaluation techniques. Multiple ranking metrics are supported, including pivoted Cosine and Okapi BM25.

- **Lucene**. Lucene [26] IR system uses the standard tf.idf weighting scheme with the cosine similarity measure, and it allows ranked and boolean queries.

- **Xapian**. [27]. It supports the Probabilistic Information Retrieval model and also supports a rich set of boolean query operators and relevance feedback. The model provided as part of the

[22]**SMART**. `ftp://ftp.cs.cornell.edu/pub/smart/`
[23]**MG**. `http://www.nzdl.org/html/mg.html`
[24]**ZPrise**. `http://www-nlpir.nist.gov/works/papers/zp2/zp2.html`
[25]**Zettair**. `http://www.seg.rmit.edu.au/zettair`
[26]**Lucene**. `http://lucene.apache.org/java/docs/`
[27]**Xapian** `http://www.xapian.org/`

search service is the Robertson/Sparck Jones probabilistic model Boolean and phrase search facilities are also provided.

- **Terrier**. [28] Performing very well at TREC Terrier includes parameter-free probabilistic retrieval approaches such as Divergence from Randomness (DFR) models (Ounis et al., 2006). Classical TF-IDF weighting scheme to the recent language modelling approach, through the well-established Okapi's BM25 probabilistic ranking formula.

- **Indri (Lemur project)**. Indri [29] (an IR component of the Lemur toolkit) is an Information Retrieval system that supports retrieval algorithms based on Language Modelling (Ogilvie & Callan, 2001). Also includes is the OKAPI retrieval algorithm and a dot-product function using TF-IDF weighting. Lemur was largely used in TREC -2002 (Ogilvie & Callan, 2001) (Nyberg et al., 2003) (Ahn et al., 2006).

- **JIRS**. The JAVA Information Retrieval System (JIRS) software (Soriano et al., 2005) is used to retrieve relevant passages related to a question. JIRS[30] was specially designed for Question Answering (QA). This system gets passages with a high similarity between the largests n-grams of the question and the ones in the passage. It has 3 modes: simple n-gram model, term weight n-gram model, and distance n-gram model.

- **IR-n**. IR-n (Pascual, 2002) is a passage retrieval software that uses the sentence as the unit to define the passages. IR-n retrieves passages of variable size and allows passage overlapping.

- **Zebra**. Zebra [31] is a high-performance, general-purpose structured text indexing and retrieval engine. It allows exact boolean search expressions and relevance-ranked queries using a the standard tf·idf weighting scheme.

Indexing

(Rijsbergen, 1979) defines an index language as the language used to describe documents and requests. The elements of the index language are index terms, which may be derived from the text of the document to be described, or attached to it. Usually, documents are indexed using its words as an *indexed terms*. In the indexing phase some dimensional reduction techniques (Term Normalization) are applied. The most popular indexing technique is the use of Inverted Indexes, that consists in having a inverted list for each index term. Some pre-process over the terms before indexing include:

- **Stopwords Removal** tries to avoid the indexing of irrelevant information by filtering words with a so high frequency of occurrences is text that they lose their utility as search keywords. Usually articles, prepositions, pronouns, etc, are stop words-

- **Stemming**. A stemmer is an algorithm that given a word form determines its stem form. The stem is not necessarily identical to the root of the word. As an example, for English, an

[28] **Terrier**. `http://ir.dcs.gla.ac.uk/terrier/`
[29] **Indri**. `http://www.lemurproject.org/`
[30] **JIRS.** `http://leto.dsic.upv.es:8080/jirs`
[31] **Zebra**. `http://www.indexdata.dk/zebra`

stemmer will possibly identify the string "build" as the stem of the following word forms: "building", "builders". The Porter algorithm is very widely used as a standard stemmer for English (Porter, 1997). This method removes the commoner morphological and inflexional endings from words in English.

- **Lemmatization**. A lemmatizer is an algorithm that given a word form determines its lemma by using the part of speech of the word in a sentence. It requires a lexicon that store the necessary knowledge of the language (i.e. a lemma and its associated lexeme, the pair <word form, part-of-speech>). Lemmatization differs from Stemming in the fact that requires the knowledge of the POS tag of the word in the sentence and needs a knowledge base of lexemes. Stemming does not take into account the function of the word in the sentence, does not require a great knowledge of the language, and normally works by stripping morphological and inflexional endings of the words. As an example, the word "went" has "go" as a lemma, but its stem is the word form itself.

- **Multi-word indexing**. Indexing multi-words is a method to improve the performance of IR systems by avoiding that multi-word terms such as "open-minded" could be indexed separately. (Jacquemin et al., 1997) for instance, use multi-word expansion before indexing with successful results.

- **Named Entity Indexing**. Indexing Named Entities as a multi-word class can improve the recall and avoid noise in the retrieval. However, a high precision NERC is required in order to lose recall. (Prager et al., 2000) started this approach by indexing Named Entities and their class (predictive annotation). This method identifies potential answers in the text and then indexes their corresponding Named Entity class or Expected Answer Type (the author uses the term QA-token).

- **Semantic Indexing**. Using WordNet synsets to index collections can improve the recall of IR systems respect to word based indexing. (Gonzalo et al., 1998) used the SMART IR and SemCor (a disambiguated collection) to index by synsets with dubious results. In fact the increase in recall (29%) has a decrease in precision counterpart due to polysemy. What is true is that with accurate WSD module (currently not existing) the results could be good. (Mihalcea & Moldovan, 2000) experiments indexing by synsets reported also an improvement in IR effectiveness using the Cranfield collection. (Liu et al., 2004) used effectively WordNet to disambiguate word senses of query terms.

Searching

Searching documents in IR systems implies the use of a textual query in a boolean or ranked manner to obtain a set of ordered or unordered relevant documents. Boolean searches involve the use of logical operators such as: AND, OR, and NOT over the query terms to find a set of documents that satisfy the logical expression. Ranked retrieval, on the other side, does a ranking over a set of documents based on keywords similarities.

IR systems sometimes offer capabilities like phrasal search (searching for a phrase or a specific sequence of words (e.g. "Tom Cruise")), fuzzy matches (e.g. "*at" will match "Pat" or "rat"),

regular expression (regexp) matches or boosting terms (i.e. weighting search terms). A frequent approach in Searching is Query Expansion (QE). The QE approach is often used to increase the recall of the system by adding similar terms to the ones in the original query. WordNet has been used for this purpose by expanding terms with its synonyms, hyponyms, and hypernyms [32]. On the other hand Gazetteers, encyclopedic knowledge, and abbreviations, can be used in certain domains to realize QEs.

The number of documents to retrieve depends on the task. In QA, normally it depends on the document processing capability of the system. The processing capability depends on the computational resources available to process and the computational costs of the algorithms designed to process the documents. Sometimes deep NLP approaches might require expensive computational resources and processing time use only few documents (and/or passages), and some naïve approaches with lesser requirements can cope with more data [33].

In the Information Retrieval field, for research purpose the first top 1000 documents are taken into account to evaluate the systems (e.g. TREC, and CLEF adhoc IR tasks). In the real world, normally the user wants the search engines for no more than 50 documents. For QA, usually few documents/passages are used to extract the answer. In PR the searching process retrieves passages sometimes with overlapping and sometimes with fixed size. (Jorg Tiedemann, 2004) does comparison of different IR systems for QA, in which Zettair and Lucene obtained the best results.

An often used approach to improve searching is Relevance Feedback for IR/PR. Relevance Feedback (RF) consists in using most relevant terms collected from the top ranked documents of an initial query to compose manually or automatically a second query with more information. Blind feedback is a technique for automatic Query Expansion by using terms collected from the documents ranked at the top after initial retrieval.

2.3.4 Answer Extraction

The Answer Extraction phase has the aim of recover the answer(s) of a certain question. This phase normally takes place after Question Processing and Passage Retrieval and processing.

After passage processing the AE algorithms can rely on simple and fast answer pattern matching or sophisticated reasoning modules. The Answer Extraction phase is often composed by three subphases: Sentence Retrieval, Answer Ranking, Answer Selection.

Current approaches to Answer Extraction can be divided into the following points depending on the use of different NLP tasks:

1. **Pattern Matching.**

   Answer pattern matching is one of the most common approaches to the QA task. Answer patterns consists of series of regular expressions based on lexical, syntactic and/or semantic features that allows easily to match the answer sentence context to extract properly the answer. As an example, the following lexical pattern *<X;is/are;[a/an/the];A>* matches “Michigan’s

[32] Without a good WSD this kind of expansion has to be done very carefully for avoiding the introduction of noisy terms.

[33] In online-QA the response time is a critical constraint while in TREC or CLEF contests time process can be huge.

state flower is the apple blossom". On the other hand, the semantic pattern *<PERSON> was born in <BIRTHDAY>* matches "Mozart was born in 1756".

Several groups used manually built rules with great success. (Soubbotin, 2001); for instance, obtained the best results at the TREC 2001 QA evaluation task (MRR: 0.676) with a system that uses massively indicative lexical answer patterns for a broad range of question types.

(Ravichandran & Hovy, 2002) presented an approach for automatically learning answer patterns (regular expressions) from the web, for certain types of questions. Their method uses bootstrapping learning to build a large tagged corpus staring with only a few examples of QA pairs.

2. **Semantic Matching.**

   Semantic matching is performed using ontologies (e.g WordNet, SUMO, or CYC) sometimes helped by syntactic parsing structures. (Vicedo, 2002) used the Semantic Content of the Concept, a semantic representation of questions and sentences based on weights obtained by using *idf* weights and WordNet relationships: synonymy, hypernymy and hyponiny. (Ferrés et al., 2004) represents semantically sentences and questions with binary and unary predicates and applies an iterative relaxation approach by means of structural and hierarchical relaxation of predicates. (Lo & Lam, 2006) presented a system with a sophisticated grammatical framework that parses the question and candidate answers and the semantic relations are obtained. Then, these relations are compared base on the level of consistency as well as the linkages from the Wikipedia.

3. **Context-based Linguistic Features.** This method uses linguistic features from the candidate's context to perform a ranking of the candidates. FALCON (Harabagiu et al., 2000), for instance, was an early advanced QA system that applied these approach integrating semantic information using WordNet, Expected Answer Type, Query Expansion, syntactic parsing with Collins' parser and abductive reasoning.

4. **Lexical Matching with Expected Answer Type**

   Expected Answer Type (EAT) matching (Pasca & Harabagiu, 2001a) is a common strategy for the Answer selection process in most of the current QA systems. Detecting the EAT of a question could be useful in the Passage Retrieval and the Answer Extraction phases. A mapping of answer types to Named Entity types is required. During the PR phases it can be used filtering out the passages without concepts of the same category as the expected answer type. Finally, in the Answer Extraction phase the EAT can be used to select the candidates with the same type. (Pasca & Harabagiu, 2001b) use an answer taxonomy that includes 8707 concepts from 129 WordNet subhierarchies. Predictive Annotation and Virtual Annotation are also successful techniques for Answer Extraction introduced by (Prager et al., 2000).

5. **Statistical Modelling.**

   Statistical modelling for answer extraction relies in Statistical Machine Learning using annotated corpus of question-answer pairs to learn probability models.

(Whittaker et al., 2006) presented a non-linguistic multilingual data-driven statistical QA system [34] trained with the TREC QA evaluation datasets and the Knowledge Master KM data [35].

(Ittycheriah et al., 2001) created statistical algorithms for both expected answer type prediction and named entity tagging. The answer selection model used maximum entropy with the following feature sets: sentence features, entity features, definition features, and linguistic features.

6. **Cache-Based Services.**

   Although is a simple strategy, some QA systems such as QUARTZ (Jijkoun et al., 2004), Aranea (Lin & Katz, 2003) among others have a Database of question-answer pairs that it is consulted before using the QA algorithms given a question.

7. **Inference & Reasoning.** This methods require the use of ontologies and Bases of Knowledge for inferences. LCC's language logic prover, COGEX (Moldovan et al., 2003), is an example of abductive reasoning for QA.

8. **Web-based External Knowledge Mining**. Using the Web as a data source to extract the answer and then apply this information into the extraction process has emerged as new research line in QA. Major search engines and confident data sources as Wikipedia are often used. Systems such as: Aranea (Lin & Katz, 2003), PowerAnswer 3 (Moldovan et al., 2006), (Lo & Lam, 2006), Ephyra (Schlaefer et al., 2006), QASCU, (Kosseim et al., 2006) among others have used this technique.

2.3.5 QA Architectures at TREC 2006

In this subsection are presented the most relevant architectures of ODQA presented in the last QA track of TREC contest.

(Kaisser et al., 2006) use lexical resources like FrameNet, PropBank and VerbNet to generate potential answer sentences to a given question. Then these sentences are used to query the web to find the answers to the question. Then answer candidates are mapped to the AQUAINT corpus. They used frame Semantics for QA, based on syntactic parsing with Minipar and using FrameNet to extract semantic structures associated to the verbs. The best run for factoid achieved 0.323 of accuracy.

(Bos, 2006) used Combinatory Categorial Grammar (CCG) to generate syntactic analysis of questions and potential answer snippets, and Discourse Representation Theory (DRT) is used as formalism to match the meaning of questions and answers. WordNet and NomLex are used too. Indri is used for Passage Retrieval. A CCG parser output is used to construct a semantic representation of the question and the passages using a Discourse Representation Structure (DRS). if the potential passage contains a discourse referent of the answer type matching then a matching process will

[34] **AskEd**. `http://asked.jp`

[35] *Knowledge Master data*. Academic Hallmarks, `http://www.greatauk.com`. A non-free library of 142,000 questions about different subjects

start. The background knowledge (WordNet, Nomlex, specialized knowledge, handcrafted general inference rules) is used to assist the matching. The best run for factoid achieved 0.18 of accuracy.

(Whittaker et al., 2006) employed a data-driven and non-linguistic framework for QA. They use a statistical approach to find the answers. Two probability models are defined: the retrieval model and the filter model. The maximum accuracy achieved by this system was 0.251

(Harabagiu et al., 2006) presented the CHAUCER system at TREC 2006. This is a very complex system with six different strategies for Answer Extraction. A novel approach named predictive questions, that consists in creating question-answer pairs has been used. A set of semantic parsers based on PropBank, NomBank, and FrameNet in conjunction with the NERC CICEROLITE with more than 300 classes have been used to process the questions and the textual collections. An Answer Type Detection module is applied using a two-stage Maximum Entropy classifier based on (Li & Roth, 2002). Lucene IR system has been used for indexing and searching. A Keyword Expansion algorithm has been applied using the results of applying Topic Signatures to the Target-relevant documents. Documents are filtered trying to eliminate those ones that may be keyword-dense but may not contain any relevant candidate answer. This filter uses the Expected Answer Type of the question, the topic signature terms of the question topic and the keywords of the question to to re-rank the 200 first retrieved documents.

Answer Extraction techniques use the EAT, patterns, predicted question-answer pairs, and FrameNet to find answer candidates. Finally, a textual entailment system is used to select the best answer. This system scored very well in TREC-2006 with an accuracy of 53.8%.

(Moldovan et al., 2006) presented the PowerAnswer 3 QA system. This system integrates standard NLP tools such as NERC and syntactic parsing and sophisticated NLP components such as embedded ontologies, semantic relation extraction, advanced inference, coreference resolution, temporal contexts, and eXtended WordNet-based lexical chains. This system has different strategies to solve a specific class of question. The PR system ranks passages using lexical similarity and used web-boosting features (i.e. using information from Internet) to correct the errors in answer processing. Answer process uses semantic matching and scoring to extract the candidates.

The LCC's language logic prover, COGEX, is used to do the final re-ranking of candidate answers based on the degree of semantic entailment between the candidate answer passage and the question. This system reached the best accuracy at TREC 2006 QA task with 0.578 of accuracy.

(Shen et al., 2006) used a cascade of Language Modelling (LM) based document retrieval (i.e. Lemur), LM based sentence extraction, Maximum Entropy based answer extraction over a dependency relation representation followed by a fusion process that uses linear interpolation to integrate evidence from various data streams. The following NLP tools are used: Lingpipe, Abney's chunker, and MINIPAR.

(Wu & Strzalkowski, 2006) presented the ILQUA system used the BBN Identifinder NERC and the MINIPAR parser, and the INQUERY IR system. The Answer extraction methods are surface text pattern matching, n-gram proximity search, and syntactic dependency match. Patterns are automatically generated by a supervised learning system and represented in a format of regular expressions which contain multiple question terms. This system achieves a 0.266 of accuracy at TREC 2006.

(Schone et al., 2006) presented the QACTIS system at TREC 2006. This system uses the Lemur IR system with two retrieval filtering strategies: key phrase filtering and google-enhanced pseudo-relevance feedback (PRF). A reordering component based on SVMs trained on previous TREC

tracks and using syntactic and semantic (e.g. WordNet) information as features. The accuracy with the key.phrase filtering strategy achieved 0.266 and 0.236 with the google-enhanced PRF.

(Lo & Lam, 2006) presented a system with a sophisticated grammatical framework that includes the Minipar parser, the PET parser, and a semantic role labelling parser. The Wikipedia was used as a resource for detecting redundancy and hidden relations between the entities and its relations. This system uses a set of 62 question types detected by a rule-based approach. The LEMUR IR system is used for document retrieval and NERC is used to tag the NEs of the document sentences. For Answer Extraction, the question and candidate answers are parsed and the semantic relations are obtained. Then, these relations are compared base on the level of consistency as well as the linkages from the Wikipedia. The best run accuracy of this system was 0.261.

(Schlaefer et al., 2006) presented Ephyra, a QA system with two approaches: a classical expected answer type approach based on a Named Entity hierarchy of 70 NE types and another one with textual patterns to classify and extract answers from text snippets. The first approach uses the Open NLP NE tagger for broad NE classes and manually rule-based classifiers and lists for fine classes. The Indri IR system is used for paragraph retrieval and the Yahoo API is used to perform pattern retrieval. The best accuracy for factoid questions gives 0.196.

(Kosseim et al., 2006) developed the QASCU system. This is a modified version of Aranea (Lin & Katz, 2003) with a parse-tree base unifier. Aranea redundancy based QA system was modified in the following aspects: i) new data sources were added to the original google snippets, ii) re-ranking using the frequency over the top 50 documents from PRISE. In addition, a parse-tree matching algorithm to identify and rank candidate sentences was developed. This algorithm uses the Leacock and Chodorow's similarity to score the semantic relatedness of the question's main verb to each verb in the candidate sentences. Minipar parser is used to obtain the parse-trees. Then a parse-tree unification method tries to match sentences's and question parse-trees.

(Zhao et al., 2006) presented the InsunQA system at TREC 2006. The architecture of this system uses common NLP tools such as the Gate NERC, WordNet, Minipar parser and the INDRI Document Retrieval. The Answer Extraction system uses a combination of Stratified Sampling Logistic Regression and formalization. The accuracy over factoid questions is 0.298.

2.3.6 Cross-Lingual and non-English QA Systems

In the last years there is a growing interest in cross-lingual and non-English QA systems. The two major evaluation tasks that foster research in this direction are the Cross-Lingual Evaluation Forum (CLEF) and the NTCIR.

Some notable systems that deal with different languages are QRISTAL (Laurent et al., 2006), Priberam's QA system (Cassan et al., 2006), QUANTICO (Sacaleanu & Neumann, 2006).

QRISTAL is a cross-language QA system for French, English, Italian, Portuguese, Polish and Czech. This system was developed originally for French by Synapse Développement QRISTAL achieves a 65% of accuracy for French. This system uses massively NLP tools and 8 different indexing techniques and 86 question types. Priberam's QA system is a Portuguese QA system that have been extended to Spanish. This system is based on answer patterns and a large question type typology. This system achieves a 67% of accuracy on factoid and definition questions for Portuguese at QA@CLEF 2006.

QUANTICO is a cross-language QA system for German and English. The SMES parser is used for full syntactic analysis and an answer selection based on distance metrics defined over graph representations.

The previous systems are the best ones in its original language. In addition, INAOE Language Technologies Lab (Juárez-Gonzalez et al., 2006) presented a System for Spanish that achieves 40.9% of accuracy on factoid questions at QA@CLEF 2006. This system uses only lexical information without complex NLP tools (Named Entity Recognizers, Parsers, ontologies,...). INAOE's QA system uses the JIRS PR system and a Naïve Bayes classifier to extract the answer.

2.4 Evaluation of QA systems

2.4.1 QA Evaluation Frameworks

QA has become a popular task in the NL Processing (NLP) research community in the framework of different international ODQA evaluation contests such as: Text Retrieval Conference (TREC) for English (Voorhees, 2003), Cross-Lingual Evaluation Forum (CLEF) for European languages (Magnini et al., 2004), and NII-NACSIS Test Collection for IR Systems (NTCIR) for Asian languages (Yutaka Sasaki & Lin, 2005). QA evaluation contests usually provide test collections (data sets usable for experiments) and unified evaluation procedures for experiment results (Voorhees & Tice, 1999). Each participating group conducts research and experiments using the common data provided by the organization with various approaches. A detailed description of the three main QA evaluations is reported here:

- **Text Retrieval Conference (TREC).** The TREC [36] conference is the most popular international evaluation framework in the field of Information Retrieval for English. TREC started in 1992 as part of the TIPSTER Text program. It is co-sponsored by the National Institute of Standards and Technology (NIST) and Advanced Research and Development Activity (ARDA) center of the U.S. Department of Defense. It has different tracks (areas of IR) that propose different tasks related to IR. NIST provides participating groups with test sets and evaluates the results of the participants. Since 1999, a QA track has been carried out every year. The TREC conference has fostered and has inspired a substantial set of publications and current QA systems.

- **Cross-Lingual Evaluation Forum (CLEF).** The CLEF [37] is an international evaluation framework for IR in European Languages. CLEF provides the infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts. From 1997 to 1999, TREC included a track for the evaluation of Cross-Language IR for European languages (CLIR track at TREC). This track was coordinated jointly by the NIST and by a group of European volunteers. In 2000, the CLEF was launched as a successor to the TREC CLIR track (Peters & Braschler, 2001). CLEF coordination was moved to Europe, and a multinational consortium was set up. Since

[36] **TREC**. `http://trec.nist.gov`
[37] **CLEF**. `http://www.clef-campaign.org`

2001, CLEF became an independent accompanying measure sponsored within the Information Society Technologies Programme of the European Commission. Within the framework of the CLEF, a pilot track for non-English monolingual and cross-language QA systems was succesfully carried out in 2003. In 2005, the QA track was established having a total of 8 monolingual (bulgarian, german, spanish, italian, french, finnish, dutch, and portugese) and 73 bilingual tasks (Vallin et al., 2005).

- **NII-NACSIS Test Collection for IR Systems (NTCIR)**

  The NTCIR Workshop [38] is a series of evaluation workshops designed to enhance research in Information Retrieval, Question Answering, Text Summarization, Extraction, etc, emphasizing Japanese and other Asian languages. NTCIR provides large-scale test collections reusable for experiments and a common evaluation infrastructure allowing cross-system comparisons The First NTCIR Workshop started in 1998. The Japan Society for Promotion of Science (JSPS) and National Center for Science Information Systems (NACSIS) sponsored the event from 1997 to 2000. In 2000 JSPS and Research Center for Information Resources at National Institute of Informatics (RCIR/NII,) in FY 2000. MEXT Grant-in-Aid for Scientific Research on Priority Areas of ”Informatics” and RCIR/NII in and after FY2001.

2.5 Metrics of Factoid QA Systems

2.5.1 QA Capabilities

The main capabilities in QA were discussed by the QA road-map committee (Burger et al., 2000). The following capabilities are the most relevant.

- **Timeliness.** The answer to a question must be provided in real-time, and the question could refer to most recent events and facts.
- **Accuracy.** The precision of QA systems is extremely important as incorrect answers are worse than no answers. To be accurate, a QA system must incorporate world knowledge and mechanisms that mimic common sense inference.
- **Usability.** This capability implies the rapid prototyping of domain-specific knowledge and its incorporation in the open-domain ontologies, the use of heterogeneous data sources, deal with heterogeneous data formats and allow the user to describe the context of the question.
- **Completeness.** Complete answers to a user’s question is desirable. Some times answer fusion is required.
- **Relevance.** The answer to a user’s question must be relevant within a specific context. The evaluation of QA system must be user-centered: humans are the ultimate judges of the usefulness and relevance of QA systems and of the ease with which they can be used.

We will examine next the main issues on evaluation of the different components of QA systems.

[38] **NTCIR**. `http://research.nii.ac.jp/ntcir`

2.5.2 Question Processing

The Question Processing phase consists in the analysis of the question using NLP tools (morpho-syntactic analyzers, syntactic parsers, Named Entity Recognizers, semantic parsing,...). Although is not common, the evaluation of this part is an important step to avoid cumulative errors in the following phases. So, for example, Named Entity Recognition and Classification could be influenced by POS-tagging errors and semantic pre-processing could depend on the errors in the NERC and the syntactic parsing steps.

2.5.3 Question Classification

In the Question Classification phase normally, is evaluated its global accuracy (number of correct questions classified divided by the total number of questions) and and the accuracy of the classifiers for a specific class c.

$$Accuracy = \frac{\#correct\_predictions}{\#predictions} \tag{2.5}$$

$$Accuracy(c) = \frac{\#correct\_predictions\_of\_class\_c}{\#predictions\_of\_class\_c} \tag{2.6}$$

2.5.4 Passage Retrieval

Let Q be the question set, D the document (or passage) collection, $A_{D,q}$ the subset of D which contains correct answers for $q \in Q$, and $R^S_{D,q,n}$ be the n top-ranked documents (or passages) in D retrieved by a retrieval system S given question q. The following metrics of a retrieval system S for a question set Q and document collection D at rank n are defined:

- **Coverage** (or Accuracy): the Coverage, sometimes called Accuracy, gives the proportion of the question set for which a correct answer can be found within the top *n* documents retrieved for each question.

$$coverage^S(Q, D, n) = \frac{|\{q \in Q | R^S_{D,q,n} \cap A_{D,q} \neq \emptyset\}}{|Q|}$$

- **Redundancy**: the answer redundancy gives the average number, per question, of passages within the top n ranks retrieved which contain a correct answer.

$$redundancy^S(Q, D, n) = \frac{\sum_{q \in Q} |R^S_{D,q,n} \cap A_{D,q}|}{|Q|}$$

- **Maximum Redundancy**: the maximum answer redundancy any system could achieve.

$$maximum\_redundancy(Q, D, n) = \frac{\sum_{q \in Q} |A_{D,q}|}{|Q|}$$

- **Precision**: the precision of a system for a given question set and document collection at rank n is the average proportion of the n returned documents or passages that contain a correct answer.
- **Recall**: the Recall is the average proportion of answer bearing documents that are present in the top n returned documents or passages.

The most useful evaluation metrics to evaluate PR for QA are *coverage* and *redundancy*. On the other hand, *precision* and *recall* are not helpful for PR in a QA context (Roberts & Gaizauskas, 2004). Precision cannot capture the goodness of the overall queries, which is crucial for QA, the evaluation is done over a set of questions and these measures can be confusing. Recall is not as unhelpful as precision, because it can show how the retrieved document set approaches to the maximum redundancy obtainable. Redundancy, on the other hand, tells one only how many answering bearing passages per question are being returned on average. However, redundancy gives a measure of how many chances per question on average an answer extraction component has to extract an answer.

In addition, we designed two different measures to evaluate the Passage Retrieval for Factoid questions: the first one (called *answer*) is the accuracy taking into account the questions that have a correct answer in its set of passages. The second one (called *answer+docID*) is the accuracy taking into account the questions that have a minimum of one passage with a correct answer and a correct document identifier in its set of passages.

2.5.5 Answer Extraction

The evaluation of the Answer Extraction module can be done in different modes depending on the number of sub-tasks that has this module. When the Answer Extraction is a single module the evaluation takes into account the retrieved passages with a correct answer to perform an evaluation of the Answer Extraction accuracy.

$$AnwserExtractionAccuracy = \frac{\#questions\_with\_correct\_answer\_extracted}{\#questions\_with\_at\_least\_1\_passage\_that\_entails\_the\_answer} \tag{2.7}$$

Sometimes Answer Extraction uses two steps: Candidate Extraction (CE) module, and Answer Selection module. Then every step can be evaluated separately.

$$CandidatesExtractionAccuracy = \frac{\#questions\_with\_correct\_candidate\_extracted}{\#questions\_with\_at\_least\_1\_passage\_that\_entails\_the\_answer} \tag{2.8}$$

$$AnswerSelectionnAccuracy = \frac{\#questions\_with\_correct\_answer\_extracted}{\#questions\_with\_correct\_candidate\_from\_a\_supported\_passage} \tag{2.9}$$

2.5.6 Question Answering

QA judgements of factoid questions in current QA evaluations often consider a response as a single pair of answer-string and document identifier. If a pair <answer-string, document-identifier> pair is given as a response, the answer-string must contain nothing other than the answer, and the document

identifier must be the global identifier of a document in the collection that supports answer-string as an answer. Sometimes if the system detects that there is no answer in the collection the response pair reflects that the question answer is nil. These answers will be judged correct if there is no answer known to exist in the document collection; otherwise it will be judged as incorrect. An answer string must contain a complete, exact answer and nothing else. As with correctness, exactness will be in the opinion of the assessor. Responses will be judged by human assessors who will assign one of four possible judgments to a response:

- *incorrect*: the answer-string does not contain a correct answer or the answer is not responsive.
- *unsupported*: the answer-string contains a correct answer but the document returned does not support that answer (i.e does not textually entails the answer).
- *non-exact*: the answer-string contains a correct answer and the document supports that answer, but the string contains more than just the answer (or is missing parts of the answer).
- *correct*: the answer-string consists of exactly a correct answer and that answer is supported by the document returned.

Mean Reciprocal Rank (MRR)

The score of each question was the reciprocal of the rank for the first answer to be judged correct (1 or 0, or 0.333, or 0.5 points), depending on the confidence ranking. One of the most used evaluation measure in QA is the Mean Reciprocal Rank (MRR). MRR represents the mean score over all questions. MRR takes into consideration both recall and precision of the systems performance, and can range between 0 (no correct responses) and 1 (all the queries have a correct answer at position one). Two versions of MRR can be applied in a QA evaluation: a) "strict", where unsupported responses are counted as wrong, and b) "lenient" where unsupported responses are counted as correct.

$$MRR = \frac{\sum_{i=1}^{|Q|} \frac{1}{far(i)}}{|Q|} \tag{2.10}$$

Confidence-Weighted Score (CWS)

CWS was used in QA Track Trec 2002 (Voorhees, 2002) and QA@CLEF 2004 (Vallin et al., 2005) as a secondary measure. CWS was designed for systems that return one answer per question in order to evaluate own performance prediction. their own performance.

Given a question ranking based on confidence of a correct response, an analog of document retrieval's uninterpolated average precision can be computed. This measure rewards a system for a correct answer early in the ranking more than it rewards for a correct answer later in the ranking.

$$CWS = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{numbercorrectinfirstıranks}{\iota} \tag{2.11}$$

Accuracy

The accuracy measure is commonly used in all the QA evaluations (TREC, CLEF, NTCIR). The accuracy measures the precision giving the answer at the top-N rank of answers. Accuracy is the fraction of questions judged to have at least one correct answer in the first n answers to the questions. Let C be the correct answers.

$$accuracy^s(Q, D, n) = \frac{|\{q \in Q | A_{D,q,n}\}|}{|Q|} \tag{2.12}$$

F-measure

The F measure is controls the relative importance of recall and precision (Voorhees, 2003). The general formula of the F measure is:

$$F = \frac{\beta^2 PR}{(\beta^2 + 1)P + R} \tag{2.13}$$

The β parameter can be used to tune the relative importance of the recall and precision.

K1

K1 is a measure to evaluate exercises when just one answer per question is requested. This measure was introduced by (Herrera et al., 2004) in the CLEF 2004 pilot QA task and then was used in the QA track of CLEF 2005 (Vallin et al., 2005).

Chapter 3

Geographical Information Retrieval - State-of-the-art

3.1 Geographical Information Retrieval

Geographical Information Retrieval (GIR) consists in searching documents with geographically restricted queries. Geographical IR queries consist in requests that involve both thematic and geographic search (e.g. "rice exportation in Japan"). In (Sanderson & Kohler, 2004) a geographic query is defined as:

"A query which includes at least one of the following types of geographic terms: place names e.g. Houston, Texas, US; other locators e.g. postcode, ZIP code; adjectives of place e.g. American, international, western; terms descriptive of location e.g. state, country, city, site, street; geographic features e.g. island, lake; and directions e.g. north, south. "

Geographical Information Retrieval has become recently a popular task in the IR community due to the inclusion of a the GeoCLEF GIR track in CLEF 2005 and 2006 (Gey et al., 2005) (Gey et al., 2006b), and the organization of the International Workshop on Geographic Information Retrieval (GIR) in different international IR conferences SIGIR 2004 (Purves & Jones, 2004), CIKM 2005 (Jones & Purves, 2005), SIGIR 2006. And also due to the inclusion of special geographic search facilities in major search engines (Jones & Purves, 2005).

Some well-known past research projects related with GIR and Geographical QA are: *CITYTOUR*, *GeoQuery*, *SPIRIT*, and *START*. The *CITYTOUR* project (Andre et al., 1986) was designed to answer natural language questions about the spatial relationship between objects in a city. *GeoQuery*[1] demo is a learned Natural Language Interface to a US Geography Database (Zelle & Mooney, 1996) (Zelle, 1995). Geoquery contains a small database of information about United States geography and can be trained with machine learning for semantic parsing to map novel natural language queries. *SPIRIT*[2] (Spatially-Aware Information Retrieval on the Internet) was a 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

[1] **GeoQuery**. `http://www.cs.utexas.edu/users/ml/geo-demo.html`

[2] **SPIRIT project**. `http://www.geo-spirit.org`

or regions referred to in a query (Jones et al., 2002) (Jones et al., 2004). Finally, *START*[3], was the first Web-based question answering system (Katz et al., 2002) that was dealing efficiently with geography.

3.2 GIR Issues

Current IR systems, based on keyword search, are not suitable to index geographical structures, deal with geographical knowledge, reasoning,...

GIR require appropriate indexing and search structures and algorithms to determine spatial relevance. Spatial relevance similarity measures require that the following aspects had to be taken into account: hierarchical containment, adjacency of places, connectivity, proximity,....

Different multi-dimensional indexing approaches have been proposed to manage spatial data: such as grid indexes, quad-trees, R-trees, and k-d-trees (Martins et al., 2005). R-Tree, that allows efficient geographical search, is the most popular spatial indexing method (Guttman, 1984). On the other hand, c-squares is a grid indexing approach that uses grid representation and can be encoded in textual strings (Rees, 2003), so it could be easily implemented in a normal IR system.

On the other hand Geographical Knowledge and Reasoning is required to deal with geographical resolution problems. These are the common issues of Geographical Knowledge for IR:

- *Using efficiently Geographic Knowledge in GIR queries.* Although (Gey et al., 2005) reported on deteriorated performance when applying manual query expansion of geographic references. (Guillén, 2005) concludes that adding geographic information in the queries could not significantly improve retrieval performance. Metacarta (Kornai, 2005) improved its results using geographic bounding boxes, but with a bit low MAP. As reported by (Toral et al., 2006) in GeoCLEF 2005 three of the top-4 systems for the English monolingual run were based only on IR (the remaining one used geographic NER). This may be due to the fact that the systems who tried to apply geographic knowledge did not do it correctly.

- *Person-Location ambiguity problems.* It is common for proper name of persons and places to be the same and this leads to potential false associations between articles mentioning persons with such name and particular places

- *Multilinguality*. (i.e. Toponyms in different languages). In many gazetteers, mostly English names are used.

- *Name Variants*. (Leveling et al., 2005) defined these variants: i) endonymic names: a local name for a geographical entity (e.g. “Wien”, “Köln”, and “Milano”), exonym names: is a place name in a language used outside its region; (e.g. “Viena”, “Cologne”, and “Milán”), historical names: traditional names such as “New Amsterdam” for “New York”.

- *Composite Names*. Two or more words form the place name (e.g. “Mount Cook”, “Island of Sylt”).

[3] START.http://start.csail.mit.edu/

- *Semantic relations between toponyms and related concepts*. Concepts related to a toponym such as the language, inhabitants of a places, properties phrases are not considered in geographic tagging.
- *Temporal changes* in toponyms.
- *Metonymic usage*. Metonymy is defined as a figure of speech in which a speaker uses "one entity to refer to another that is related to it" (Leveling & Veiel, 2006). As an example in the following sentence: "At the meeting of France and Germany in Lisbon last year, Paris vetoed the decision", Paris is a metonymy of France.
- *Query expansion*. Adding terms to the original query in order to increase the retrieval performance can lead to obtain additional relevant documents (i.e. increasing Recall), possibly at the expense of Precision.
- *Gazetteers problems*. Incompleteness of the major gazetteers. (Fonseca et al., 2002) discuss about the problems of selecting and using gazetteers. As an example, the GeoNet Names Server geographical gazetteer presents the following problems: i) highly ambiguity on some names, ii) geographic entities that have a certain area/length (like rivers or large cities) but only a single latitude/longitude pair is given(Hauff et al., 2006), iii) bad data (out of range longitude/latitude pairs, parent information can overlap or is not fully accurate), iv) lack of data (Leveling et al., 2005) (e.g. lack of native language forms). v) relations or modifiers generate name variants not covered by a gazetteer (e.g. Southern Germany), vi) data representation may be inconsistent. (e.g. some streams or rivers are represented with only one point), vii) it does not provide sufficient information for a successful disambiguation from context (e.g. temporal information is missing). viii) incomplete ontological basis, ix) uncovered mame inflection.

3.3 GIR Approaches

Major approaches in GIR (Gey et al., 2005) include: adhoc techniques, QA modules, Gazetteer construction, Geoname Entity Extraction, Term expansion using WordNet geographic thesauri, toponym resolution, NLP-Geofiltering predicates, latitude-longitude assignment, gazetteer based query expansion, conventional IR systems, geographic entity recognition, Knowledge Bases, query expansion strategies (e.g. blind feedback, addition of proper names, geographic reference expansion using hierarchical information on GKB), and geo-spatial query restriction strategies: minimum bounding box based, geo-scope based.

Despite of the diversity of approaches at GIR, two major phases can be present in all the system architectures: Topic and Collection Processing and Document Retrieval.

3.3.1 Topic Processing and Collection Processing

Topic and Collection Processing consists in analyze the topics and/or documents of the collection in order to enrich them with useful information derived from Natural Language Analysis or a Geographical Analysis.

Linguistic Analysis

Natural Language Analysis (NLA) in GIR normally consists of applying linguistic analysis over the topics and/or the document collection for lexical purposes. Semantic parsing and lexical databases are rarely applied. Lexical analysis for GIR normally deals with Named Entity Recognition and Classification in order to detect place names. POS tagging is applied in most of these systems because sometimes is required for the NERC to have useful features.

NERC approaches applied in GIR include both Machine Learning approaches and Rule-based ones. Rule-based systems such as: GATE (García-Vega et al., 2006a) Alias-I LingPipe for NERC has been used by several groups (Yi Li et al., 2006) (Hu & Ge, 2006) (Bischoff et al., 2006), (Ferrández et al., 2005) used DRAMNERI, a rule based NERC.

(Leidner, 2005) uses a Maximum Entropy Classifier (Curran & Clark, 2003) trained with the MUC-7 data. (Lana-Serrano et al., 2005) uses a NERC based approach with a lexicon (with the GNIS and GNS gazetteers) and a grammar. (Ferrández et al., 2005) NERUA a weighted voting strategy with KNN, Maximum Entropy and HMM. (Overell et al., 2006) used ESpotter, a domain-adaptative NERC (Zhu et al., 2005).

(Ferrés et al., 2005a) used ABIONET, an Adaboost based system.

(Buscaldi et al., 2005) used the WordNet ontology in the geographical domain, by applying a query expansion method, based on the synonymy and meronymy relationships, to geographical terms. Description of the synonymy and meronymy (substitution of one word of another with which it is associated e.g. substitution of Washingnton for USA). (Buscaldi et al., 2006) used the WordNet lexical database to perform an index expansion based on synonymy and holonymy relations.

(Leveling & Veiel, 2006) employed multilayered extended semantic networks for the representation of knowledge, queries and documents for GIR with the syntactico-semantic parser (WOCADI).

Geographical Analysis

Geographical Analysis of the topics and documents may consists on using Geographical Knowledge Bases (GKB) and Toponym Resolution algorithms. GKBs are used in order to detect geographical place names and its possible referents. Toponym Resolution is applied to decide which referent is used in a certain context.

Geographical Knowledge Bases. Geographical Knowledge Bases can be defined as geospatial dictionaries of geographic names with some relationships among place names. Usually these places can be political and administrative areas, natural features, and man-made structures. Relationships among place names are commonly downward (parent-child) relations (e.g. Asia - China) and upward (e.g. Germany - Europe). On the other hand some approaches define other relationships, (Hu & Ge, 2006) GKB includes relationships between entities such as part-of adjacency and similar (e.g. if two entities have a similarity such as being administrative divisions of the same country or if they are countries, ...). (Lana-Serrano et al., 2005) provided a flexible structure that allows define other types of relationships between resources: based on its languages(latin america, anglo-saxon countries) or religion (catholic, protestant,...). Geographical tagging, annotation scheme that allows us to specify the geographical path to the entity.

The most commonly used GKBs in GeoCLEF evaluations are publicly available huge gazetteers such as: GeoNet Names Server, GNIS, WorldGazetteer .... WorldGazetteer is widely used due to its population statistics (Cardoso et al., 2005), (Leidner, 2005), (Ferrés & Rodríguez, 2006b) (Gey et al., 2005). Some groups are using the Wikipedia to collect information (Cardoso et al., 2005). Only (Overell et al., 2006) and (Yi Li et al., 2006) used the Getty Thesaurus of Geographic Names, a private Gazetteer.

GIR systems often tend to merge some these gazetteers into a unique one. (Hauff et al., 2006) used a merge of GNS, GNIS, and World Gazetteer (WG), that provides information about the parent-child relationships. (Andogah, 2006) used geographic resources such as Wikipedia, World-Gazetteer, GeoNet names server, and WordNet. (Hu & Ge, 2006) joined several resources to build a GKB: a) FIPS 10-4 for countries and administrative divisions, b) World Factbook for border countries, coastlines, country capital cities, c) Wikipedia for oceans, seas, gulfs, rivers and regions, d) A set of large cities collected from TravelGis.com, e) The Standard Country and Area Codes Classifications (M49) for regions and continents, f) The ESRI Gazetteer server developed by the Environmental Systems Research Institute, Inc. for Minimum Boundary Rectangle (MBR) of countries, and g) WordNet for variant places names. (Toral et al., 2006) used Geonames DB [4]. (Leveling & Veiel, 2006) GNS was also employed to extract geographical knowledge.

Toponym Resolution. Toponym Resolution is used in several GIR approaches. TR Algorithms usually decide the best referent candidate among a set of possible referents for a place name applying a set of heuristics (see Chapter 4 for a detailed explanation of the TR methods). (Cardoso et al., 2005) at GeoCLEF 2005 and (Martins et al., 2006) at GeoCLEF 2006 used the one single scope per document heuristic (Martins & Silva, 2005) with a PageRank variation graph based algorithm. (Leidner, 2005) used a maximum-population heuristic. (Overell et al., 2006) applied co-occurrence models trained with Wikipedia for place name disambiguation with a Naïve Bayes. (Yi Li et al., 2006) used a probabilistic approach for toponym resolution based on the following evidences: *local contextual information, population information, Trigger Words, global contextual information,* and *Mutual disambiguation.*

(Leveling & Veiel, 2006) implemented a metonymic location classifier training with the manual annotated data from the GERMAN CONLL-2003 shared task. and a subset of the GeoCLEF newspaper corpus. The features uses were shallow (post tags, position of words in a sentence, word length and base forms of verbs). The classifier achieves a performance of 81.7% of F1-measure in differentiating between literal and metonymic senses of location names.

On the other hand few groups apply geo-disambiguation to resolve the *personorganization - location* ambiguity (i.e distinguish if the candidate was correctly tagged as a toponym or is really a person name or an organization name). (Ferrés et al., 2005a) apply a NEC correction filter to correct these errors. This filter stores in a hash table all the tokens that compose the NEs classified as *person*. Then *location* or *organization* NEs are checked against the hash table. (Li et al., 2006) apply a set of rules for resolving the location-person ambiguity.

[4]Geonames. `http://www.geonames.org`

3.3.2 Document Retrieval

The main goal of this phase is to retrieve a set of relevant documents to the topic. The main process of this phase is the Information Retrieval process which normally requires the use of an IR system. This phase can be complemented by a Query Expansion phase and a post phase of Document Filtering.

Query Expansion

Query Expansion techniques in IR usually consist in adding related terms to the query manually or automatically in order to retrieve more relevant documents. In GIR is also normal to use normal IR QE techniques in order to modify the thematic search. For instance, (García-Vega et al., 2006a) performed a thesaurus-based expansion using words with a high rate of document co-occurrence. But for geographical IR, normally terms geographically related to the topic terms are added to the query. The GIR Query Expansion can be done by several heuristics based on spatial relations and location type.

Before Query Expansion, the desired keywords are extracted to compose the query. Some groups apply special algorithms for this Query Processing or Query Parsing step. (Hu & Ge, 2006) did Query parsing consisted in removing guidance information (e.g. “documents about”) and stopwords, and abbreviations are expanded using WordNet API. (Toral et al., 2006) collected required words and geographical items. Required words are all the nouns of the topic, description and narrative without geographic ones , stopwords and guidance information,

Sometimes document expansion is applied previously. Document expansion and query expansion techniques are used to match the location in a query to all its gazetteer children and nearby locations. (Yi Li et al., 2006) used a geographic-based query expansion, using a gazetteer to extend geospatial terms to “nearby” locations, and included sublocations. A geo-term in the query may be expanded upwards (for “close/near “ relations, influencing all or some of its ancestors) or downwards (for “in” relations, extending the influence to all of its descendants in the gazetteer hierarchy).

(Lana-Serrano et al., 2006) applied a geographic and spatial relation identifier and a expander to compute the points located in a geographic are whose centroid is known. The expansion is determined by the geographic type of the feature and its spatial relation. (Andogah, 2006) uses query expansion with geo-references without significant improvement over the simple thematic textual search. (Leidner, 2005), (Buscaldi et al., 2005), and (Leveling & Veiel, 2006) applied query expansion with meronyms (e.g. for California, “Orange County” and “Los Angeles” are included), and (Toral et al., 2006) and (García-Vega et al., 2006a) used automatic query expansion consisting in expanding the locations of the topics with geographical information from Geonames gazetteer. (Leveling & Veiel, 2006) also employed multilayered extended semantic networks for the representation of knowledge, queries and documents for GIR. Geographical concepts from the query network are expanded with semantically connected via topological, directional, and proximity relations. (Overell et al., 2006) and (Hauff et al., 2006) did manual query processing for some topics at GeoCLEF 2006.

3.3.3 Information Retrieval

Information Retrieval approaches for GIR often use combined search (i.e. both thematic and geographical search). There are few systems that do not use Geographical Knowledge (GK) in IR (Gey et al., 2005), (Guillén, 2005), ,(Guillén, 2006), and (Toral et al., 2006) (with a system without GK). But these systems, based only in pure IR techniques achieve the best results in the GeoCLEF evaluations. It seems that Geographical Knowledge is not properly used by most of the GIR groups involved in the GeoCLEF tasks.

Boolean models are rarely used for GIR, if used only for geographical searches (Ferrés & Rodríguez, 2006b) (Bischoff et al., 2006). Most of the IR engines at GeoCLEF are based on the Vector Space Model (Lucene, SMART, Zettair, Zebra, etc.) or Probabilistic frameworks (Lemur (Indri), Terrier, Zapian, etc). Lucene with a TFIDF weighting scheme is used frequently by many groups (Leidner, 2005), (Buscaldi et al., 2005), (Buscaldi et al., 2006), (Hu & Ge, 2006), (Andogah, 2006), among others. This system is preferentially used for thematic search rather than for geographical search. The Lemur toolkit (Indri) was also used for several groups: (Guillén, 2006), (García-Vega et al., 2006a), and (Hauff et al., 2006). Passage Retrieval was used by few groups: (Ferrés & Rodríguez, 2006b) used JIRS for thematic and geographical search, and (Ferrández et al., 2005) (Toral et al., 2006) used IR-n. On the other hand also RDBMS systems were used specially for geographical isolated search (i.e. queries with only geographical terms): Postgres (used by (Overell et al., 2006)), MySQL (used by (Hu & Ge, 2006)), and (Toral et al., 2006) used SQL queries over the Geonames DB.

Normal textual indexing is vastly used for all the systems. Some of them take profit of the "field search" capabilities of some IR search engines. Several systems indexed separately textual terms and geographical terms. (Andogah, 2006), for instance, indexed and searched separately geographical relevant terms (place names, geo-spatial relations, geographic concepts and geographic adjectives) and thematic terms. (Ferrés et al., 2005a) and (Yi Li et al., 2006) used hierarchically expanded geo-terms indexing (i.e. a concatenated string consisting of a candidate and its ancestors in the gazetteer). (Li et al., 2006) performed this idea with a different way: utilizes the inverted index to store all the explicit and implicit locations of documents.

Only three systems employed indexing structures specially designed for Geographical IR: R-Tree structures were used by (Overell et al., 2006), (Li et al., 2006) used grid indexing with a textual index IRengine dividing the surface of the earth into 1000x2000 grids, and (Kornai, 2005) used the Metacarta search engine with a bounding box derivation scheme.

Relevance Feedback (which consists in performing a new retrieval loop with a set of manually or automatically collected terms from the initial retrieved passages) has emerged as a efficient method for improving the results in GIR. Systems such as (Guillén, 2005) and (Gey et al., 2005) achieved the best results at GeoCLEF 2005 using Relevance Feedback(RF) techniques with a Probabilistic IR approach. RF techniques include: Blind Feedback ((Martins et al., 2006) and (Bischoff et al., 2006)), Automatic Relevance Feedback ((Ruiz et al., 2006)), and Pseudo-Relevance Feedback (García-Vega et al., 2006a).

Term weighting schemas applied for GIR systems are: TF-IDF, BM25, DFR, binary TF, and Boolean. TF-IDF and Okapi's BM25 are the most widely used.

Pre-processing techniques such as stemming, stopwords removal are extensively used in most

of the systems. Porter's stemmer in combination with the SMART stop words list are used in some GIR systems: (Guillén, 2005), (Hu & Ge, 2006), and (Ruiz et al., 2006), etc. Finally, (Ferrés & Rodríguez, 2006b) uses lemmatization instead of stemming.

Document Filtering

Document filtering strategies for GIR try to filter out geographically irrelevant documents by using GKBs. (Hauff et al., 2006) retrieved by content and subsequently filtered by geographical relevance using a gazetteer and coordinates restrictions. (Leidner, 2005) used Geographic Filtering to filter out documents that do not fall in the area of interest using Minimal bounding Rectangles (MBR) to approximate the polygons described by the locations in the query.

Document Ranking

The Document Ranking (DR) phase consists in combining scores from thematic search and geographic search (i.e. geographically isolated terms search). Relevant approaches include linear interpolation (Leidner, 2005) (Andogah, 2006) and geographic similarity ranking (Martins et al., 2006).

3.4 Geographical Information Retrieval Evaluations

This Section describes the GIR evaluation task existing GeoCLEF.: the GIR task is reported and the results of GeoCLEF 2005 and GeoCLEF 2006 are presented.

3.4.1 GeoCLEF

GeoCLEF is Geographical IR task in the CLEF evaluation framework. GeoCLEF started as a cross-language geographic retrieval task at the CLEF 2005 campaign (Gey et al., 2005). The goal of the GeoCLEF task is to find as many relevant documents as possible from the document collections, using a topic set . Topics at GeoCLEF 2005 were textual descriptions with the following fields: title, description, narrative, location (e.g. geographical places like continents, regions, countries, cities, etc.) and a geographical operator (e.g. spatial relations like in, near, north of, etc.) (see an example of a topic in Figure 3.1). In GeoCLEF2006 the topics do not contain explicit expressions with geographic references and geographic operators (see an example of a topic in Figure 3.2. This implies that geographical references (geographic places, and geographic relations) are embedded in the title , description, and narrative. In addition, new geographic relationship were added, such as geographic distance (e.g. within 100km of Frankfurt) and complex geographic expressions (e.g. Northern Germany).

The relevance judgements are binary, i.e. the document either meets the information need expressed in a topic (1) or not (0) (Leidner, 2005).

Several kinds of geographical topics can be considered (Gey et al., 2006a):

1. *non-geographic subject restricted to a place* (e.g. music festivals in Germany).

```
<num> GC001 </num>
<orignum> C084 </orignum>
<EN-title> Shark Attacks off Australia and California </EN-title>
<EN-desc> Documents will report any information relating to shark
attacks on humans. </EN-desc>
<EN-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. </EN-narr>
<EN-concept> Shark Attacks </EN-concept>
<EN-spatialrelation> near </EN-spatialrelation>
<EN-location> Australia </EN-location>
<EN-location> California </EN-location>
```

Figure 3.1: Example of a topic of the GeoCLEF 2006 edition.

2. *geographic subject with non-geographic restriction* (rivers with vineyards)
3. *geographic subject restricted to a place* (cities in Germany)
4. *non-geographic subject associated to a place* (independence, concern, economic handlings to favour/harm that region, etc.) Examples: independence of Quebec, love for Peru (as often remarked, this is frequently, but not necessarily, associated to the metonymical use of place names)
5. *non-geographic subject that is a complex function of place* (for example, place is a function of topic) (European football cup matches, winners of Eurovision Song Contest)
6. *geographical relations among places* (how are the Himalayas related to Nepal? Are they inside? Do the Himalaya mountains cross Nepal's borders? etc.)
7. *geographical relations among (places associated to) events* (Did Waterloo occur more north than the battle of X? Were the findings of Lucy more to the south than those of the Cromagnon in Spain?)
8. *relations between events which require their precise localization* (Was it the same river that flooded last year and in which killings occurred in the XVth century?)

3.4.2 GIR Systems at GeoCLEF Evaluations

The results of the GeoCLEF evaluations (GeoCLEF 2005 and GeoCLEF 2006) are presented in Table 3.1 and Table 3.2. The best results were obtained by Probabilistic IR systems (including Logistic Regression) that use BM25 or DFR term weighting models and do not use Geographical Knowledge: (Gey et al., 2005) and (Guillén, 2005) at GeoCLEF 2005 and (Gey et al., 2006b),(Guillén, 2006), (Toral et al., 2006) at GeoCLEF 2006. Finally, it must be take into account that some of the top-performing systems (Gey et al., 2005) and (Martins et al., 2006) achieved good results using manual Query Expansion.

```
<num>GC027</num>
<EN-title>Cities within 100km of Frankfurt</EN-title>
<EN-desc>Documents about cities within 100 kilometers of the city of
Frankfurt in Western Germany</EN-desc>
<EN-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</ENnarr>

<num> GC034 </num>
<EN-title> Malaria in the tropics </EN-title>
<EN-desc> Malaria outbreaks in tropical regions and preventive
vaccination </EN-desc>
<EN-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.</EN-narr>
```

Figure 3.2: Example of a topic of the GeoCLEF 2006 edition.

3.5 IR Evaluation Measures

There are various ways to measure how well the retrieved information matches the desired information:

3.5.1 Precision

The proportion of retrieved and relevant documents to all the documents retrieved:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{3.1}$$

Precision can also be evaluated at a given cut-off rank, denoted *P@n*, instead of all retrieved documents.

3.5.2 Recall

The proportion of relevant documents that are retrieved, out of all relevant documents available:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{3.2}$$

3.5.3 Fall-Out

The probability to find an irrelevant among the retrieved documents.

IR-Approach	Group	IR-Weighting	GeoKB	QE	RF	Best MAP
Logistic Regression	Berkeley 1 (Gey et al., 2005)	BM25	no	auto	BF	0.3936
		BM25	no	manual	BF	0.3550
	Berkeley 2 (Gey et al., 2005)	BM25	yes	auto	-	0.2924
		BM25	yes	auto	-	0.3879
Language Modelling	(Guillén, 2005)	-	no	auto	PRF	0.2694
	(Guillén, 2005)	-	yes	auto	PRF	0.1362
Vector Space Model	(Leidner, 2005)	TFIDF	yes	yes	-	0.1850
	(Cardoso et al., 2005)	TFIDF	yes	manual	-	0.2253
	(Kornai, 2005)	TFIDF	yes	auto	-	0.1700
	(Buscaldi et al., 2005)	TFIDF	yes	auto	-	0.1464
	(Ferrés et al., 2005a)	TFIDF	yes	auto	-	0.2231
	(Ferrández et al., 2005)	BM25	no	auto	-	0.3495
	(Kornai, 2005)	TFIDF	yes	auto	-	0.1700
Probabilistic Model	(Lana-Serrano et al., 2005)	trie	yes	auto	-	0.2653
	(Guillén, 2005)	DFR	no	auto	PRF	0.3616
	(Guillén, 2005)	DFR	yes	auto	PRF	0.3032

Table 3.1: GIR approaches in the context of the GeoCLEF 2005 evaluation.

$$\text{fall-out} = \frac{|\{\text{irrelevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{3.3}$$

3.5.4 F1-measure

The weighted harmonic mean of precision and recall.

$$F = \frac{PR}{2P + R} \tag{3.4}$$

3.5.5 Mean Average Precision

Over a set of queries, find the mean of the average precisions, where Average Precision is the average of the precision after each relevant document is retrieved.

Where r is the rank, N the number retrieved, rel() a binary function on the relevance of a given rank, and P() precision at a given cut-off rank:

$$AveP = \frac{\sum_{r=1}^{N}(P(r) \times \text{rel}(r))}{\text{number of relevant documents}} \tag{3.5}$$

IR-System	Group	IR-Weighting	GeoKB	QE	RF	Best MAP
Logistic Regression	Berkeley 1 (Gey et al., 2006b)	BM25	no	auto	BF	0.2656
		BM25	no	manual	BF	0.2887
Language Modelling	(Hauff et al., 2006)	-	yes	auto	-	0.1875
	(Ferrés & Rodríguez, 2006b)	-	yes	auto	-	0.1370
Vector Space Model	(Martins et al., 2006)	BM25	yes	manual	-	0.2080
		BM25	yes	manual	BF	0.2150
	(Buscaldi et al., 2006)	TFIDF	yes	auto	-	0.2660
	(Hu & Ge, 2006)	TFIDF	yes	auto	-	0.2758
	(Lana-Serrano et al., 2006)	BM25	yes	auto	-	0.2000
	(Andogah, 2006)	TFIDF	yes	auto	-	0.2195
	(Yi Li et al., 2006)	BM25	yes	auto	-	0.2464
	(García-Vega et al., 2006a)	mix	yes	auto	-	0.2403
	(García-Vega et al., 2006b)	BM25	yes	auto	PRF	0.2403
	(Ruiz et al., 2006)	TFIDF	yes	manual	RF	0.2446
	(Ruiz et al., 2006)	TFIDF	yes	auto	RF	0.2344
	(Bischoff et al., 2006)	boolean	yes	auto	BF	0.1875
	(Overell et al., 2006)	binTF	yes	auto	-	0.1953
	(Li et al., 2006)	BM25	yes	manual	-	0.2395
	(Li et al., 2006)	BM25	yes	auto	-	0.2000
Probabilistic Model	(Lana-Serrano et al., 2005)	trie	yes	auto	-	0.2653
	(Guillén, 2006)	DFR	no	auto	-	0.2857
	(Toral et al., 2006)	DFR	no	auto	-	0.2985
		DFR	yes	auto	-	0.1201

Table 3.2: GIR approaches in the context of the GeoCLEF 2006 evaluation.

Chapter 4

Geographical Information Resolution - State-of-the-art

Geographical Information Resolution (GIRE) means the automatic understanding of the geographical concepts appearing in an electronic text. GIRE implies that every geographical concept in the text must be recognized, classified in a fine geographical ontology and disambiguated into its geographical world referent.

GIRE technologies start by initially finding Named Entities using broad classes (i.e. distinguish among LOCATION, PERSON, ORGANIZATION and so). The main issue of this phase for GIRE systems is to have a good accuracy of the Geo/Non-Geo classification system. Large-scale geographical gazetteers ontologies and other lexical resources, as the Alexandria Digital Library Gazetteer (Hill, 2000) covering about 5 million of geographical terms, or the Metacarta GazDB (Axelrod, 2003) are used for Geographical NERC purpose.

Then, applying some fine subclassification using extended NE hierarchies as the Perseus system (Smith & Crane, 2001) or (Ferrés et al., 2004b) for geographical NEs. (Sekine et al., 2002) uses an extended NE hierarchy of 150 types and (Manov et al., 2003) use 97 classes for the location sub-ontology.

At this point sometimes a geographical place name type disambiguation procedure must be applied to decide at which subclass pertains the place name (e.g. in some contexts the Named Entity "Buffalo" could be a city or a River). A Geographical Name Place Class Disambiguator normally tries to disambiguate those NEs using features from the document in which the Named Entity appears and optionally features from external resources to decide in which subclass pertains. Finally, the last process is the Grounding of geographical NEs, i.e. mapping a geographical NE to its appropriate physical (spatial) location (coordinates, area, etc.), as in (Leidner et al., 2003).

From the different approaches to resolve this task we can see that three kind of knowledge to extract features has normally been used:

- Local linguistic context: this features are commonly used by all the systems that try to disambiguate NEs. The features are internal or external context evidences. Some approaches use hand-written rules (Smith & Crane, 2001) or Machine Learning techniques such as Bayesian Learning, Decision Trees (Fleischman, 2001), Inductive Logic Programming (ILP) (Ferrés

et al., 2004b).

- Document context and General Knowledge: (Smith & Crane, 2001) use an heuristic technique of calculating weighted centroids of geographic focus in documents. Similar heuristics have been applied in (Rauch et al., 2003) using the fact that exists a high degree of spatial correlation in geographic references that occur in textual proximity.

- Domain Knowledge: (Rauch et al., 2003) use population heuristics to disambiguate NEs using the supposition that a place with high population is more likely to be mentioned than a place with a lower one.

4.1 Geographical Gazetteers

Geographical Gazetteers can be defined as geospatial dictionaries of geographic names. Normally these places can be political and administrative areas, natural features, and man-made structures. They contain large lists of geographical entities, normally enriched with some information such as: place name class (e.g. city, country), location (e.g. geographical coordinates such as longitude and latitude), elevation, population, language, inclusive relations (e.g. referent of the state or country where is located). We shortly describe some of the most relevant geographical gazetteers.

- **GEOnet Names Server (GNS[1])**. A worldwide database of geographic feature names, excluding the United States and Antarctica, with 5.3 million entries. The coordinate system for data served by GNS is WGS84. Each gazetteer entry contains a geographical name (toponym) and its geographical coordinates (latitude, longitude), language of the geographical name and other features as country, first administrative division, etc.

- **Geographic Names Information System** (GNIS[2]). A gazetteer with 2.0 million entries about geographic features of the United States and its territories.

- **Alexandria Digital Gazetteer (DGL).** [3] (Frew et al., 1998) . The ADL gazetteer, a geospatially defined geographic name datasets or a place-name index under 4 million entries, allows a user to find earth features by typing in the name associated with that feature, e.g., find the city of Santa Barbara or find all references to the name "Santa Barbara" worldwide. The gazetteer database may be used as a spatial finding aid or as a stand-alone reference tool.

- **GeoWorldMap**[4] gazetteer with approximately 40,594 entries (countries, regions and important cities).

- **UN-LOCODE**. The official gazetteer by the United Nations [5], with more than 36.000 locations in 234 countries.

[1] **GNS**. http://earth-info.nga.mil/gns/html
[2] **GNIS**. http://geonames.usgs.gov/geonames/stategaz
[3] **ADL**. `http://www.alexandria.ucsb.edu/gazetteer/`
[4] **Geobytes Inc.**: `http://www.geobytes.com/`
[5] **UN-LOCODE.** http://www.unece.org/cefact/locode/service/main.htm

- **Getty Thesaurus of Geographic Names (TGN)**. This gazetteer was compiled by the Getty Research Institute. The TGN includes names and associated information about places. Places in TGN include administrative political entities (e.g., cities, nations) and physical features (e.g., mountains, rivers). Current and historical places are included. The TGN is a structured vocabulary currently containing around 1,102,000 names and other information about places. Names for a place may include names in the vernacular language, English, other languages, historical names, names and in natural order and inverted order. Among these names, one is flagged as the preferred name. There are around 911,000 places in the TGN hierarchy with geographic coordinates, notes, sources for the data, and place types, role of the place (e.g., inhabited place and state capital) and temporal information coverage.

- **Heavens-Above GmbH Gazetteer**. Heavens Above is a private company which offers a gazetteer data to specify geographic location in order to orient sky charts, satellite fly-overs, etc.

- **Global Gazetteer**. Worldwide directory about 3,397,140 cities and towns (excluding U.S.A.), sorted by country and linked to a map for each town.

- **World Gazetteer**. [6]: a gazetteer with approximately 171,021 entries of towns, administrative divisions and agglomerations with their features and current population.

- **Geonames**[7]. The *Geonames* geographical database contains over eight million geographical names and consists of 6.2 million unique features whereof 2.2 million populated places and 1.8 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes. Geonames is integrating geographical data such as names, altitude, population and others from various sources. All lat/long coordinates are in WGS84 (World Geodetic System 1984). The sources used by this KB are: *NGA*: National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names (most names except US and CA), *GNIS* : U.S. Geological Survey Geographic Names Information System (names in US), `www.geobase.ca` (names in CA), *gtopo30* (elevation data), *Wikipedia*.

- **Pertaynims Gazetteers**. A set of nationalities-countries lists were obtained automatically from WordNet, (pertaynims). As shown in (Greenwood, 2004), pertaynims are useful for IR queries for QA, because answers to questions which include a location often occur in close proximity to the adjective form of the location, hence including the adjective form in the IR query increase the coverage of the retrieved documents.

4.2 Toponym Resolution

Toponym Resolution (TR) means grounding a place name to its real world physical location (coordinates). TR algorithms normally decide the best referent candidate among a set of possible referents for a place name applying a set of heuristics.

[6] **World Gazetteer**. `http://www.world-gazetteer.com/`
[7] **Geonames.** `http://www.geonames.org`

Geographical ambiguity problems treated by TR systems include:

- **Referent ambiguity problem.** This problem occurs when the same name is used for several locations (of the same or different class). Some authors (Li et al., 2003) note the similarity of this problem to the Word Sense Disambiguation (WSD) problem.

  In a question, sometimes it is impossible to solve this ambiguity, and, in this case, we have to accept as correct all of the possible interpretations (or a superclass of them). Otherwise, a trigger phrase pattern can be used to resolve the ambiguity (e.g. ”Madrid” is an ambiguous NE, but in the phrase, ”comunidad de Madrid” (State of Madrid), ambiguity is solved).

  The basic approaches to this problem are:

  1. **One referent per discourse**. a similar approach to the WSD work ”one sense per discourse” ((Gale et al., 1992)). In this method for WSD, it is assumed that a word appearing in a discourse refers to the same sense throughout the discourse. The approach for geographical referent disambiguation is to assume that a place name used in a discourse refers to the same location throughout the discourse ((Leidner et al., 2003)).
  2. **Proximity of place names**. ”There is a high degree of spatial correlation in geographic references that are in textual proximity”. (Rauch et al., 2003) uses some heuristics increase c(p,n) based on how many and which points (and enclosing regions) are mentioned in the same document s n and their proximity.
  3. **Spatial minimality heuristic**. This approach tries to disambiguate places assuming that the small region that is able to ground the whole set of places appearing in the discourse is the correct interpretation of these toponyms ((Leidner et al., 2003).
  4. **Contextual Pattern Matching**. Applying contextual patterns (e.g. location1 at South of location2, city of X) is the most widely used approach (Li et al., 2002), (Rauch et al., 2003), (Manov et al., 2003).
  5. **Population heuristics**. Population data in geographical gazetteers is used in different ways: ignoring small places and/or promoting dense populated place. (Rauch et al., 2003) assumed that ”A place with a high population is more likely mentioned than a place with a lower one”.
  6. **Co-occurrence models**. (Li et al., 2003) used discourse features based on co-occurring toponyms (e.g., a document with “Buffalo” , “Albany” and “Rochester” will likely have those toponyms disambiguated to New York state). (Overell et al., 2006) applied co-occurrence models trained with Wikipedia for place name disambiguation with a Naïve Bayes.
  7. **Use of default**. Some methods set a default location when a place name is ambiguous, the most common heuristic to decide the default place is the use of the candidate with the largest population.

- **Reference ambiguity problem.** This problem occurs when the same location can have more than one name (in Spanish texts this frequently occurs as many place names occur in languages other than Spanish, as Basque, Catalan or Galician). Knowledge sources as GNS or

TGN are used to deal with this problem. For instance, (Luque et al., 2006) applies a grouping process over GNS to create groups of place names that refer to the same locations. On the other hand, (Leveling & Veiel, 2006) implemented a metonymic location classifier trained with the manual annotated data from the GERMAN CONLL-2003 shared task. The classifier achieves a performance of 81.7% of F1-measure in differentiating between literal and metonymic senses of location names.

- **Referent Class Ambiguity**. The same name can be used for locations and also for other classes of Named Entities like persons or organizations. (Ferrés et al., 2005a) apply a NEC correction filter to correct these errors. This filter stores in a hash table all the tokens that compose the NEs classified as *person*. Then *location* or *organization* NEs are checked against the hash table. (Li et al., 2006) apply a set of rules for resolving the location-person ambiguity.

4.2.1 Toponym Resolution Architectures

This part presents some of the most relevant Toponym Resolution architectures:

- (Rauch et al., 2003) use data mining procedures and domain knowledge repositories (such as first names) to generate sets of contexts with positive or negative indicators . Positive context for geographic names could be trigger words before of after a name (e.g. ”city”,”mayor”, ”community college”).

- (Overell et al., 2006) applied co-occurrence models trained with Wikipedia for place name disambiguation with a Naïve Bayes classifier.

- (Garbin & Mani, 2005) describes a corpus-based method for disambiguating toponyms with and unsupervised machine learning system that develops disambiguation rules. They used the ALTAS Gazetteer and the World Gazetteer and the LexScan tool. They used a Human Annotated Corpus of news (from TimeBank 1.2, and Gigaword NYT Sept. 2001 and June 2) (Section 5). This corpus contains 83,872 words with 1275 place names (435 distinct) annotated with 3 geographical classes: *national capital*, *civil politicaladministrative region*, and *populated place*. This method achieves a 78.5% of accuracy in the human-annotated corpus.

- (Leidner, 2006) Presented the first systematic account of the utility of different heuristics for the toponym resolution task, based on experimental comparison on two novel large-scale gold-standard annotated corpora: *TR-CoNLL* (a gold-standard corpus of nearly 1,000 news articles from CoNLL 2003 with the correct referents annotated by humans) and *TR-MUC4* (an annotated corpus of 100 MUC-4 documents focused on Central America). Both corpora were annotated with these populated place classes: city, state, country, and continent. (Leidner, 2006) replicated two methods: Perseus (Smith & Crane, 2001) and LSW03 (Leidner et al., 2003) for a set of large-scale experiments. LSW03 outperformed Perseus in both corpora. LSW03 achieved 0.4736 and 0.4598 of Toponym Score (see the explanation of these evaluation metric in the next Section) in TR-CoNLL and TR-MUC4 respectively. Perseus achieved 0.3431 and 0.4023 of Toponym Score in the same corpora.

- (Yi Li et al., 2006) used a probabilistic approach for toponym resolution based on a five-level normalization of the gazetteer. Assigning more probabilities to the top levels (country or nations). Initial probabilities are also adjusted based on the following evidences: *local contextual information*: for example, geo-types in close proximity to each other (e.g. city, state), *population information*, *Trigger Words*. (e.g. "county", "river, etc.), *global contextual information*, occurrences in the document of country geo-terms that are gazetteer ancestors to the candidate, and *Mutual disambiguation*: Candidates taht are closely related to each other in the gazetteer hierarchy boost each others' probability assignment for their respective terms. They used a hand annotated subset of the GeoCLEF corpus to determine the performance of the Named Entity Classification System, and the toponym disambiguation algorithm. The corpus consists of a set of 106 Glasgow Herald and 196 LA times news articles, which contained 2311 tagged locations in total. LingPipe achieved a 50% of Precision and a 65% of recall . The TR algorithm achieved an accuracy of 90.3% on the 1502 place names identified by LingPipe. The disambiguation accuracy with respect to the total number of total locations achieve an accuracy of 60.8%.

4.3 Evaluation Metrics

(Leidner, 2006) proposed an adaptation of traditional NERC methods for the toponym resolution task. The following methods can be re-casted:

4.3.1 Precision

The proportion of toponyms resolved correctly to all the toponyms resolved:

$$\text{Precision (P)} = \frac{\text{\#toponyms resolved correctly}}{\text{\#toponyms resolved}} \tag{4.1}$$

4.3.2 Coverage

The proportion of toponyms resolved correctly to all the toponyms resolved:

$$\text{Coverage (C)} = \frac{\text{\#toponyms resolved}}{\text{\#total number of toponyms}} \tag{4.2}$$

4.3.3 Toponym Score

The *Toponym Score* is a measure that relates geometrically *Precision* and *Coverage*:

$$\text{Toponym Score } T_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{C}} \tag{4.3}$$

Chapter 5

TALP-QA Question Answering Approach

5.1 TALP-QA Question Answering Approach

This chapter describes TALP-QA, a multilingual open-domain Question Answering (QA) system under development at UPC for the past 3 years. A first version of TALP-QA for Spanish was used to participate in the CLEF 2004 Spanish QA track (see (Ferrés et al., 2004)). From this version, a new version for English was built and was used in TREC 2004 (Ferrés et al., 2005). Then an improvement of this version was used to participate in CLEF 2005 (Ferrés et al., 2005c) and TREC 2005 (Ferrés et al., 2005b).

5.2 TALP-QA Question Answering Architecture

The system architecture for factoid questions has three subsystems (as shown in Figure 5.1 that are executed sequentially without feedback: Question Processing (QP), Passage Retrieval (PR) and Answer Extraction (AE). This section describes the three main subsystems and a Collection Pre-processing process.

5.2.1 Collection Pre-processing

We pre-processed the document collections with linguistic tools for Spanish and English (described in the next subsection) to mark the part-of-speech (POS) tags, lemmas, Named Entities (NE), and syntactic chunks. The EFE 1994 and EFE 1995 collections where used in CLEF evaluations for Spanish language and the AQUAINT collection (i.e. about 1 million documents) was used for the TREC QA track.

Then, the collections were indexed separately and we computed the idf weight at document level for the whole collection. We used the *Lucene*[1] [2] Information Retrieval (IR) engine to create

[1] **Lucene**. `http://jakarta.apache.org/lucene`

[2] In previous versions of the system, MG was used instead of Lucene.

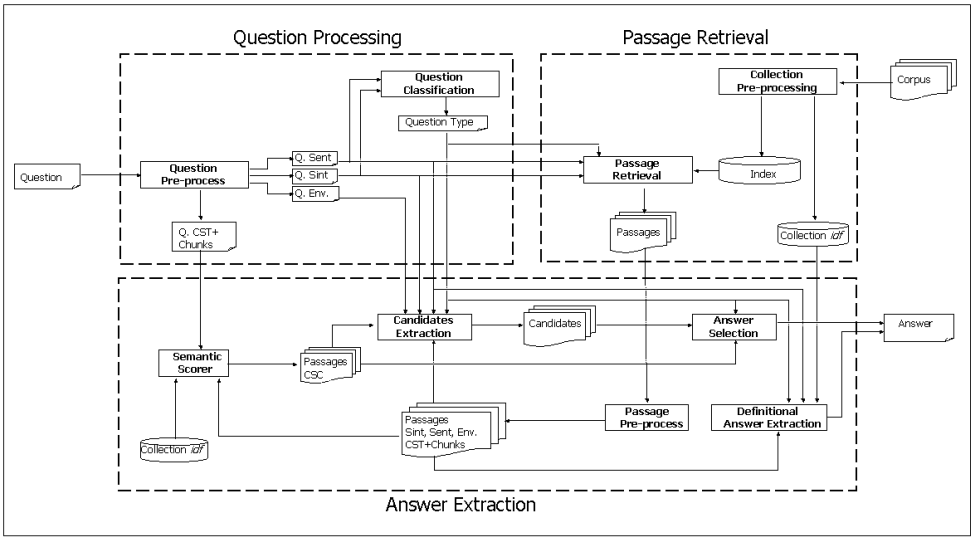

Figure 5.1: Architecture of TALP-QA system.

the indexes.

In CLEF QA (Spanish) the index contains two fields per document: i) the lemmatized text with NERC and syntactic information, ii) the original text (forms) with NER (not classified) and syntactic information. For English (at TREC evaluations) we built an index with two fields per document: i) the lemmatized text with POS tags, and the recognized Named Entities with its class, ii) the original text (forms) with Named Entity Recognition. The first field is used in a search by lemma, and the information of both fields is retrieved when a query succeeds.

5.2.2 Question Processing

Our first QP module, used for English and Spanish at CLEF 2004 and TREC 2004 contests was based on a ILP approach (FOIL) for learning independent binary classifiers for each kind of question complemented with manually created default rules for assuring full coverage. The results were acceptable for English but very low for Spanish. For CLEF 2005, We decided to build a new Question Processing (QP) module with two objectives: i) improving the accuracy of our QC component and ii) providing better material for allowing a more accurate semantic pre-processing of the question. The QP module is structured into five components, we will describe next these components focusing in those having changed from our previous system (see (Ferrés et al., 2004) for details).

- **Question Pre-processing**.

  The main goal of this subsystem is to detect the expected answer type and to generate the information needed for the other subsystems. For PR, the information needed is basically lexical (POS and lemmas) and syntactic, and for AE, lexical, syntactic and semantic. We use a language-independent formalism to represent this information. We use the same semantic

primitives and relations for both languages (English and Spanish) processed by our system.

For Spanish we used a set of general purpose tools produced by the UPC NLP group (see (Carreras et al., 2004) and (Atserias et al., 1998)):

- **FreeLing**, which performs tokenization, morphological analysis (including identification of quantities, dates, multi-word terms, etc.), POS tagging and lemmatization. See (Carreras et al., 2004).
- **Tacat**, a partial parser that recognizes shallow nominal, prepositional and verbal phrases. See (Atserias et al., 1998).
- **ABIONET**, a Named Entity Recognizer and Classifier that classifies NEs in basic categories (person, place, organization and other) (Carreras et al., 2002). This NERC was trained with the CoNLL-2002 Spanish data set.
- **EuroWordNet (EWN)**, used to obtain the following semantic information: a list of synsets (with no attempt at Word Sense Disambiguation), a list of hypernyms of each synset (up to the top of each hypernymy chain), the EWN's Top Concept Ontology (TCO) class (Rodríguez et al., 1998), and Magnini's Domain Codes (DC) (Magnini & Cavagliá, 2000).
- **Gazetteers**, with the following information: acronyms, obtained using a Decision Tree approach (Ferrés et al., 2004a), location-nationality relations (e.g. España-español, Spain-Spanish) and actor-action relations (e.g. escribir-escritor, write-writer).
- **Geographical gazetteers**. Due to the limited amount of context in questions, the accuracy of our NER and NEC components suffers a severe fall, specially serious when dealing with locatives (a 46% of NEC errors in the CLEF 2004 questions analysis were related with locatives). For this reason, we used geographical gazetteers to improve the accuracy of the NEC task. The gazetteers used were: a subset of 126,941 non-ambiguous places from the GEOnet Names Server (GNS)[3], the *GeoWorldMap* gazetteer with approximately 40,594 entries (countries, regions and important cities), and *Albayzin Gazetteer* (a gazetteer of 758 place names of Spain existing in the speech corpus Albayzin (Diaz et al., 1998)).
- **FreeLing Measure Recognizer and Classifier**. A module for a fine-grained classification of measures and units has been created. This module was added to *Freeling* and it recognizes the following measure classes: *acceleration*, *density*, *digital*, *dimension*, *energy*, *extent*, *flow*, *frequency*, *power*, *pressure*, *size*, *speed*, *temperature*, *time*, and *weight*.
- **Temporal expressions grammar**. This process recognizes complex temporal expressions both in the questions and in the passages. It is a recognizer based on a grammar of temporal expressions (composed by 73 rules) which detects four types of such expressions:

[3] **GNS**. `http://gnswww.nima.mil/geonames/GNS/index.jsp`

  - *Date*: A specific day, including day, day of the week (most times calculated), month and year (and eventually the time).
  - *Date_range*: Period of time, spanning between two specific dates or expressions such as ”in 1910” (which would be equivalent to the period between January 1st 1910 and December 31st 1910), but also the seasons or other well-known periods of the year.
  - *Date_previous*: the period previous to a specific date.
  - *Date_after*: the period subsequent to a specific date.

  Moreover, in all the four types, not only absolute dates or periods are detected, but also dates relative to the current date, in expressions such as ”el próximo viernes” (next Friday),”ayer” (yesterday), or ”a partir de mañana” (from tomorrow on). These relative dates are converted into absolute according to the date of the document in which they are found.

The following tools were used to process English:

- **Morphological components**, an statistical POS tagger (*TnT*) (Brants, 2000) and the WordNet lemmatizer (version 2.0) are used to obtain POS tags and lemmas. We used the *TnT* pre-defined model trained on the Wall Street Journal corpus.
- **Spear**. A modified version of the Collins parser, which performs full parsing and robust detection of verbal predicate arguments (Collins, 1999). For the purpose of question answering, we have limited the number of predicate arguments to three: agent, direct object (or theme), and indirect object (benefactive or instrument), and use a series of robust heuristics to identify them. For example, one heuristic labels a noun phrase as agent if it precedes an active verb within a sentence construct. Furthermore, we have retrained the parser on a corpus of questions (SBARQ and SQ phrases) which are lacking in the original Penn Tree-Bank. From the previous TREC evaluations (TREC 8 to TREC 11) we have constructed an additional corpus of 1769 questions for training and a corpus of 537 questions for testing. Using this training corpus in addition to the Tree-Bank, our parser boosts its F-measure on the question test corpus from 81.82% to 95.10%.
- **ABIONET**, trained for English with the CoNLL-2003 data set.
- **Alembic**, a Named Entity Recognizer and Classifier that identifies and classifies NEs with MUC classes (person, place, organization, date, time, percent and money). See (Aberdeen et al., 1995).
- **WordNet 1.5**.
- **Three Gazetteers**, with the following information: acronyms, obtained using a Decision Tree approach (Ferrés et al., 2004a); location-nationality relations (e.g. Spain-Spanish) and actor-action relations (e.g. write-writer).

These tools are used for the linguistic processing of both questions and passages. See the example in Figure 5.2. The application of the language dependent linguistic resources and tools to the text of the question results in two structures:

– **Sent**, which provides lexical information for each word: form, lemma, POS tag (Eagles tag-set for Spanish and Penn-Tree-Bank (PTB) tag-set for English), semantic class of NE, list of EWN synsets and, finally, whenever possible the verbs associated with the actor and the relations between some locations (specially countries) and their gentiles (e.g. nationality).

– **Sint**, composed of two lists, one recording the syntactic constituent structure of the question (basically nominal, prepositional and verbal phrases) and the other collecting the information of dependencies and other relations between these components.

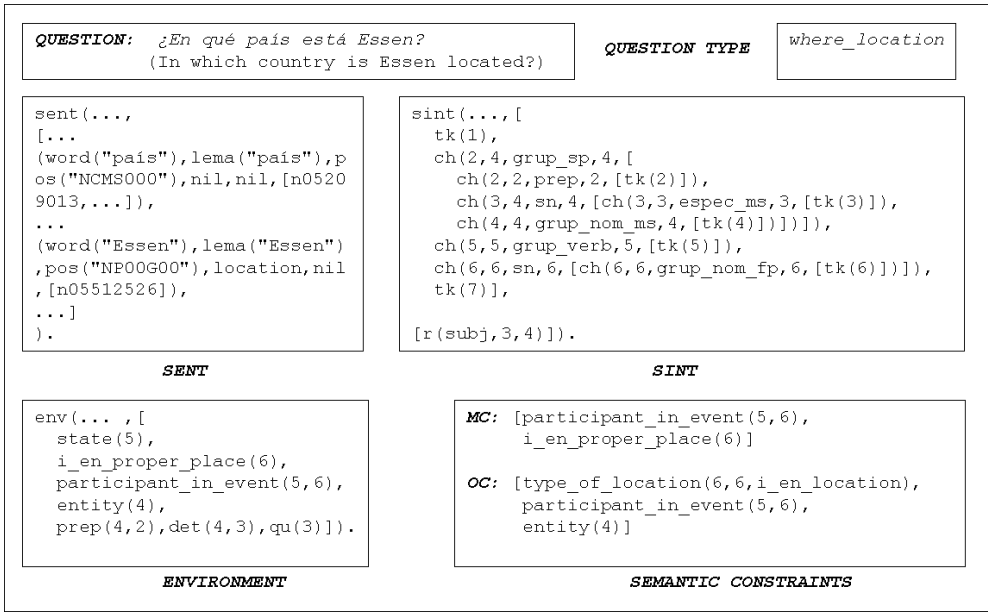

Figure 5.2: Results of pre-processing of a question.

- **Question Refinement**. This module contains two components: a tokenizer and a parser (processing the lexical structure of Question Pre-processing step). The tokenizer refines and sometimes modifies the *sent* structure. Basically the changes can affect the NEs occurring in the question and their local context (both the segmentation and the classification can be affected). Taking evidences from the local context a NE can be refined (e.g. its label can change from location to city), reclassified (e.g. passing from location to organization), merged with another NE, etc. Most of the work of the tokenizer relies on a set of trigger words associated to NE types, especially locations. We have collected this set from the Albayzin corpus (a corpus of about 6,887 question patterns in Spanish on Spain's geography domain, (Diaz et al., 1998)). The parser uses a DCG grammar learned from the Albayzin corpus and tuned with the CLEF 2004 questions. In addition of triggers, the grammar uses a set of introducers, patterns of lemmas as "dónde" (where), "qué ciudad" (which city), etc. also collected from Albayzin corpus.

- **Environment Building.** The semantic process starts with the extraction of the semantic relations that hold between the different components identified in the question text. These relations are organized into an ontology of about 100 semantic classes and 25 relations (mostly binary) between them. Both classes and relations are related by taxonomic links. The ontology tries to reflect what is needed for an appropriate representation of the semantic environment of the question (and the expected answer). The environment of the question is obtained from *Sint* and *Sent*. A set of about 150 rules was built to perform this task. Refer to (Ferrés et al., 2004) for details.

- **Question Classification**.

  The most important information we need to extract from the question text is the Question Type (QT), which is needed by the system when searching the answer. Failure to identity the QT practically disables the correct extraction of the answer. Currently we are working with about 26 QTs (we have used the same categories used in TREC 2003 (Massot et al., 2003)).

  The Question Types used were:

```
- abreviation
- abreviation_expansion
- definee
- definition
- event_related_to
- feature_of_person
- howlong_event
- howlong_object
- howmany_objects
- howmany_people
- howmuch_action
- non_human_actor_of_action
- subclass_of
- synonymous
- theme_of_event
- translation
- when_action
- when_begins
- when_person_died
- where_action
- where_location
- where_organization
- where_person_died
- where_quality
- who_action
- who_person_quality
```

  The QT focuses the type of expected answer and provides additional constraints. For instance, when the expected type of the answer is a person, two types of questions are considered, *Who_action*, which indicates that we are looking for a person who performs a certain

action and *Who_person_quality*, that indicates that we are looking for a person having the desired quality. The action and the quality are the parameters of the corresponding QT. The following are examples of questions correctly classified respectively as *Who_person_quality* and *Who_action* type:

- *Who was the head of the XII Israel government?*
- *Who won the Nobel Prize for Literature in 1994?*

The results of QC in CLEF 2004 were rather low (only 58.33% accuracy). As was explained in (Ferrés et al., 2004) the low accuracy obtained is basically due to two facts: i) the dependence on errors of previous tasks (Ferrés et al., 2004), ii) the question classifier was trained with the manual translation of questions from TREC 8 and TREC 9 (about 900 questions). The classifier performs better in English (74% (171/230)) than in Spanish (58.33% (105/180)), probably due to the artificial origin of the training material. Next, we will describe our two approaches for Question Classification.

- **ILP-based approach**. In order to determine the QT our system at CLEF 2004 and TREC 2004 used an Inductive Logic Programming (ILP) learner that learns a set of weighted rules from a set of positive and negative examples. We used as learner the FOIL system (Quinlan & Cameron-Jones, 1993). A binary classifier (i.e. a set of rules) was learned for each QT. As training set we used the set of questions from TREC 8 and 9 ($\sim$900 questions) manually tagged and as test set the 500 questions from TREC 11. For each classifier we have used as negative examples the questions belonging to the other classes. For the classification task the following features were used: form, position in the question, lemma, POS, semantic class of NE, synsets together with all their hypernyms, TCO, DC and subject and object relations.

  The set of rules for each class was manually revised and completed with a set of manually built rules (with lower priority) in order to ensure a greater coverage. See below a couple of such rules:

  * A learned rule:

```
rule(non_human_actor_of_action,A,weight_1):-
  first_position(A,B),
  next_position(B,C),
  is_tco(cObject,C),
  is_domain(dTransport,C).
```

  * The same rule after transformation (performed for the sake of efficiency):

```
rule(non_human_actor_of_action,A,weight_1,
    [],TT) :-
  sent(A,_,TT), TT=[_,W2|_],
  has_tco(W2,cObject),
  has_domain(W2,dTransport).
```

  This rule performs as follows: using the question number (*A*), the *Sent* of the sentence is retrieved (*TT*). Then, the information about the second token from the sentence is obtained. Finally, we check that this token has a TCO (EWN top concept

ontology value) corresponding to the class *Object* and a Domain Code corresponding to the class *Transport*.

* A manual rule:

```
rule(non_human_actor_of_action,A,weight_994,
    [T1,T3],T) :-
  sent(A,_,[T1|T]),
  the_lema(T1,lema("which")),
  has_chunk_with_hyperonym(_,T,[T2|TT],
    [sArtifact,sObject,sAnimal],T3),
  the_pos(T2,pos("IN")),
  not(has_term_with_pos(TT,pos("JJS"),_)).
```

The manual rule performs as follows: using the question number (*A*), the first token of the sentence (*T1*) and the following tokens of the sentence (*T*) are retrieved. Then, we check that the first token has "which" as lemma and the token's list *T* has a chunk with a token having an hypernym corresponding to one of the following synsets: artifact, object or animal. Finally, we check that the first token of the chunk (*T2*) is a preposition or a subordinating conjunction and does not contain a superlative adjective in its text.

– **Manual-rules approach based on linguistic introducers and trigger words**. This approach was used in CLEF 2005 and TREC 2005. This component uses 72 hand made rules to extract the Question Type (QT). These rules use a set of introducers (e.g. 'where'), and the predicates extracted from the environment (e.g. location, state, action,...) to detect the QT (currently, 25 types). The QT is needed by the system when searching the answer. The QT focuses the type of expected answer and provides additional constraints.

- **Semantic Constraints Extraction.** Depending on the QT, a subset of useful items of the environment has to be selected in order to extract the answer. Sometimes additional relations, not present in the environment, are used and sometimes the relations extracted from the environment are extended, refined or modified. We define in this way the set of relations (the semantic constraints) that are supposed to be found in the answer. The Semantic Constraints Set (SCS) is the set of semantic relations that are supposed to be found in the sentences containing the answer. The SCS of a question is built basically from its environment. The environment tries to represent the whole semantic content of the question while the SCS should represent a part of the semantic content of the sentence containing the answer. Mapping from the environment into the SCS is not straightforward. Some of the relations belonging to the environment are placed directly in the SCS, some are removed and some are modified (usually to become more general) and, finally, some new relations are added (e.g. *type_of_location*, *type_of_temporal_unit*,..., frequently derived from the question focus words). Relations of SCS are classified into two classes: Mandatory Constraints (MC) and Optional Constraints (OC). MC relations have to be satisfied in the passage. If a OC relation is satisfied the score of the answer is higher.

  In order to build the semantic constraints for each question a set of rules (typically 1 or 2 for each type of question) has been manually built. A set of 88 rules is used. The environment

```
get_semantic_constraints(Question,MC,OC,Environment,where_location,1) :-
  …
  state(C,Question,Environment),
  get_related_tokens_in_environment(C,Environment,ListRelatedTokens),
  filter_tuple_tokens(ListRelatedTokens,MC,_,OC,
          [theme_of_event,time_of_event,location_of_event,which_entity],
          []),
  …
  filter_related_tokens(ListRelatedTokens,
          [
               [human_participant_in_event(C,_X)],
               [participant_in_event(C,_X), i_en_proper_person(_X)],
               [participant_in_event(C,_X), i_en_proper_organization(_X)],
               [participant_in_event(C,_X), i_en_proper_named_entity(_X)]
          ],
          MCRelations),
  …
  extend_mandatory(ListRelatedTokens,MCRelations,MC,OC,Question,Environment).
```

Figure 5.3: A rule to obtain the Semantic constraints of a question.

is basically a first order formula with variables denoted by natural numbers (corresponding to the tokens in the question). Several auxiliary predicates over this kind of formulas are provided and can be used in these rules. Usually these predicates allow the inclusion of filters, the possibility of recursive application and other generalization issues. A fragment of the rule applied in the example is presented in Figure 5.3. The rule can be paraphrased as follows: If the relation *state(C)* holds in the environment, get recursively all the predicates related to C, then filter the appropriate ones to be included in MC and OC and finally extend these sets for the sake of completeness. The application of the rule results in the constraints shown in Figure 5.2. The binary and unary predicates that compose the environment are shown in this Figure. The unary predicates extracted are:

- *state(4)*: which corresponds to the verb ”is”.
- *i_en_proper_place(5)*: which corresponds to the Named Entity ”Essen”. This unary predicate specifies that Essen is a NE classified as a location.
- *entity(3)*: a common noun corresponding to ”country”.
- *qu(2)*: corresponds to the interrogative pronoun ”which”.

The binary predicates extracted are:

- *participant_in_event(6,5)*: a semantic relation between the Named Entity ”Essen” and the verb ”locate”.
- *prep(3,1)* and *det(2,1)*: syntactic relations without semantic content.

From the binary and unary predicates the SCS extracted is:

- *participant_in_event(locate,Essen)*. (MC)
- *i_en_proper_place(Essen)*. (MC)

– *type_of_location(country,country, i_en_location)*. (OC)

– *entity(country)*. (OC)

5.2.3 Passage Retrieval

The main function of the passage retrieval component is to extract small text passages that are likely to contain the correct answer. Document retrieval is performed using the *Lucene* Information Retrieval system. For practical purposes we currently limit the number of documents retrieved for each query to 1000. The passage retrieval algorithm uses a data-driven query relaxation technique: if too few passages are retrieved, the query is relaxed first by increasing the accepted keyword proximity and then by discarding the keywords with the lowest priority. The reverse happens when too many passages are extracted. Each keyword is assigned a priority using a series of heuristics fairly similar to (Moldovan et al., 1999). For example, a proper noun is assigned a higher priority than a common noun, the question focus word (e.g. ”state” in the question ”What state has the most Indians?”) is assigned the lowest priority, and stop words are removed.

The basic Passage Retrieval subsystem has been improved with the following components:

- **Temporal Constraints Keywords Search**. When a keyword is a temporal expression, the PR system returns passages that have a temporal expression that satisfies the constraint detected by our temporal grammar.

- **Coreference resolution**. We apply a coreference resolution algorithm to the retrieved passages. This algorithm is applied to enhance the recall in the Answer Extraction modules. We use an adaptation of the limited-knowledge algorithm proposed in (Saiz, 2002). We start by clustering the Named Entities in every passage according to the similarity of their forms (trying to capture phenomena as acronyms). For Named Entities classified as Person we use a first name gazetteer[4] to classify them as masculine or feminine. By the clustering procedure we get the gender information for the occurrences of the name where the first name does not appear. After that, we detect the omitted pronouns and the clause boundaries using the method explained in (Ferrández & Peral, 2000), and then apply the criteria of (Saiz, 2002) to find the antecedent of reflexive, demonstrative, personal and omitted pronouns among the noun phrases in the 4 previous clauses.

5.2.4 Factoid Answer Extraction

After PR, for factoid AE, two tasks are performed in sequence: Candidate Extraction (CE) and Answer Selection (AS). In the first component, all the candidate answers are extracted from the highest scoring sentences of the selected passages. In the second component the best candidate is chosen as answer.

[4]By Mark Kantrowitz, `http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names`

- **Candidate Extraction**. The answer extraction process is carried out on the set of passages obtained from the previous subsystem. These passages are segmented into sentences and each sentence is scored according to its semantic content (see (Massot et al., 2003)). The linguistic process of extraction is similar to the process carried out on questions and leads to the construction of the environment of each candidate sentence. The rest is a mapping between the semantic relations contained in this environment and the semantic constraints extracted from the question. The mandatory restrictions must be satisfied for the sentence to be taken into consideration; the satisfaction of the optional constraints simply increases the score of the candidate. The final extraction process is carried out on the sentences satisfying this filter.

  The knowledge source used for this process is a set of extraction rules with a credibility score. Each QT has its own subset of extraction rules that leads to the selection of the answer. The application of the rules follows an iterative approach (see Figure 5.5). In the first iteration all the semantic constraints must be satisfied by at least one of the candidate sentences. If no sentence has satisfied the constraints, the set of semantic constraints is relaxed by means of structural or semantic relaxation rules, using the semantic ontology. Two kinds of relaxation are considered: i) moving some constraint from MC to OC and ii) relaxing some constraint in MC substituting it for another more general in the taxonomy. If no candidate sentence occurs when all possible relaxations have been performed the question is assumed to have no answer.

  An example of an extraction rule is presented in Figure 5.4. The rule can be paraphrased as follows: Look in MC for predicates *state(C)* and *location(X)* satisfied in the environment. Then look in the environment for the predicates related to C, *location_of_event* and *location*. Make sure that the two locations are different and adjust the corresponding score.

```
extract_contextual_answer_from_tokens(DS,SS,_,_,Env, where_location,1, MT,A1,Sc2,_):-
  satisfy_MT_esp_obl([state(C),location(X)],MT,_),Sc=10,
  satisfy_strict([location_of_event(C,A,DS,Env),location(A,DS,Env)]),
  X\==A,
  nth(A,SS,A1),
  nth(X,SS,A2),
  smooth_scr(SS,X,A,Sc,Sc1),
if(
satisfy_MT_esp_obl([type_of_location(_,_,TL)],MT,_),
(check_type_of_location(A1,TL,A2,Sc3),Sc3 > 0.4, Sc2 is (Sc1 + Sc3 * 10)/ 2),
Sc2 is Sc1).
```

Figure 5.4: One of the extraction rules used in the example.

- **Answer selection**. In order to select the answer from the set of candidates, the following scores are computed for each candidate sentence: i) the rule score (which uses factors such as the confidence of the rule used, the relevance of the OC satisfied in the matching, and the similarity between NEs occurring in the candidate sentence and the question), ii) the passage score, iii) the semantic score (defined previously) , iv) the relaxation score (which takes into account the level of rule relaxation in which the candidate has been extracted). For each candidate the values of these scores are normalized and accumulated in a global score. The

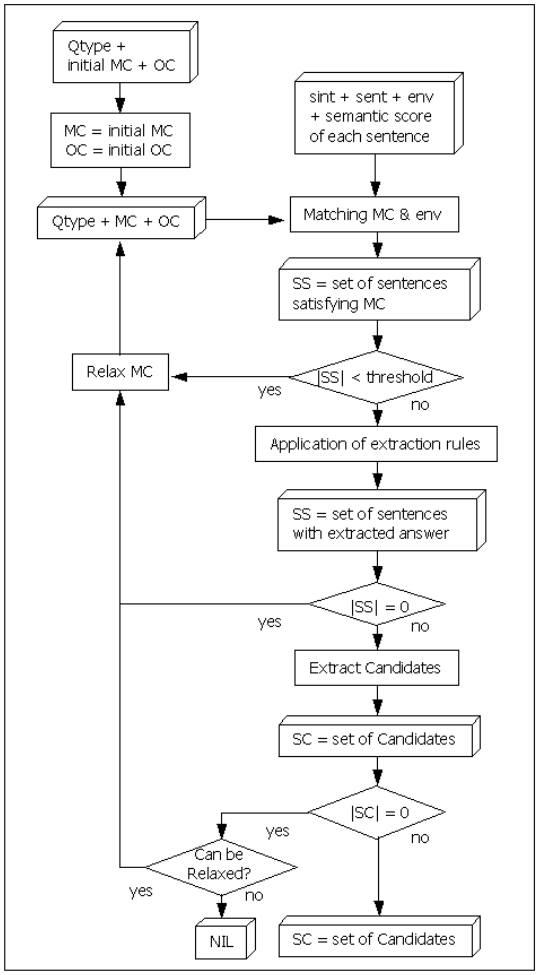

Figure 5.5: Candidates Extraction Relaxation Loop.

answer to the question is the candidate with the best global score.

5.3 Evaluation and Results at CLEF 2004

This section evaluates the behaviour of our system in CLEF 2004. From the 200 questions proposed in these evaluation, 180 were factoid and 20 were definitional. We evaluated the three main components of our system and the global results:

- **Question Processing**. This subsystem has been manually evaluated for factoid questions (see Table 5.1) and the following components: basic NLP tools (POS, NER and NE Classification (NEC)), semantic pre-processing (Environment, MC and OC construction) and finally,

Question Classification. These results are accumulatives.

Table 5.1: Results of Question Processing evaluation.

Subsystem	Total units	Correct	Incorrect	Accuracy	Error
POS-tagging	1667	1629	38	97.72%	2.28%
NE Recognition	183	175	8	95.63%	4.37%
NE Classification	183	137	46	74.86%	25.14%
Environment	180	81	99	45.00%	55.00%
MC	180	77	103	42.78%	57.22%
OC	180	131	49	72.78%	27.22%
Q. Classification	180	105	75	58.33%	41.67%

- **Passage Retrieval**. The evaluation of this subsystem was performed using the set of correct answers given by the CLEF organization (see Table 5.2). We submitted two runs. In both runs we retrieved only the 1000 top documents (no passages) for definition questions. These runs differ only in the parameters of the passage retrieval module for factoid questions:

  - Windows proximity: in run1 the proximity of the different windows that can compose a passage was lower than run2's (from 60 lemmas to 80).
  - Threshold for minimum passages: the PR algorithm relaxes the query to obtain more passages if the number of extracted passages is lower than this threshold. These values are: 4 (run1) and 1 (run2) passages.
  - Number of passages retrieved: we have chosen a maximum of 3000 passages in run1 and 50 passages in run2.

Table 5.2: Passage Retrieval results.

Question type	Measure	run1	run2
FACTOID	Accuracy (*answer*)	64.37% (103/160)	59.37% (95/160)
	Accuracy (*answer+docID*)	48.12% (77/160)	43.12% (69/160)

In this part we computed two measures: the first one (called *answer*) is the accuracy taking into account the questions that have a correct answer in its set of passages. The second one (called *answer+docID*) is the accuracy taking into account the questions that have a minimum of one passage with a correct answer and a correct document identifier in its set of passages.

- **Answer Extraction**. The evaluation of this subsystem for factoid questions has been done in three parts: evaluation of the Candidate Extraction (CE) module, evaluation of the Answer Selection (AS) module and finally evaluation of the AE subsystem's global accuracy for

factoid questions in which the answer appears in our selected passages.

Table 5.3: Factoid Answer Extraction results.

Subsystem	Measure	run1	run2
Candidate Extraction	Accuracy (*answer*)	33.00% (34/103)	35.78% (34/95)
Answer Selection	Accuracy (*answer*)	70.58% (24/34)	79.41% (27/34)
Answer Extraction	Accuracy (*answer*)	23.30% (24/103)	28.42% (27/95)

- **Global Results**. The overall results of our participation in CLEF 2004 are listed in Table 5.4.

Table 5.4: Results of TALP-QA system at CLEF 2004.

Measure	run1	run2
Total Num. Answers	200	200
Right/Wrong	48/150	52/143
IneXact/Unsupported	1/1	3/2
Overall accuracy	24.00% (48/200)	26.00% (52/200)
Accuracy over Factoid	18.89% (34/180)	21.11% (38/180)
Accuracy over Definition	70.00% (14/20)	70.00% (14/20)
Answer-string ”NIL” returned correctly	19.23% (10/52)	20.37% (11/54)
Confidence-weighted Score	0.08780 (17.560/200)	0.10287 (20.574/200)

In the CLEF 2004 Spanish monolingual QA evaluation task, Out of 200 questions, our system provided the correct answer to 48 questions in run1 and 52 in run2. Hence, the global accuracy of our system was 24% and 26% for run1 and run2 respectively.

The accuracy over factoid questions is 18.89% (run1) and 21.11% (run2). In comparison with the other participants of the CLEF 2004 Spanish QA track (see (Magnini et al., 2004)), our system has obtained the best results in the following type of questions: location, person and objects. On the other hand, our system has a poor performance in the classes: manner, measure, organization, other and time.

- **Question Processing**. The Question Classification subsystem has an accuracy of 58%, a similar accuracy as the *environment*, MC and OC constraints. These values are influenced by the previous errors in the POS, NER and NEC subsystems.

- **Passage Retrieval**. In the PR we evaluated that 64.37% (run1) and 59.37% (run2) of questions have a correct answer in their passages. Taking into account the document identifiers the evaluation shows that 48.12% (run1) and 43.12% (run2) of the questions are really supported.

- **Answer Extraction**. The accuracy of the AE module for factoid questions for which the answer occurred in our selected passages was of 23.32% (run1) and 28.42% (run2).

5.4 Evaluation and Results at CLEF 2005

This section evaluates the behaviour of our system in CLEF 2005. From the 200 questions proposed in these evaluation, 118 were factoid, 50 definitional, and 32 temporally restricted factoid. We evaluated the three main components of our factoid QA system and the global results:

- **Question Processing**. This subsystem has been manually evaluated for factoid questions (see Table 5.5) in the following components: POS-tagging, NER and NE Classification (NEC) and QC. These results are accumulatives.

Table 5.5: Results of Question Processing evaluation.

Question Type	Subsystem	Total units	Correct	Incorrect	Accuracy	Error
FACTOID	POS-tagging	1122	1118	4	99.64%	0.36%
	NE Recognition	132	129	3	97.73%	2.27%
	NE Classification	132	87	45	65.91%	34.09%
	Q. Classification	118	78	40	66.10%	33.89%
TEMPORAL	POS-tagging	403	402	1	99.75%	0.25%
	NE Recognition	64	56	8	87.50%	12.50%
	NE Classification	64	53	11	82.81%	17.19%
	Q. Classification	32	27	5	84.37%	15.62%

- **Passage Retrieval**. This subsystem was evaluated using the set of correct answers given by the CLEF organization (see Table 5.6). We computed two measures: the first one (called *answer*) is the accuracy taking into account the questions that have a correct answer in its set of passages. The second one (called *answer+docID*) is the accuracy taking into account the questions that have a minimum of one passage with a correct answer and a correct document identifier in its set of passages. For factoid questions the two runs submitted differ in the parameters of the passage retrieval module: i) the maximum number of documents retrieved was 1200 (run1) and 1000 (run2), ii) the windows proximity was: (run1: 60 to 240 lemmas; run2: 80 to 220 lemmas), iii) the threshold for minimum passages: 4 (run1) and 1 (run2), iv) the maximum number of passages retrieved: 300 (run1) and 50 (run2).

Table 5.6: Passage Retrieval results (accuracy).

Question type	Measure	run1	run2
FACTOID	Acc. (*answer*)	78.09% (82/105)	76.19% (80/105)
	Acc. (*answer+docID*)	64.76% (68/105)	59.05% (62/105)
TEMPORAL	Acc. (*answer*)	50.00% (13/26)	46.15% (12/26)
	Acc. (*answer+docID*)	34.61% (9/26)	30.77% (8/26)

- **Answer Extraction**. The evaluation of this subsystem (see Table 5.7) uses the *answer+docID* and *answer* accuracies described previously.

Table 5.7: Factoid Answer Extraction results (accuracy).

Question Type	Accuracy Type	run1	run2
FACTOID	Acc. (*answer*)	29.27% (24/82)	26.25% (21/80)
	Acc. (*answer+docID*)	35.29% (24/68)	33.87% (21/62)
TEMPORAL	Acc. (*answer*)	15.38% (2/13)	33.33% (4/12)
	Acc. (*answer+docID*)	22.22% (2/9)	50.00% (4/8)

- **Global Results**. The overall results of our participation in CLEF 2005 Spanish monolingual QA task are listed in Table 5.8.

Table 5.8: Results of TALP-QA system at CLEF 2005 Spanish monolingual QA task.

Measure	run1	run2
Total Num. Answers	200	200
Right	58	54
Wrong	122	133
IneXact	20	13
Unsupported	0	0
Overall accuracy	29.00% (58/200)	27.00% (54/200)
Accuracy over Factoid	27.97% (33/118)	25.42% (30/118)
Accuracy over Definition	36.00% (18/50)	32.00% (16/50)
Accuracy over Temporal Factoid	21.88% (7/32)	25.00% (8/32)
Answer-string ”NIL” returned correctly	25.92% (14/54)	22.41% (13/58)
Confidence-weighted Score	0.08935 (17.869/200)	0.07889 (15.777/200)

In the CLEF 2005 Spanish monolingual QA evaluation task, out of 200 questions, our system provided the correct answer to 58 questions in run1 and 54 in run2. Hence, the global accuracy of our system was 29% and 27% for run1 and run2 respectively. In comparison with the results of the last evaluation (CLEF 2004), our system has reached a small improvement (24% and 26% of accuracy). Otherwise, we had 20 answers considered as inexact. We think that with a more accurate extraction phase we could extract correctly more questions and reach easily an accuracy of 39% . We conclude with a summary of the system behaviour for the three question classes:

- **Factoid questions**. The accuracy over factoid questions is 27.97% (run1) and 25.42% (run2). Although no direct comparison can be done using another test collection, we think that we have improved slightly our factoid QA system with respect to the results of the CLEF 2004 QA evaluation (18.89% and 21.11%) in Spanish. In comparison with the other participants

of the CLEF 2005 Spanish QA track, our system has obtained good results in the following type of questions: location and time. On the other hand, our system has obtained a poor performance in the classes: measure and other.

  - **Question Processing**. In this subsystem the Question Classification component has an accuracy of 66.10%. This result means that there is no great improvement with respect to the classifier used in CLEF 2004 (it reached a 58% of accuracy). These values are influenced by the previous errors in the POS, NER and NEC subsystems. On the other hand, NEC errors have increased substantially with respect to the previous evaluation. NEC component achieved an error rate of 34.09%. This is the most serious drawback of the QP phase and needs an in depth analysis for the next evaluation.
  - **Passage Retrieval**. We evaluated that 78.09% (run1) and 76.19% (run2) of questions have a correct answer in their passages. Taking into account the document identifiers the evaluation shows that 64.76% (run1) and 59.05% (run2) of the questions are really supported. This subsystem has improved substantially its results in comparison with the CLEF 2004 evaluation (48.12% and 43.12% of *answer+docID* accuracy).
  - **Answer Extraction**. The accuracy of the AE module for factoid questions for which the answer and document identifier occurred in our selected passages was of 35.29% (run1) and 33.87% (run2). This means that we have improved our AE module, since the results for this part in CLEF 2004 were 23.32% (run1) and 28.42% (run2), evaluated only with answer accuracy. This is the subsystem that performs worst and needs a substantial improvement and tuning.

- **Temporal Factoid Questions**. The accuracy over temporal factoid questions is 21.88% (run1) and 25.00% (run2). We detected poor results in the PR subsystem: the accuracy of PR with answer and document identifiers is 34.61% (run1) and 30.77% (run2). These results are due to the fact that some questions are temporally restricted by events. These questions need a special treatment, different from the one for factoid questions.

5.5 Evaluation and Results at TREC 2004

This section evaluates the behaviour of our system in TREC 2004. The QA track Main task at TREC 2004 consisted in resolving a series of questions related to a target. These questions are factoid, list, and 'other' questions. 'Other' questions ask for relevant information about a target that was not reported by the previous factoid and list questions. This Section focuses on the results obtained over the set of 230 factoid questions. We evaluated the three main components of our system and the global results:

- **Question Processing**. This subsystem has been manually evaluated for factoid questions (see Table 5.9) and the following components: target analysis (NERC) and substitution in the original question, basic NLP tools (POS, NER and NEC), semantic pre-processing (Environment, MC and OC construction) and finally, question classification.

In the following components the errors are cumulative: basic NLP tools (NE Recognition is influenced by POS-tagging errors and NE Classification is influenced by NE Recognition and POS-tagging errors), semantic pre-processing (the construction of the environment depends on the errors in the basic NLP tools and the syntactic analysis, the MC and OC errors are influenced by the errors in the environment), and question classification (is influenced by the errors in the basic NLP tools and the syntactic analysis).

Subsystem	Accuracy
Target Substitution	91.52% (151/165)
Target Analysis	72.31% (47/65)
POS-tagging	97.89% (1621/1656)
NE Recognition	88.89% (184/207)
NE Classification	82.13% (170/207)
Environment	45.22% (104/230)
MC	41.74% (96/230)
OC	82.61% (190/230)
Q. Classification	74.34% (171/230)

Table 5.9: Results of Question Processing evaluation.

- **Passage Retrieval**. The evaluation of this subsystem was performed using the set of correct answers given by the TREC organization (see Table 5.5). We submitted two runs. In both runs we retrieved only the 50 top passages for factoid questions. These passages were selected from the 1000 top documents.

Accuracy Measure	Result
Factoid (*answer*)	72.41% (147/203)
Factoid (*answer+docID*)	58.62% (119/203)

Table 5.10: Results of Passage Retrieval for Factoid questions.

- **Answer Extraction**. The evaluation of this subsystem for factoid questions has been done in three parts: evaluation of the Candidates Extraction (CE) module, evaluation of the Answer Selection (AS) module and finally we performed an evaluation of the AE subsystem's global accuracy for factoid questions in which the answer appears in our selected passages. The results are presented in Table 5.11.
- **Global Results**. The overall results of our participation in TREC 2004 are listed in Table 5.5.

Our system obtained a final score of 0.128 in run1 and 0.136 in run2. We conclude with a summary of the system behaviour for the the Factoid questions. The accuracy over factoid questions is 15.7%.

Subsystem	Accuracy (*answer*)
Candidates Extraction	25.17% (37/147)
Answer Selection	83.78% (31/37)
Answer Extraction	21.08% (31/147)

Table 5.11: Factoid Answer Extraction results.

Measure	Results
Factoid Total	230
Factoid Right	36
Factoid Wrong	190
Factoid IneXact	4
Factoid Unsupported	0
Factoid Precision NIL	0.089 (5/56)
Factoid Recall NIL	0.227 (5/22)
Accuracy over Factoid	0.157
Average F-score List	0.031
Average F-score Other (Run1)	0.165
Average F-score Other (Run2)	0.197
Final score (run1)	0.128
Final score (run2)	0.136

Table 5.12: Results of TALP-QA system at TREC 2004.

- **Question Processing**. The Question Classification subsystem has an accuracy of 74.34%. We improved slightly the results of this component with respect to the previous TREC evaluation. In the previous evaluation we obtained an accuracy of 69%.

- **Passage Retrieval**. In the PR we evaluated that 72.41% of questions have a correct answer in their passages. The evaluation taking into account the document identifiers shows that 58.62% of the questions are definitively supported. The accuracy of our PR subsystem has improved because in the TREC 2003 evaluation we obtained an accuracies of 62.10% and 42.36% for the previous measures respectively.

- **Answer Extraction**. The accuracy of the AE module for factoid questions for which the answer occurred in our selected passages is 21.08%. We achieved a significant improvement of our AE module, since the results of this component in TREC 2003 were 8.9%. We expect to improve these results by reducing the error rate in the construction of the *environment*, MC and OC.

5.6 Evaluation and Results at TREC 2005

This section presents the evaluation of the TALP-QA system for factoid questions and the global results at TREC 2005.

In TREC 2005 we combined the results of three heterogeneous factoid QA Systems: TALP-QA (a precision-oriented QA system), Sibyl (a recall-oriented QA system) and ARANEA (a recall-oriented and Web-based QA system). Three runs (sets of results) were submitted to the evaluation. The first one using the TALP-QA system, the second one combining TALP-QA and Sybil, and finally the third one combining three systems: TALP-QA, Sybil, and ARANEA.

The QA track Main task at TREC 2005 had the same form of the previous edition (see previous Section). In the TREC 2005 edition were 362 factoid questions. On the other hand a Document Ranking task was attached to the Main task. The results of these task are also reported.

- **Question Processing**. This subsystem has been manually evaluated for factoid questions (see Table 5.13) and the following components: target substitution in the original question, basic NLP tools (POS, NER and NEC), semantic pre-processing (Environment, MC and OC construction) and finally, Question Classification (QC).

  In the following components the errors are cumulative: basic NLP tools (NER is influenced by POS-tagging errors and NEC is influenced by NER and POS-tagging errors), semantic pre-processing (the construction of the environment depends on the errors in the basic NLP tools and the syntactic analysis, the MC and OC errors are influenced by the errors in the environment), and QC (is influenced by the errors in the basic NLP tools and the syntactic analysis).

Subsystem	Accuracy
Target Substitution	89.83% (309/344)
POS-tagging	98.87% (3149/3185)
NE Recognition	93.53% (434/464)
NE Classification	82.11% (381/464)
Environment	49.45% (179/362)
MC	31.77% (115/362)
OC	58.01% (210/362)
Q. Classification	76.79% (278/362)

Table 5.13: Results of Question Processing evaluation for the TALP-QA system.

- **Passage Retrieval**. The evaluation of this subsystem was performed using the set of correct answers given by the TREC organization (see Table 5.14).

- **Answer Extraction**. We evaluated the Candidates Extraction (CE) module, the Answer Selection (AS) module and finally we performed an evaluation of the AE subsystem's global accuracy for factoid questions in which the answer appears in our selected passages.

Question	Accuracy	Result
Factoid (run1)	(*answer*)	62.60% (216/345)
	(*answer+docID*)	46.37% (160/345)

Table 5.14: TALP-QA Passage Retrieval results.

Subsystem	Accuracy (*answer*)
Candidates Extraction	8.11% (28/345)
Answer Selection	71.42% (20/28)
Answer Extraction	5.79% (20/345)

Table 5.15: TALP-QA Answer Extraction results.

- **Global Results**. The overall results of our participation in the TREC 2005 Main QA Task are listed in Table 5.16. The results of Document Ranking Evaluation Task are listed in Table 5.17.

Measure	run1	run2	run3
Factoid Total	362	362	362
Factoid Right	27	53	62
Factoid Wrong	330	288	279
Factoid IneXact/Uns.	4/1	17/4	17/4
Factoid Precision NIL	7/172	5/76	5/77
Factoid Recall NIL	7/17	5/17	5/17
Accuracy over Factoid	0.075	0.146	0.171
Average F-score List	0.024	0.026	0.028
Average F-score Other	0.172	0.164	0.079
Final score	0.088	0.125	0.116

Table 5.16: Results of TALP's runs at TREC 2005.

This section summarizes the evaluation of our participation in the TREC 2005 Main QA and Document Ranking tasks.

- **Question Answering Task**. Our system obtained a final score of 0.088 in *run1*, 0.125 in *run2*, and 0.116 in *run3* (see Table 5.16). The accuracy over factoid questions is 7% in *run1*, 14.6% in *run2*, and 17.1% in *run3* (see 5.16). The results of the TALP-QA system (run1) are low due to errors in the Candidates Extraction module. Otherwise, the voting scheme is useful as seen in runs 2 and 3.

  The TALP-QA system (run1) has been evaluated in its three phases:

Run	run1	run2
AvgP.	0.1191	0.1468
R-Prec.	0.1287	0.1685
Docs. Retrieved	781	1619
Recall (%)	11.68%	20%
Recall	184/1575	375/1575
Δ AvgP. Diff.(%) over all runs AvgP.	-32.15%	-7.22%

Table 5.17: TREC 2005 Document Ranking Task.

1. **Question Processing**. The Question Classification subsystem has an accuracy of 76.79%. We improved slightly the results of this component with respect to the TREC 2004. In the previous evaluation we obtained an accuracy of 74.34%. These are good results if we take into account that in TREC 2005 has increased the average length of both questions and targets.
2. **Passage Retrieval**. We evaluated that 62.60% of questions have a correct answer in their passages. The evaluation taking into account the document identifiers shows that 46.37% of the questions are definitively supported. The accuracy of our PR subsystem has decreased in comparison with the TREC 2004 evaluation (72.41% and 58.62% of accuracy for the previous measures respectively). This drop may be due to the increase of the average question length at TREC 2005.
3. **Answer Extraction**. The accuracy of the AE module for factoid questions for which the answer occurred in our selected passages is 5.79%. This poor accuracy is due to a technical error in the AE module. Otherwise, we expect to improve these results by reducing the error rate in the construction of the *environment*, MC and OC.

- **Document Ranking Task**. The results of the Document Ranking task are presented in Table 5.17. Our system obtained an Average Precision of 0.1191 (run1) and 0.1468 (run2 and run3), a R-Precision of 0.1287 (*run1*) and 0.1685 (*run2* and *run3*). The Document Raking Median of over all runs of TREC 2005 was 0.1574. We obtained an Average Precision Difference over all runs of -32.15% (*run1*) and -7.22% (*run2* and *run3*).

The resulting voting scheme with system combination has been successful, improving the accuracy over *run1* (with only TALP-QA) with 108% in *run2* and with 144% in *run3*. The results in factoid questions were 7% of accuracy in the run without voting, and 14.6% and 17.1% in the runs with voting. While these numbers are low (due to technical problems in the Answer Extraction phase of TALP-QA system) they indicate that voting is a successful approach for performance boosting of QA systems.

Chapter 6

GeoTALP-QA Geographical Question Answering Approach

This chapter describes an approach to adapt an existing multilingual Open-Domain Question Answering (ODQA) system for factoid questions to a Restricted Domain, the Geographical Domain. The adaptation of this ODQA system involved the modification of some components of our system such as: Question Processing, Passage Retrieval and Answer Extraction. The new system uses external resources like GNS Gazetteer for Named Entity (NE) Classification and Wikipedia or Google in order to obtain relevant documents for this domain. The system focuses on a Geographical Scope: given a region, or country, and a language we can semi-automatically obtain multilingual geographical resources (e.g. gazetteers, trigger words, groups of place names, etc.) of this scope. The resulting multilingual Geographical Domain Question Answering (GDQA) system is called GeoTALP-QA. This Restricted Domain Question Answering (RDQA) system has been built over an existing ODQA system, TALP-QA. The system has been trained and evaluated for Spanish in the scope of the Spanish Geography.

We outline below the organization of the chapter. In the next section we present the overall architecture of GeoTALP-QA and describe briefly its main components, focusing on those components that have been adapted from an ODQA to a GDQA. Then, the Scope-Based Resources needed for the experimentation and the experiments are presented in Sections 2 and 3. In section 4 we present the results obtained over a Geographical Domain corpus.

6.1 System Description

GeoTALP-QA has been developed within the framework of ALIADO[1] project. The system architecture uses a common schema with three phases that are performed sequentially without feedback: Question Processing (QP), Passage Retrieval (PR) and Answer Extraction (AE). More details about this architecture can be found in (Ferrés et al., 2005) and (Ferrés et al., 2004).

Before describing these subsystems, we introduce some additional knowledge sources that have been added to our system for dealing with the geographic domain and some language-dependent

[1] **ALIADO**. http://gps-tsc.upc.es/veu/aliado

NLP tools for English and Spanish. Our aim is to develop a language independent system (at least able to work with English and Spanish). Language dependent components are only included in the Question Pre-processing and Passage Pre-processing components, and can be easily substituted by components for other languages.

6.1.1 Additional Knowledge Sources

One of the most important task to deal with the problem of GDQA is to detect and classify NEs with its correct Geographical Subclass (see classes in Section 6.3). We use Geographical scope based Knowledge Bases (KB) to solve this problem. These KBs can be built using these resources:

- **GEOnet Names Server (GNS[2])**. A worldwide gazetteer, excluding the USA and Antarctica, with 5.3 million entries.
- **Geographic Names Information System** (GNIS[3]). A gazetteer with 2.0 million entries about geographic features of the USA.
- **Grammars for creating NE aliases**. Geographic NEs tend to occur in a great variety of forms. It is important to take this into account to avoid losing occurrences. A set of patterns for expanding have been created. (e.g. <toponym>_Mountains, <toponym>_Range, <toponym>_Chain).
- **Trigger Words Lexicon**. A lexicon containing trigger words (including multi-word terms) is used for allowing local disambiguation of ambiguous NE, both in the questions and in the retrieved passages.

Working with geographical scopes avoids many ambiguity problems, but even in a scope these problems occur:

- **Referent ambiguity problem.** This problem occurs when the same name is used for several locations (of the same or different class). In a question, sometimes it is impossible to solve this ambiguity, and, in this case, we have to accept as correct all of the possible interpretations (or a superclass of them). Otherwise, a trigger phrase pattern can be used to resolve the ambiguity (e.g. ”Madrid” is an ambiguous NE, but in the phrase, ”comunidad de Madrid” (State of Madrid), ambiguity is solved). Given a scope, we automatically obtain the most common trigger phrase patterns of the scope from the GNS gazetteer.
- **Reference ambiguity problem.** This problem occurs when the same location can have more than one name (in Spanish texts this frequently occurs as many place names occur in languages other than Spanish, as Basque, Catalan or Galician). Our approach to solve this problem is to group together all the geographical names that refer to the same location. All the occurrences of the geographical NEs in both questions and passages are substituted by the identifier of the group they belong to.

[2] **GNS**. http://earth-info.nga.mil/gns/html
[3] **GNIS**. http://geonames.usgs.gov/geonames/stategaz

We used the geographical knowledge available in the GNS gazetteer to obtain this geographical NEs groups. First, for each place name in the scope-based GNS gazetteer we obtained all the NEs that have the same feature designation code, latitude and longitude. For each group, we then selected an identifier choosing one of the NE included in it using the following heuristics: the information of the GNS field ”native” tells if a place name is native, conventional, a variant, or, is not verified. So we decided the group representative assigning the following order of priorities to the names: native, conventional name, variant name, unverified name. If there is more than one place name in the group with the same name type we decide that the additional length gives more priority to be cluster representative. It is necessary to establish a set of priorities among the different place names of the group because in some retrieval engines (e.g. web search engines) is not possible to do long queries.

6.1.2 Language-Dependent Processing Tools

A set of general purpose NLP tools are used for Spanish and English. The same tools are used for the linguistic processing of both the questions and the passages (see (Ferrés et al., 2005) and (Ferrés et al., 2004) for a more detailed description of these tools). The tools used for Spanish are:

- *FreeLing*, which performs tokenization, morphological analysis, POS tagging, lemmatization, and partial parsing.
- *ABIONET*, a NE Recognizer and Classifier (NERC) on basic categories.
- *EuroWordNet*, used to obtain a list of synsets, a list of hypernyms of each synset, and the Top Concept Ontology class.

The following tools are used to process English:

- *TnT*, a statistical POS tagger.
- *WordNet lemmatizer 2.0*.
- *ABIONET*.
- *WordNet 1.5*.
- *Spear*. A modified version of the Collins parser.
- *Alembic*, a NERC with MUC classes.

6.1.3 Question Processing

The main goal of this subsystem is to detect the Question Type (QT), the Expected Answer Type (EAT), and the question analysis. This information is needed for the other subsystems. We use a language-independent formalism to represent this information. We apply the processes described above to the the question and passages to obtain the following information:

- Lexical and semantic information for each word: form, lemma, POS tag (Eagles or PTB tag-set), semantic class and subclass of NE, and a list of EWN synsets.
- Syntactic information: syntactic constituent structure of the sentence and the information of dependencies and other relations between these components.

Once this information is obtained we can find the information relevant to the following tasks:

- **Environment Building.** The *Environment* of a question is the set of semantic relations that hold between the different components identified in the question text (the *Environment* is described in Section 5.1.2. The ontology has been extended for the GD (see below the classes related with this domain).

```
 ENTITY
     ENTITY_PROPER_PLACE
         GEOLOGICAL_REGION
             ARCHIPELAGO
             ISLAND
             LAND_FORM
                 MOUNTAIN
             SEA_FORM
                 CAPE
                 GULF
                 SEA
             WATER_FORM
                 RIVER
         POLITICAL_REGION
             CITY
             CONTINENT
             COUNTY
             COUNTRY
             STATE
ENTITY_QUANTITY
          NUMERIC
MAGNITUDE
          AREA
          LENGTH
          FLOW
          WEIGHT
```

- **Question Classification**. Our ODQA system uses 25 QTs. For the GD we only used 10 Question Types (see Table 6.1). Only 5 QTs are common with the ODQA QTs, 5 QTs have been specially created for this domain.

  In order to determine the QT our system uses a Prolog DCG Parser. This parser uses the following features: word form, word position in the question, lemma and part-of-speech (POS). A set of DCG rules was manually configured in order to ensure a sufficient coverage.

  The parser uses external information: geographical NE subclasses, trigger words for each Geographical subclass (e.g. ”poblado” (*ville*)), semantically related words of each subclass (e.g. ”water” related with *sea* and *river*), and introductory phrases for each Question Type (e.g. ”which extension” is a phrase of the QT *What_area*).

Question Type	Expected Answer Type
Count_objects	NUMBER
How_many_people	NUMBER
What_area	MEASURE_AREA
What_flow	MEASURE_FLOW
What_height	MEASURE_HEIGHT
What_length	MEASURE_LENGTH
Where_action	LOCATION_SUBCLASS
Where_location	LOCATION_SUBCLASS
Where_quality	LOCATION_SUBCLASS
Default_class	LOCATION

Table 6.1: QTs and Expected Answer Types.

- **Semantic Constraints Extraction.** The Semantic Constrains Set is the set of Mandatory and Optional constraints extracted from the question (see Section 5.1.2 for an extended explanation of these concepts). An example of the constraints extracted from an environment is shown in Table 6.2. This example shows the question type predicted, the initial predicates extracted from the question, the Environment predicates, the MCs and the OCs. MCs are *entity(4)* and *i_en_city(6)*. The first predicate refers to token number 4 (”autonomia” (*state*)) and the last predicate refers to token number 6 (”Barcelona”).

Question	¿ A qué autonomía pertenece Barcelona? (Which state Barcelona pertains to?)
Q. Type	*where_location*
Predicates	*city(’Barcelona’),state(X), pertains(’Barcelona’,X)*
Environment	*action(5), participant_in_event(5,4), theme_of_event(5,6),prep(4,2),entity(4), i_en_proper_place(6),det(4,3),qu(3)*
Mandatory Constraints	*entity(4),i_en_city(6)*
Optional Constraints	*action(5),theme_of_event(5,6), participant_in_event(5,4),prep(4,2), type_of_location(5,5,i_en_state), property(5,5,pertenecer,3,6)*

Table 6.2: Question Analysis example.

6.1.4 Passage Retrieval

We use two different approaches for Passage Retrieval. The first one uses a pre-processed corpus as a document collection. The second one uses the web as document collection.

Off-line Corpus Retrieval

This approach uses a pre-processed and indexed corpus with Scope-related Geographical Information as a document collection for Passage Retrieval. The processed information was used for indexing the documents. Storing this information allows us to avoid the pre-processing step after retrieval. The Passage Retrieval algorithm used is the same of our ODQA system: a data-driven query relaxation technique with dynamic passages implemented using Lucene IR engine API (See (Ferrés et al., 2005) for more details).

Online Web Snippet Retrieval

The other approach uses a search-engine to get snippets with relevant information. We expect to get a high recall with few snippets. In our experiments, we chose Google as the search-engine using a boolean retrieval schema that takes advantage of its phrase search option and the Geographical KB to create queries that can retrieve highly relevant snippets. We try to maximize the number of relevant sentences with only one query per question.

The algorithm used to build the queries is simple. First, some expansion methods described below can be applied over the keywords. Then, stop-words (including normal stop-words and some trigger words) are removed. Finally, only the Nouns and Verbs are extracted from the keywords list. The expansion methods used are:

- **Trigger Words Joining (TWJ).** Uses the trigger words list and the trigger phrase pattern list (automatically generated from GNS) to join trigger phrases (e.g. ”isla Conejera” o ”Sierra de los Pirineos”).

- **Trigger Words Expansion (TWE).** This expansion is applied to the NEs that were not detected as a trigger phrase. The expansion uses its location subclass to create a keyword with the pattern: $TRIGGER + NE$ (e.g. ”Conejera” is expanded to: (”isla Conejera” OR ”Conejera”)).

- **GNS Grouping Expansion (CE).** Noun Phrase expansion based on the groups generated from GNS Gazetteer.

- **Question-based Expansion (QBE).** This method appends keywords or expands the query depending on the question type. As an example, in the case of a question classified as *What_length*, trigger words and units associated to the question class like ”*longitud*” (*length*) and ”*kilómetros*” (*kilometers*) are appended to the query.

6.1.5 Answer Extraction

We used two systems for Answer Extraction: our ODQA system (adapted for the GD) and a frequency based system.

ODQA Extraction

The Candidates Extraction phase is based on a relaxation process of the set of semantic constraints that is performed by means of structural or semantic relaxation rules, using the semantic ontology (consult Section 5.1.4 for more details about this process). Then an extraction process applies a set of extraction rules on the set of sentences that have satisfied the Mandatory Constraints. Finally an Answer selection phase scores the candidates and extracts the answer (see Section 5.1.4 for more details).

Frequency-Based Extraction

This extraction algorithm is quite simple. First, all snippets are pre-processed. Then, we make a ranked list of all the tokens satisfying the expected answer type of the question. The score of each token in the snippets is computed using the following formula:

$$Score(tk_i) = \sum_{o \in Occurrence(tk_i)} \frac{1}{snippet\_rank(o)}$$

Finally, the top-ranked token is extracted.

6.2 Resources for Scope-Based Experiments

In this section we describe how we obtained the resources needed to carry out experiments in the Spanish Geography domain using Spanish language. These resources were: the question corpus (validation and test), the document collection required by the off-line ODQA Passage Retrieval, and the geographical scope-based resources. Finally, we describe the experiments performed.

6.2.1 Language and Scope Based Geographical Question Corpus

We obtained a corpus of Geographical questions from Albayzin, a speech corpus (Diaz et al., 1998) that contains a geographical subcorpus with utterances of questions about the geography of Spain in Spanish. We obtained from Albayzin a set of 6887 question patterns. We analyzed this corpus and we extracted the following type of questions: Partial Direct, Partial Indirect, and Imperative Interrogative factoid questions with a simple level of difficulty (e.g. questions without nested questions). We selected a set of 2287 question patterns. As a question corpus we randomly selected a set of 177 question patterns from the previous selection (see Table 6.3). These patterns have been randomly instantiated with Geographical NEs of the Albayzin corpus. Then, we searched the answers in the Web and the Spanish Wikipedia (SW). The results of this process were: 123 questions with answer in the SW and the Web, 33 questions without answer in the SW but with answer using the Web, and finally, 21 questions without answer (due to the fact that some questions when instantiated cannot be answered (e.g. which sea bathes the coast of Madrid?)). We divided the 123 questions with answer in the SW in two sets: 61 questions for development (setting thresholds and other parameters) and 62 for test.

¿A qué comunidad autónoma pertenece el <PICO>?
At which state pertains <PEAK>?
¿Cuál es el capital de <COMUNIDAD>?
Which is the capital of <STATE>?
¿Cuál es la comunidad en la que desemboca el <RÍO>?
What is the state in which <RIVER> flows into?
¿Cuál es la extensión de <COMUNIDAD>?
Which is the extension of <STATE>?
Longitud del río <RÍO>.
Length of river <RIVER>.
¿Cuántos habitantes tiene la <COMUNIDAD>?
How many people does <STATE> has?

Table 6.3: Some question patterns from Albayzin.

6.2.2 Document Collection for ODQA Passage Retrieval

In order to test our ODQA Passage Retrieval system we need a document collection with enough geographical information to solve the questions of Albayzin corpus. We used the filtered Spanish Wikipedia[4]. First, we obtained the original set of documents (26235 files). Then, we selected two sets of 120 documents about the Spanish geography domain and the non-Spanish geography domain. Using these sets we obtained a set of Topic Signatures (TS) (Lin & Hovy, 2000) for the Spanish geography domain and another set of TS for the non-Spanish geography domain. Then, we used these TS to filter the documents from Wikipedia, and we obtained a set of 8851 documents belonging to the Spanish geography domain. These documents were pre-processed and indexed.

6.2.3 Geographical Scope-Based Resources

A Knowledge Base (KB) of Spanish Geography has been built using four resources:

- GNS: We obtained a set of 32,222 non-ambiguous place names of Spain.
- Albayzin Gazetteer: a set of 758 places.
- A Grammar for creating NE aliases. We created patterns for the summit and state classes (the ones with more variety of forms), and we expanded this patterns using the entries of Albayzin.
- A lexicon of 462 trigger words.

We obtained a set of 7632 groups of place names using the grouping process over GNS. These groups contain a total of 17617 place names, with an average of 2.51 place names per group. See in Figure 6.1 an example of a group where the canonical term appears underlined.

In addition, a set of the most common trigger phrases in the domain has been obtained from the GNS gazetteer (see Table 6.4).

[4]**Spanish Wikipedia**. http://es.wikipedia.org

{*Cordillera Pirenaica*, *Pireneus*, *Pirineos*, *Pyrenaei Montes*, *Pyrénées*, *Pyrene*, *Pyrenees*}

Figure 6.1: Example of a group obtained from GNS.

	Geographical Scope	
	Spain	UK
Top-ranked Trigger Phrases	*TRIGGER* de *NE*	*NE TRIGGER*
	TRIGGER NE	*TRIGGER NE*
	TRIGGER del *NE*	*TRIGGER* of *NE*
	TRIGGER de la *NE*	*TRIGGER* a' *NE*
	TRIGGER de las *NE*	*TRIGGER* na *NE*

Table 6.4: Sample of the top-ranked trigger phrases automatically obtained from GNS gazetteer for the geography of Spain and UK.

6.3 Experiments

We have designed some experiments in order to evaluate the accuracy of the GDQA system and its subsystems (QP, PR, and AE). For PR, we evaluated the web-based snippet retrieval using Google with some variants of expansions, versus our ODQA Passage Retrieval with the corpus of the SW. Then, the passages (or snippets) retrieved by the best PR approach were used by the two different Answer Extraction algorithms. The ODQA Answer Extractor has been evaluated taking into account the answers that have a supported context in the set of passages (or snippets). Finally, we evaluated the global results of the complete QA process with the different Answer Extractors: ODQA and Frequency-Based.

6.4 Results

This section evaluates the behavior of our GDQA system over a test corpus of 62 questions and reports the errors detected on the best run. We evaluated the three main components of our system and the global results.

- **Question Processing**. The Question Classification task has been manually evaluated. This subsystem has an accuracy of 96.77%.
- **Passage Retrieval**. The evaluation of this subsystem was performed using a set of correct answers (see Table 6.5). We computed the *answer accuracy*: it takes into account the number of questions that have a correct answer in its set of passages.
- **Answer Extraction**. The evaluation of the ODQA Answer Extractor subsystem is shown in Table 6.6. We evaluated the accuracy taking into account the number of correct and supported answers by the passages divided by the total number of questions that have a supported answer

Retrieval Mode	Accuracy at N passages/snippets			
	N=10	N=20	N=50	N=100
Google	0.6612	0.6935	0.7903	0.8225
+TWJ	0.6612	0.6774	0.7419	0.7580
+TWJ+TWE	0.6612	0.6774	0.7419	0.7580
+CE	0.6612	0.6774	0.7741	0.8064
+QBE	0.8064	0.8387	0.9032	0.9354
+TWJ+QB+CE	0.7903	0.8064	0.8548	0.8870
Google+All	0.7903	0.8064	0.8548	0.8870
ODQA+Wiki	0.4354	0.4516	0.4677	0.5000

Table 6.5: Passage Retrieval results (refer to section 3.4.2 for detailed information of the different query expansion acronyms).

in its set of passages. This evaluation has been done using the results of the top-ranked retrieval configuration over the development set: the *Google+TWJ+QB+CE* configuration of the snippet retriever.

Accuracy at N Snippets		
N=10	N=20	N=50
0.2439 (10/41)	0.3255 (14/43)	0.3333 (16/48)

Table 6.6: Results of the ODQA Answer Extraction subsystem (accuracy).

In Table 6.7 are shown the global results of the two QA Answer Extractors used (ODQA and Frequency-Based). The passages retrieved by the *Google+TWJ+QB+CE* configuration of the snippet retriever were used.

	Accuracy	
Num. Snippets	ODQA	Freq-based
10	0.1774 (11/62)	0.5645 (35/62)
20	0.2580 (16/62)	0.5967 (37/62)
50	0.3387 (21/62)	0.6290 (39/62)

Table 6.7: QA results over the test set.

We analyzed the 23 questions that fail in our best run. The analysis detected that 10 questions had no answer in its set of passages. In 5 of these questions it is due to have a non common question or location. The other 5 questions have problems with ambiguous trigger words (e.g. *capital*) that confuse the web-search engine. On the other hand, 13 questions had the answer in its set of passages, but were incorrectly answered. The reasons are mainly due to the lack of passages with the answer (8), answer validation and spatial-reasoning (3), multilabel Geographical NERC (1), and more context in the snippets (1).

Out of 62 questions, our system provided the correct answer to 39 questions in the experiment with the best results.

Our Passage Retrieval for ODQA offers less attractive results when using the SW corpus. The problem of using SW to extract the answers is that it gives few documents with the correct answer, and, it is difficult to extract the answer because the documents contain tables, lists, ill-formed sentences, etc. Our ODQA AE needs a grammatically well-structured text to extract correctly the answers. The QA system offers a low performance (33% of accuracy) when using this AE over the web-based retrieved passages. In some cases, the snippets are cut and we could expect a better performance retrieving the whole documents from Google.

On the other hand, web-based snippet retrieval, with only one query per question, gives good results in Passage Retrieval. The QA system with the Frequency-Based AE obtained better results than with the ODQA AE (62.9% of accuracy).

Chapter 7

TALP-GeoIR Geographical IR Approach

This section describes TALP-GeoIR, a multilingual Geographical Information Retrieval (GIR) system. The chapter focuses on the GIR systems that have been used for our participation in the GeoCLEF Monolingual English tasks at CLEF 2005 (Gey et al., 2005) and GeoCLEF 2006 (Gey et al., 2006b) (see the GeoCLEF evaluations Section in Chapter 3). The approaches for both evaluations were similar but with changes in the retrieval modes. The GIR system at GeoCLEF 2005 (called GeoTALP-IR) was based on *Lucene*, uses a modified version of the Passage Retrieval module of the TALP Question Answering (QA) system presented at CLEF 2004 (Ferrés et al., 2004) and TREC 2004 (Ferrés et al., 2005). We designed a Keyword Selection algorithm based on a Linguistic and Geographical Analysis of the topics. A Geographical Thesaurus (GT) has been build using a set of Geographical Gazetteers and a Geographical Ontology. Our GIR system at GeoCLEF 2006 (called TALP-GeoIR) was a modified version of the system presented in GeoCLEF 2005 (Ferrés et al., 2005a) with some changes in the retrieval modes and the Geographical Knowledge Base (KB). This system had four phases performed sequentially: i) a Keyword Selection algorithm based on a linguistic and geographical analysis of the topics, ii) a geographical document retrieval with *Lucene*, iii) a textual document retrieval with the *JIRS* Passage Retrieval (PR) software, and iv) a Document Ranking phase. A Geographical KB has been build using a set of publicly available geographical gazetteers and the *Alexandria Digital Library (ADL) Feature Type Thesaurus*.

In this chapter we present the overall architecture of GeoTALP-IR in the different GeoCLEF evaluation tasks and describe briefly its main components. We also present an evaluation of the system used in the GeoCLEF 2005 and GeoCLEF 2006 evaluations.

7.1 GeoCLEF 2005 System Description

7.1.1 Overview

The system architecture presented at GeoCLEF 2005 has two phases that are performed sequentially (as shown in Figure 7.1): Topic Analysis (TA) and Document Retrieval (DR). A collection pre-processing process was carried out in advance.

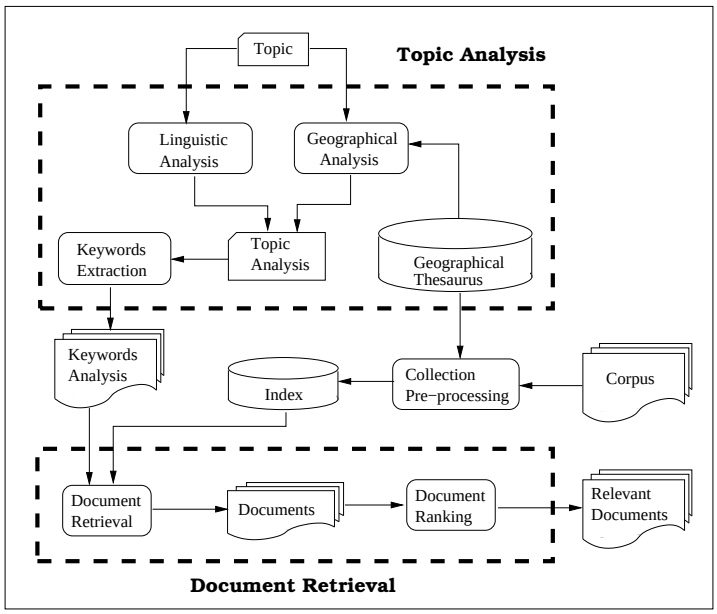

Figure 7.1: Architecture of GeoTALP-IR system.

7.1.2 Collection Pre-processing

We have used the *Lucene*[1] Information Retrieval (IR) engine to perform the DR task. Before GeoCLEF 2005 we indexed the entire English collections: Glasgow Herald 1995 (GH95) and Los Angeles Times 1994 (LAT94) (i.e. 169,477 documents). We pre-processed the whole collection with linguistic tools (described in the next sub-section) to mark the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process the collection is analyzed with a Geographical Thesaurus (described in the next sub-section). This information was used to build an index (see an example in Figure 7.2) that contains the following fields for each document:

- **Form Field**: this field stores the original text (word forms) with the Named Entities recognized.

[1] http://jakarta.apache.org/lucene

- **Lemma Field**: this part is built using the lemmas of the words, the POS tags, and the results of the Named Entity Recognition and Classification (NERC) module and the Geographical Thesaurus.

- **Geo Field**: it contains all NEs classified as *location* or *organization* that appear in the Geographical Thesaurus. This part has the geographical information about these NE: including geographical coordinates and geographical relations with the corresponding places of its path to the top of the geographical ontology (i.e. a city like "Barcelona" contains its state, country, sub-continent and continent). If a NE is an ambiguous location, all the possible ambiguous places are stored in this field.

Field	Indexed Content
Form	Watson flew off with his wife for a weekend in Barcelona, returned to London on Monday,
Lemma	Watson#NNP#PERSON fly#VBD off#RP with#IN his#PRP$ wife#NN for#IN a#DT weekend#NN in#IN Barcelona#NNP#LOCATION#city ,#, return#VBD to#TO London#NNP#LOCATION#capital on#IN monday#NNP ,#,
Geo	Europe#Europe#Spain#Cataluña#Barcelona#41.383_2.183 Europe#Europe#United_Kingdom#England#London#51.517_-0.105

Figure 7.2: Example of an indexed document.

7.1.3 Topic Analysis

The goal of this phase is to extract all the relevant keywords from the topics enriching them as a result of the analysis. These keywords are then used by the Document Retrieval phase. The Topic Analysis phase has three main components: a Linguistic Analysis, a Geographical Analysis and a Keyword Selection algorithm.

Linguistic Analysis

This process extracts lexico-semantic and syntactic information using the following set of Natural Language Processing tools:

- **Morphological components**, a statistical POS tagger (*TnT*) (Brants, 2000) and the WordNet 2.0 (Fellbaum, 1998) lemmatizer are used to obtain POS tags and lemmas. We used the *TnT* pre-defined model trained on the Wall Street Journal corpus.

- **Spear**, which performs full parsing and robust detection of verbal predicate arguments (Collins, 1999). See Section 5.2.2 for more details.

- **A Maximum Entropy based NERC**, a Named Entity Recognizer and Classifier that identifies and classifies NEs in basic categories (person, place, organization and other). This

NERC has been trained with the CONLL-2003 shared task English data set (Tjong Kim Sang & De Meulder, 2003b).

- **Gazetteers**, with the following information: location-nationality relations (e.g. Spain-Spanish) and actor-action relations (e.g. write-writer).

Geographical Analysis

The Geographical Analysis is applied to the Named Entities provided by the location tag (<EN-location>), and the Named Entities from the Title and Description tags that have been classified as *location* or *organization* by the NERC module. This analysis has two main components:

- **Geographical Thesaurus:** this component has been built joining three gazetteers that contain entries with places and their geographical class, coordinates, and other information: *GeoNet Names Server* (GNS), *Geographic Names Information System* (GNIS) (using only a subset of 39,906 of the most important geographical names), and *GeoWorldMap*.

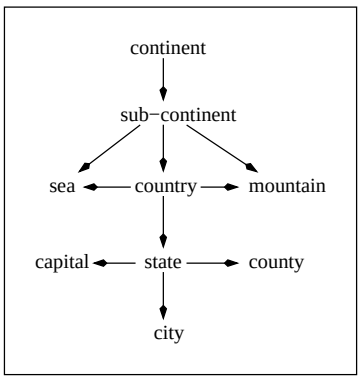

Figure 7.3: Geographical ontology.

Each one of these gazetteers have a different set of classes. We have mapped this sets to our set of classes (see Figure 7.3), which includes the most common classes and the most important ones (e.g. country is not common, but is important). The resulting thesaurus contains approximately 3.7 million places with its geographical class. This approach is similar to that used in (Manov et al., 2003), but they used a limited number of locations (only the 50,000 most important ones).

- **NEC correction filter**: a filter to correct some common errors in the *location-person* and *organization-person* ambiguity classes has been implemented. This filter stores all the NEs classified as *person* in the document; for each one of these NEs it extracts and stores in a hash table all the tokens that compose the NE. Then, for each NE of the document classified as *location* or *organization* it checks whether the NE exists in the document hash. If the NE exists then its class is changed to *person*.

Topic Keywords Selection

We designed an algorithm to extract the most relevant keywords of each topic (see an example in Figure 7.4). These keywords are then passed to the Document Retrieval phase. The algorithm is applied after the Linguistic and Geographical analysis and has the following steps:

1. Initial Filtering. First, all the punctuation symbols and stopwords are removed from the analysis of the title, description and geographical tags.

2. Title Words Extraction. All the words from the title tag are obtained.

3. Description Chunks Filtering. All the Noun Phrase base chunks from the description tag that contain a word with a lemma that appears in one or more words from the title are extracted.

4. Description Words Extraction. The words belonging to the chunks extracted in the previous step and do not have a lemma appearing in the words of the title are extracted.

5. Append Title, Description and Location Words Analysis. The words extracted from the title and description and the geographical tag are appended.

Topic	EN-title	Environmental concerns in and around the Scottish Trossachs
	EN-desc	Find articles about environmental issues and concerns in the Trossachs region of Scotland.
	EN-location	the Scottish Trossachs
Keyword Selection	Title Stopword Filtering	Environmental concerns Scottish Trossachs
	Title Extracted words	Environmental, concerns, Scottish, and Trossachs
	Description Chunks	[environmental issues] [Trossachs region]
	Description Words Extraction	issues and region
	Selected Keywords	Environmental#environmental#JJ concerns#concern#NNS issues#issue#NNS region#region#NN scottish#Scottish#NNP#misc#location(”Scotland”) Trossachs#trossachs#NNP

Figure 7.4: Keyword Selection example.

7.1.4 Document Retrieval

The main function of the Document Retrieval component is to extract relevant documents that are likely to contain the information needed by the user. Document retrieval is performed using the

Lucene Information Retrieval system. *Lucene* uses the standard tf.idf weighting scheme with the cosine similarity measure, and it allows ranked and boolean queries. The document retrieval algorithm uses a data-driven query relaxation technique: if too few documents are retrieved, the query is relaxed by discarding the keywords with the lowest priority. The reverse happens when too many documents are extracted. Each keyword is assigned a priority using a series of heuristics fairly similar to (Moldovan et al., 1999). For example, a proper noun is assigned a higher priority than a common noun, the adverb is assigned the lowest priority, and stop words are removed.

The main options of the Document Retrieval phase are:

- Query types:
  - Boolean: all the keywords must appear in the documents retrieved. *Lucene* allows boolean queries and returns a score for each retrieved document.
  - Ranked: *Lucene* does ranked queries with tf-idf and cosine similarity.
  - Boolean+Ranked: this mode joins documents retrieved from boolean and ranked queries, giving priority to the documents from the boolean query.
- Geographical Search Mode:
  - Lemma Field: this search mode implies that all the keywords that are Named Entities detected as *location* are searched in the "Lemma" field part of the index.
  - Geo Field: this search means that the NEs tagged as *location* and detected as keywords will be searched at the "Geo" index field.
- Geographical Search Policy:
  - Strict: this search policy can be enabled when the "Geo" Field search is running, and is used to find a *location* with exactly all this ontological path and coordinates for the following classes: country and region. In example, the form used to search "Australia" in the index is:
    Oceania#Oceania#Australia#-25.0_135.0
  - Relaxed: this search policy can also be enabled when the "Geo" field search is running. This mode searches without coordinates. The form used to search "Australia" in the index for this kind of search policy is:
    Oceania#Oceania#Australia
    In this case, the search is flexible and all the cities and regions of Australia will be returned. An example of a location found with the previous query is:
    Oceania#Oceania#Australia#Western_Australia#Perth#-31.966_115.8167

7.1.5 Document Ranking

This component joins the documents provided by the Document Retrieval phase. If the Query type is *boolean* or *ranked* it returns the first 1000 top documents with their *Lucene* score. In the

case of a query mode *boolean+ranked*, it first gives priority to the documents retrieved from the boolean Query and holds their score. The documents provided by the ranked query are added to the list of relevant documents, but their score is then re-scaled using the score of the last boolean document retrieved (the document with lower score of the boolean retrieval). Finally, the first 1000 top documents are selected.

7.2 GeoCLEF 2006 System Description

This section describes our system for Geographical Information Retrieval (GIR) in the context of our participation in the CLEF 2006 GeoCLEF Monolingual English task.

Our GIR system is a modified version of the system presented in GeoCLEF 2005 (Ferrés et al., 2005a) with some changes in the retrieval modes and the Geographical Knowledge Base (KB). The system has four phases performed sequentially: i) a Keywords Selection algorithm based on a linguistic and geographical analysis of the topics, ii) a geographical document retrieval with *Lucene*, iii) a textual document retrieval with the *JIRS* Passage Retrieval (PR) software, and iv) a Document Ranking phase. A Geographical KB has been build using a set of publicly available geographical gazetteers and the *Alexandria Digital Library (ADL) Feature Type Thesaurus*.

7.2.1 Collection Processing

We processed the entire English collections: Glasgow Herald 1995 and Los Angeles Times 1994 with linguistic tools (described in the next sub-section) to mark the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process, the collection is analyzed with a Geographical KB (described in the next sub-section). This information was used to build two indexes: one with the geographical information of the documents and another one with the textual and geographical information of the documents. We have used two Information Retrieval (IR) systems to create these indexes: *Lucene*[2] for the geographical index and *JIRS*[3] for the textual and geographical index (see a sample of both indexes in Figure 1). These indexes are described below:

- **Geographical Index**: this index contains the geographical information of the documents and its Named Entities. The Geographical index contains the following fields for each document:
  - **docid**: this field stores the document identifier.
  - **ftt**: this field indexes the feature type of each geographical name and the Named Entity classes of all the NEs appearing in the document.
  - **geo**: this field indexes the geographical names and the Named Entities of the document. It also stores the geographical information (feature type, hierarchical ancestors' path, and coordinates) about the place names. Even if the place is ambiguous all the possible referents are indexed.

[2] **Lucene**. `http://jakarta.apache.org/lucene`
[3] **JIRS.** `http://leto.dsic.upv.es:8080/jirs`

- **Textual and Geographical Index**: this index stores the lemmatized content of the document and the geographical information (feature type, hierarchical ancestors' path, and coordinates) about the geographic place names appearing in the text. If the geographical place is ambiguous then this information is not added to the indexed content.

System	Indexed Content	
Lucene	docid	GH950102000000
	ftt	regions@land_regions@continents administrative_areas@political_areas@countries_1st_order_divisions administrative_areas@populated_places@cities administrative_areas@political_areas@countries ...
	geo	Europe Asia@Western_Asia@Saudi_Arabia@Hejaz@24.5_38.5 America@Northern_America@United_States@South_Carolina @Lodge@32.9817_-80.952 America@Northern_America@United_States@38.91_-96.19 ...
JIRS	... the role of the wheel in lamatrekking , and where be the good place to air your string vest. pity the crew who accompany him on his travel as sayle of Arabia *countries_1st_order_divisions* **Asia Western_Asia Kuwait Arabia 25.0_45.0** along the Hejaz *countries_1st_order_divisions***Asia Western_Asia Saudi_Arabia Hejaz 24.5_38.5** railway line from Aleppo *countries_1st_order_divisions* **Asia Middle_East Syria Aleppo 36.0_37.0** in Northern_Syria *countries* **Asia Middle_East Syria 35.0_38.0** to Aqaba *cities* **Asia Western_Asia Jordan Maán Aqaba 29.517_35** in Jordan *countries* **Asia Western_Asia Jordan 31.0_36.0**. as he journey through the searing heat in an age East German ' biscuit tin ' , his good humour be sorely test ...	

Figure 7.5: Samples of an indexed document with *Lucene* and *JIRS*.

7.2.2 Topic Analysis

The goal of this phase is to extract all the relevant keywords (with its analysis) from the topics. These keywords are then used by the document retrieval phases. The Topic Analysis phase has three main components: a Linguistic Analysis, a Geographical Analysis, and a Keyword Selection algorithm.

Linguistic Analysis.

This process extracts lexico-semantic and syntactic information using these Natural Language Processing tools: i) *TnT* part-of-speech tagger (Brants, 2000), ii) *WordNet 2.0 lemmatizer*, iii) *Spear*[4]

[4]**Spear.** `http://www.lsi.upc.edu/~surdeanu/spear.html`

(a modified version of the *Collins parser* (Ferrés et al., 2005)), and iv) a *Maximum Entropy* based Named Entity Recognizer and Classifier (NERC) trained with the CONLL-2003 shared task English data set (Tjong Kim Sang & De Meulder, 2003b).

Geographical Analysis.

The Geographical Analysis is applied to the NEs from the title, description, and narrative tags that have been classified as *location* or *organization* by the NERC tool. This analysis has two components:

- **Geographical Knowledge Base:** this component has been built joining four geographical gazetteers: *GEOnet Names Server (GNS)*, *Geographic Names Information System (GNIS)* (using only a subset of 39,906 entries with the most important places), *GeoWorldMap Gazetteer*, and the *World Gazetteer* (adding only the 29,924 cities with more than 5,000 inhabitants). A detailed description of these gazetteers can be found in Section 4.1.
- **Geographical Feature Type Thesaurus**: the feature type thesaurus of our Geographical KB is the *ADL Feature Type Thesaurus (ADLFTT)*. The *ADL Feature Type Thesaurus* is a hierarchical set of geographical terms used to type named geographic places in English (Hill, 2000). Both *GNIS* and *GNS* gazetteers have been mapped to the *ADLFTT*, with a resulting set of 575 geographical types. Our *GNIS* mapping is similar to the one exposed in (Hill, 2000).

Topic Keywords Selection.

This algorithm extracts the most relevant keywords of each topic. The algorithm was designed for GeoCLEF 2005 (Ferrés et al., 2005a) (consult Section 7.1.3. for more details about this algorithm). Once the keywords are extracted, three different Keyword Sets (KS) are created (see an example in Figure 7.6):

- **All**: all the keywords extracted from the topic tags.
- **Geo**: geographical places or feature types appearing in the topic tags.
- **NotGeo**: all the keywords extracted from the topic tags that are not geographical place names or geographical types.

7.2.3 Geographical Document Retrieval with *Lucene*

Lucene is used to retrieve geographically relevant documents given a specific Geographical IR query. *Lucene* uses the standard tf-idf weighting scheme with the cosine similarity measure and allows ranked and boolean queries. We used boolean queries with a *Relaxed geographical search policy* (see Section 7.1.3 for more details). This search policy allows to retrieve 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 the U.S. places (e.g. America@Northern_America@United_States@Ohio).

Topic	EN-title	Wine regions around rivers in Europe
	EN-desc	Documents about wine regions along the banks of European rivers.
	EN-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.
Keywords Set (KS)	Not Geo	wine European
	Geo	Europe#location#regions@land_regions@continents#Europe regions hydrographic_features@streams@rivers
	All	wine regions rivers European Europe

Figure 7.6: Keyword sets sample of Topic 026.

7.2.4 Document Retrieval using the *JIRS* Passage Retriever

The *JIRS* Passage Retrieval System (Soriano et al., 2005) is used to retrieve relevant documents related to a GIR query. *JIRS* is a Passage Retriever specially designed for Question Answering (QA). This system gets passages with a high similarity between the largests n-grams of the question and the ones in the passage. We used *JIRS* considering a topic keyword set as a question. Then, we retrieved passages using the n-gram distance model of *JIRS* with a length of 11 sentences per passage. We obtained the first 100.000 top-scored passages per topic. Finally, a process selects the relevant documents from the set of retrieved passages. Two document scoring strategies were used:

- **Best**: the document score is the score of the top-scored passage in the set of the retrieved passages that belong to this document.
- **Accumulative**: the document score is the sum of the scores of all the retrieved passages that belong to this document.

7.2.5 Document Ranking

This component ranks the documents retrieved by *Lucene* and *JIRS*. First, the top-scored documents retrieved by *JIRS* that appear in the document set retrieved by *Lucene* are selected. Then, if the set of selected documents is less than 1,000 the top-scored documents of *JIRS* that not appear in the document set of *Lucene* are selected with a lower priority than the previous ones. Finally, the first 1,000 top-scored documents are selected. On the other hand, when the system uses only *JIRS* for retrieval only the first 1,000 top-scored documents by *JIRS* are selected.

7.3 Experiments and Results at GeoCLEF 2005

We designed a set of four experiments that consist in applying different query strategies and tags to an automatic GIR system (see Table 7.1). Two baseline experiments have been performed: the runs *geotalpIR1* and *geotalpIR2*. These runs differ uniquely in the Query type used: a *boolean+ranked* retrieval in *geotalpIR1* run and only *ranked* retrieval in *geotalpIR2* run. These runs consider the

Title and Description tags, and they use the ”lemma” index field. The third run (*geotalpIR3*) differs from the previous ones in the use of the Location tag (considering Title, Description and Location) and uses the ”Geo” field instead of the ”lemma” field. The ”Geo” field is used with a Strict Query search policy. This run also performs a *boolean+ranked* retrieval. The fourth run (*geotalpIR4*) is very similar to the third run (*geotalpIR3*), but uses a Relaxed Query search policy.

Table 7.1: Description of the Experiments at GeoCLEF 2005.

Run	Run type	Tags	Query Type	Geo. Index	Geo. Search
geotalpIR1	automatic	TD	Boolean+Ranked	Lemma	-
geotalpIR2	automatic	TD	Ranked	Lemma	-
geotalpIR3	automatic	TDL	Boolean+Ranked	Geo	Strict
geotalpIR4	automatic	TDL	Boolean+Ranked	Geo	Relaxed

We can expect a considerable difference between the two first runs and the last ones, because the other ones used an index with geographical knowledge. The fourth run is expected to be better than the third, due to the use of a relaxed search policy, that can increase the recall. On the other hand, we avoided the use of the operation tag (e.g. south, in, near,...) because our system is not prepared to deal with this information. Finally, the use of the location tag in the last runs is not so relevant, because our NERC and Geographical Thesaurus are able to detect the place names from the Title and Description tags with high performance.

The results of the GeoTalpIR system at the GeoCLEF 2005 Monolingual English task are summarized in Table 7.2. This table shows the following IR measures for each run: *Average Precision*, *R-Precision*, *Recall*, and the increment over the median of the average precision (0.2063) obtained by all the systems that participated in the GeoCLEF 2005 Monolingual English task.

Table 7.2: GeoCLEF 2005 results.

Run	Tags	AvgP.	R-Prec.	Recall (%)	Recall	Δ AvgP. Diff.(%) over GeoCLEF AvgP.
geotalpIR1	TD	0.1923	0.2249	49.51%	509/1028	-6.78%
geotalpIR2	TD	0.1933	0.2129	49.22%	506/1028	-6.30%
geotalpIR3	TDL	0.2140	0.2377	62.35%	641/1028	+3.73%
geotalpIR4	TDL	**0.2231**	**0.2508**	**66.83%**	**687/1028**	**+8.14%**

The results show a substantial difference between the two first runs and the two last ones, specially in the recall measure: 49.51% and 49.22% respectively in the first and second run (*geotalpIR1* and *geotalpIR2*) and 62.35% and 66.38% respectively in the third and fourth run (*geotalpIR3* and *geotalpIR4*). The recall is also improved by the use of Geographical Knowledge and a relaxed policy over the ”Geo” Field as it is seen in run four (*geotalpIR4*). Finally, in the last run (*geotalpIR4*) we obtained results about +8.14% better than the median of the average obtained by all runs (0.2063).

7.4 Experiments and Results at GeoCLEF 2006

We designed a set of five experiments that consist in applying different IR systems, query keyword sets, and tags to an automatic GIR system (see Table 7.3). Basically, these experiments can be divided in two groups depending on the retrieval engines used:

- **JIRS**. Two baseline experiments have been done in this group: the runs *TALPGeoIRTD1* and *TALPGeoIRTDN1*. These runs differ uniquely in the use of the narrative tag in the second one. Both runs use one retrieval system, *JIRS*, and they use all the keywords to perform the query. The experiment *TALPGeoIRTDN3* is similar to the previous ones but uses a Cumulative scoring strategy to select the documents with *JIRS*.

- **JIRS & Lucene**. The runs *TALPGeoIRTD2* and *TALPGeoIRTDN2* use *JIRS* for textual document retrieval and *Lucene* for geographical document retrieval. Both runs use the *Geo* keywords set for *Lucene* and the *NotGeo* keywords set for *JIRS*.

Table 7.3: Description of the experiments at GeoCLEF 2006.

Automatic Runs	Tags	IR System	JIRS KS	Lucene KS	*JIRS* Score
TALPGeoIRTD1	TD	JIRS	All	-	Best
TALPGeoIRTD2	TD	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN1	TDN	JIRS	All	-	Best
TALPGeoIRTDN2	TDN	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN3	TDN	JIRS	All	-	Cumulative

The results of the TALP-GeoIR system at the CLEF 2006 GeoCLEF Monolingual English task are summarized in Table 7.4. This table has the following IR measures for each run: *Average Precision*, *R-Precision*, and *Recall*.

The results show a substantial difference between the two sets of experiments. The runs that use only *JIRS* have a better *Average Precision*, *R-Precision*, and *Recall* than the ones that use *JIRS* and *Lucene*. The run with the best *Average Precision* is *TALPGeoIRTD1* with 0.1342. The best *Recall* measure is obtained by the run *TALPGeoIRTDN1* with a 68.78% of the relevant documents retrieved. This run has the same configuration that the*TALPGeoIRTD1* run but uses the narrative tag. Finally, we obtained poor results in comparison with the mean average precision (0.1975) obtained by all the systems that participated in the GeoCLEF 2006 Monolingual English task.

We have applied *JIRS*, a state-of-the-art PR system for QA, to the GeoCLEF 2006 Monolingual English task. We also have experimented with an approach using both *JIRS* and *Lucene*. In this approach *JIRS* was used only for textual document retrieval and *Lucene* was used to detect the geographically relevant documents. The approach with only *JIRS* was better than the one with *JIRS* and *Lucene* combined.

Comparatively with the Mean Average Precision (MAP) of all the runs participating at GeoCLEF 2006 Monolingual English task our MAP is low. This fact can be due to several reasons: i)

Table 7.4: TALP-GeoIR results at GeoCLEF 2006 Monolingual English task.

Automatic Runs	Tags	IR System	AvgP.	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	JIRS	**0.1342**	**0.1370**	60.84%	230/378
TALPGeoIRTD2	TD	JIRS+Lucene	0.0766	0.0884	32.53%	123/378
TALPGeoIRTDN1	TDN	JIRS	0.1179	0.1316	**68.78%**	**260/378**
TALPGeoIRTDN2	TDN	JIRS+Lucene	0.0638	0.0813	47.88%	181/378
TALPGeoIRTDN3	TDN	JIRS	0.0997	0.0985	64.28%	243/378

the *JIRS* PR system may be was not used appropriately or is not suitable for GIR, ii) our system is not dealing with geographical ambiguities, iii) the lack of textual query expansion methods, iv) the need of Relevance Feedback methods, and v) errors in the Topic Analysis phase.

Chapter 8

Geographical Named Entity Subclassification

This chapter describes our two approaches to deal with the Geographical Named Entity Subclassification (GNES) task. Both approaches apply Machine Learning techniques for a finer grained classification of NEs that have been previously classified as locations by a general purpose NERC system. The first approach uses the Inductive Logic Programming (ILP) paradigm and the second one uses Support Vector Machines (SVMs). Both approaches use features from the local linguistic context of the Named Entity to perform the sub-classification.

Our approaches to Geographical Named Entity Subclassification need the following Natural Language processing steps:

- *Part-of-speech (POS) tagging*. This basic step recognizes the word forms and selects their part-of-speech tags.

- *Lemmatization*. This step consists in obtaining the lemma of a word.

- *Named Entity Recognition (NER)*. Recognizing consists on locating a sequence of one or more contiguous words that can be considered candidate to be a NE and deciding if it is an actual one.

- *Named Entity Classification (NEC)*. Classifying implies assigning a class from a closed dataset to the NE.

- *Named Entity Subclassification (NES)*. Given a named entity and its Named Entity class (LOCATION in this case), the system decides which subclass of the previous classes is used in the context. For instance, in a LOCATION class, the Named Entity *Buffalo* could be a city, or a river. The NES system tries to subclassify this NE using features from the document in which the Named Entity appears and optionally features from external resources to decide in which subclass pertains.

8.1 Inductive Logic Programming Approach

Our approach is based on the acquisition of pre-classified knowledge from gazetteers and the use of a Machine Learning algorithm to train binary classifiers for each geographical class. This approach uses only local linguistic context.

8.1.1 Knowledge Acquisition

In order to obtain classifiers to disambiguate geographical places, is very important to have a training corpus with good examples. Three gazetteers and a huge news corpus have been used to obtain these examples. In the system presented here, the following steps have been applied to acquire examples to train the classifiers:

- **Gazetteer creation.** First, a geographical gazetteer has been built joining two huge gazetteers that contain entries with the places and their geographical class, coordinates, and other information:

  – GEOnet Names Server (GNS)[1]: a gazetteer covering worldwide excluding the United States and Antarctica, with 5.3 million entries.

  – Geographic Names Information System (GNIS)[2], contains information about physical and cultural geographic features in the United States and its territories. This gazetteer has 2.0 million entries.

  The core of our system is an Inductive Logic Programming (ILP) learner that learns, from a set of positive and negative examples, a ranked list of rules to obtain a binary classifier for each geographical class. Both natural geographical entities (Sea, Mountain, River, etc.) and political or organizational divisions (Country, State, Province, City, etc.) are considered. Our learner (we have used Quinlan's FOIL (Quinlan, 1990)) follows a supervised schema, so a training set has been collected and automatically tagged. What has to be learned is the dependence of the different types of location on the context of their occurrences.

8.1.2 Learning Methodology

In the system presented here, the following learning methodology has been applied:

– Firstly, an initial set of sources of highly confident classified resources has been selected. We have used the MUC6 Reference Gazetteer complemented with location names extracted from five different web sites (see Table 8.1). The information extracted includes not only the basic terminological information (i.e. lists of tagged NEs) but also some spatial relations (e.g. states in a country, islands in a sea, etc.). Up to 133,744 geographical names classified into 18 classes have been extracted in this way (with a very

[1] **GNS**. http://gnswww.nima.mil/geonames/GNS/index.jsp
[2] **GNIS**. http://geonames.usgs.gov/geonames/stategaz

irregular distribution, from 117,598 cities to only 3 forests). Table 8.2 shows the number of names per class.

http://www.world-gazetteer.com/
http://people.depauw.edu/djp/
http://www.worldatlas.com/
http://en.wikipedia.org/
http://www.gazeteer.com/

Table 8.1: Web sites used to extract the gazetteer.

Classes	Number
Airport	729
City	117,598
Country	303
Country-zone	220
Desert	43
Forest	3
Gulf	22
Island	917
Island-sea	698
Lake	47
Mountains	27
Peak	2,218
Port	4,641
Province	5,331
River	333
Sea	45
State	530
Volcano	39
Total	133,744

Table 8.2: Number of geographical names per class.

– From this initial set we have removed all the NEs belonging to more than one class in order to reduce, as much as possible, the use of contexts corresponding to ambiguous NEs.
– We have merged the classes with few members and semantically related (e.g. port and airport, mountain and peak), dropped out poorly represented classes and selected a maximum of 500 names per class. In addition, we have performed a shallow manual revision. A total of 11 classes remained after this step: mountain or peak, river, sea, lake, island, desert, port or airport, city, country, state and province.
– We have looked for the first 500 occurrences of the members of these lists in the AQUAINT [3] corpus. A previous preprocess was carried out including POS tagging

[3]The corpus has been used for our participation in TREC-2003. More information about AQUAINT corpus can be obtained at http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31

with *TnT* (Brants, 2000) and NERC with *Abionet* (Carreras et al., 2003a). Restricting the number of names per class and the number of examples per name is needed for getting the resulting set as balanced as possible.

– From these corpus we have extracted the context, up to 10 tokens on each side, of each occurrence as well as the needed morphological information. This procedure resulted in a total of 110,576 examples (see last column in Table 8.4).

With this material we have fed FOIL (Quinlan, 1990) to learn one classifier for each class. FOIL is a relational learning system aimed at inductively learning first-order rules (in prolog format) from positive and negative examples. By default, FOIL considers the *close-world assumption* to automatically generate the set of negative examples, meaning that all non-positive elements are negative ones. We have used, however, the examples corresponding to each particular class as positive examples for learning this class and the examples related to the rest of classes as negative ones. This experimental setting has proved to provide better results than the multi-class approach with *close-world assumption*. With respect to the background knowledge used to learn, FOIL requires each of the examples, positive and negative ones, to be represented as a set of predicates. For our particular learning problem we have used the features presented in table 8.3 from which the following set of propositional predicates has been designed:

– Context predicates:
  * diw_x: the ith word in the direction d (right or left) is x.
  * dip_x: the ith POS-tag in the direction d is x.
  * dis_x: x is the NE class (LOC, ORG, PER and MISC) about ith word in the direction d.
– Internal NE predicates:
  * zi_x: the ith token of the NE is x (i could be 1 or 0, for the two last tokens of the NE)

In all cases, with the exception of case three, i can be omitted, it means that the information appears in any position in the direction d.

8.1.3 Experiments

We have designed a set of three experiments to decide which predicates are the best to learn to classify geographical NEs. In these experiments we have only changed the features related to the NE (i.e. internal predicates as tokens *z1_Lake* and *z0_Garda* in the case of *Lake_Garda*), and we do not have modified context predicates. All the experiments have had the same set of context predicates. The following experiments have been done:

1. Experiment with all the predicates previously explained.
2. Experiment only with context predicates.

Feature type	Features
lexical information	- Bag of words of positions -5 to +5 (NE not included). - Words in position from -1 to -3. - Words in position from +1 to +2. - Two last Tokens included in the NE.
morphological information	- Bag of POS of positions -5 to +5 (NE not included). - POS in position from -1 to -3. - POS in position from +1 to +2.
semantic information	- NE class of positions -1 to -3. - NE class of positions +1 to +2

Table 8.3: Features used by FOIL.

3. Experiment with all context predicates and using internal predicates only for NEs having more than 1 token (i.e. *Lake_Garda*):

FOIL has learned a set of binary classifiers for each class. We have used the k-Fold Cross-Validation measure to evaluate these classifiers. The k parameter means the number of sets to split the examples, k has been set to 5. We have balanced the number of examples used to learn the classifiers taking a threshold of 1200 in classes with many examples (see column 2 of Table 8.4).

Classes	#Examp. (5CV)	#Examp. (total)
Airport+Port	376	376
City	1,200	25,000
Country	1,200	25,000
Desert	517	517
Island	1,200	7,259
Lake	1,447	1,447
Mountains+Peak	1,186	1,186
Province	1,200	23,399
River	1,200	5,189
Sea	1,200	20,050
State	850	850
Total	11,576	110,273

Table 8.4: Number of examples used in 5-CV and total.

8.1.4 Results

The results of the three 5-fold cross-validation experiments are summarized in tables 8.5, 8.6 and 8.7. These tables contain the following average evaluation measures of 5-fold test sets

for each class: precision, recall, F_1[4] and the variance of F_1. As shown in these tables, experiment 1 achieves the best overall performance with an 0.9553 average measure of F_1. It has produced 613 rules for all classes (an average of 55.72 rules per class). However, experiment 1 uses internal predicates that produce over-fitting. These predicates are the last two tokens of the NE (i.e. *z0_York*, *z1_New*, *z_York* and *z_New*). Using these predicates can be useful to capture some relevant features of the NE (i.e. capturing *z0_River* in *Colorado_River*), especially in these classes: airport+port (*z1_Airport*,*z0_Port*), desert (*z0_Desert*), lake (*z1_Lake*), mountains+peak (*z1_Mount*,*z0_Peak*), river (*z0_River*) and sea (*z0_Sea*, *z0_Ocean*). Besides, in the case of NEs having only one token it can affect negatively to the learning rules (i.e. capturing *z0_Sahara* in *Sahara*). Concluding, we cannot obtain robust rules using internal predicates with NEs having only one token.

Classes	Precision	Recall	F_1	varF_1
airport+port	0.7850	0.9632	0.8509	0.0086
city	0.8293	0.9475	0.8821	0.0006
country	0.8796	0.9017	0.8883	0.0014
desert	0.9980	0.9942	0.9961	0.0000
island	0.9908	0.9817	0.9862	0.0000
lake	1.0000	1.0000	1.0000	0.0000
mountains+peak	0.9873	0.9601	0.9729	0.0004
province	0.9341	0.9550	0.9440	0.0003
river	0.9992	0.9950	0.9971	0.0000
sea	1.0000	0.9983	0.9992	0.0000
state	0.9873	0.9953	0.9912	0.0000
Total Avg	0.9446	0.9720	0.9553	0.0011

Table 8.5: Results of 5-fold cross validation with internal predicates (Experiment 1).

Classes	Precision	Recall	F_1	varF_1
airport+port	0.7544	0.8222	0.7729	0.0249
city	0.6460	0.8183	0.7146	0.0007
country	0.6657	0.8833	0.7557	0.0010
desert	0.6271	0.7851	0.6954	0.0009
island	0.7228	0.8600	0.7839	0.0005
lake	0.6107	0.8527	0.7044	0.0008
mountains+peak	0.6552	0.6655	0.6587	0.0118
province	0.6306	0.9075	0.7391	0.0021
river	0.7842	0.9108	0.8400	0.0005
sea	0.7294	0.8817	0.7959	0.0010
state	0.6311	0.8188	0.7108	0.0004
Total Avg	0.6779	0.8369	0.7429	0.0041

Table 8.6: Results of 5-fold cross validation with only context predicates (Experiment 2).

[4] F_β is the harmonic mean of recall (ρ) and precision (π) (Rijsbergen, 1979). The F_β function formula is: $F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho}$.

Classes	Precision	Recall	F_1	varF_1
airport+port	0.8161	0.9082	0.8413	0.0087
city	0.6975	0.8367	0.7606	0.0003
country	0.7465	0.8408	0.7894	0.0012
desert	0.8201	0.7935	0.7981	0.0055
island	0.7290	0.8917	0.8004	0.0006
lake	1.0000	1.0000	1.0000	0.0000
mountains+peak	0.9791	0.9516	0.9644	0.0008
province	0.6431	0.7225	0.6803	0.1162
river	0.9950	0.9908	0.9929	0.0000
sea	0.9224	0.9442	0.9311	0.0007
state	0.7556	0.9059	0.8223	0.0009
Total Avg	0.8277	0.8896	0.8528	0.0123

Table 8.7: Results of 5-fold cross validation with reduced internal predicates (Experiment 3).

More and most robust rules have been learned with experiments 2 and 3. These rules have been reported the following measures of F_1 in average 0.7429 and 0.8528 respectively. These latter experiments have produced 4695 and 2139 rules, respectively, with an average of 426.62 and 194.45 rules per class, respectively. Examples of the best ranked rules obtained for desert class can be seen in figures 8.1 and 8.2.

One of the advantage of using an ILP system as FOIL is the readability of learned rules. This property allows to easily analyze these rules and modify or remove those which are considered irrelevant ones. This is the case of rules:

desert(A) :- l1w_the(A), rp_RB(A), not(lp_NN(A)).

desert(A) :- lw_villages(A).

The following rules can be considered relevant rules:

desert(A) :- l1w_the(A), not(z0_River(A)), not(z0_Sea(A)), not(z0_Mountains(A)), not(rs_NNP(A)), not(ls_VB(A)), not(z0_State(A)), rs_IN(A), not(r1p_IN(A)).

desert(A) :- rw_desert(A).

desert(A) :- rw_desert(A).
desert(A) :- l1w_the(A), rp_RB(A), not(lp_NN(A)).
desert(A) :- l1w_the(A), not(rp_NNP(A)), lp_VBN(A), not(lp_NN(A)), not(rp_VB(A)).
desert(A) :- l1w_the(A), not(rp_NNP(A)), not(rp_JJ(A)), rp_VBZ(A), not(lp_JJ(A)), not(r2p_VBZ(A)).
desert(A) :- lw_the(A), not(lp_NNS(A)), l3p_RB(A), not(lp_VBN(A)).
desert(A) :- not(rp_NNP(A)), l2w_in(A), not(r1p_IN(A)), rp_NN(A), not(rp_,(A)), not(r1p_NN(A)).
desert(A) :- lw_the(A), rp_,(A), rp_DT(A), not(r2w_the(A)), not(l2p_IN(A)), not(lw_in(A)).
desert(A) :- lw_the(A), rp_,(A), rp_RB(A), not(l3p_NN(A)), not(rp_VBD(A)).
desert(A) :- lw_desert(A), not(rw_of(A)).
desert(A) :- l2p_IN(A), rw_in(A), not(l3p_NNS(A)), not(l3p_NN(A)).

Figure 8.1: First rules obtained with only context predicates (Experiment 2).

Although no direct evaluation on a test corpus has been performed, we can guess that small variance of F_1 measure resulting in most of the classes and experiments is a clear indicator

desert(A) :- l1w_the(A), not(z0_River(A)), not(z0_Sea(A)), not(z0_Mountains(A)), not(rp_NNP(A)), not(lp_VB(A)), not(z0_State(A)), rp_IN(A), not(r1p_IN(A)).
desert(A) :- z0_Valley(A).
desert(A) :- rw_desert(A).
desert(A) :- l1w_the(A), not(z0_River(A)), not(z0_Sea(A)), not(rs_IN(A)), lw_of(A), not(r2p_RB(A)).
desert(A) :- not(z0_River(A)), not(z0_Sea(A)), l1w_the(A), not(z0_Mountains(A)), not(r1p_NN(A)), not(l3p_NN(A)), not(z0_State(A)), not(r2p_NNP(A)), not(l3s_LOC(A)), not(rw_an(A)).
desert(A) :- lw_the(A), not(z0_River(A)), not(z0_Sea(A)), not(rw_of(A)), l2p_IN(A), rs_,(A), not(rs_,(A)), not(r2p_NN(A)).
desert(A) :- lw_south(A), not(l3w_south(A)).
desert(A) :- l1w_western(A).
desert(A) :- rw_Israel(A).
desert(A) :- lw_villages(A).

Figure 8.2: First rules obtained with internal NE predicates, but only NEs having more than 1 token (Experiment 3).

that we can obtain similar results by training with the whole training set and testing with a test corpus.

8.2 SVM Approach for GNES

The experiments described in this Section are the continuation of the previous ones experiments but with some improvements:

- Highly-confident data sources: in the previous work we used as a source of training lists of places collected from Internet.. In the actual, we use high-confident gazetteers from geological organizations.
- Machine Learning with SVM.
- Different set of features. The number of features have changed, also the number of examples and the size of the context used to train.
- Location sub-ontology: we used an ontology to map our geographical classes, and we have learnt at different levels of the ontology. In our previous work we had not learnt at different levels of the ontology, and we used a set of 11 classes at the bottom level to learn.

The core of our system is a SVM learner that learns, from a set of positive and negative examples, a binary classifier for each geographical class. Both natural geographical entities (Sea, Mountain, River, etc.), political or organizational divisions (Country, County, Province, City) and facilities (Airport, Building, Park, etc.) are considered. Our learner follows a supervised schema, so a training set has been collected and automatically tagged. What has to be learned is the dependence of the different types of location on the context of their occurrences and its internal features.

Each one of these gazetteers have a different set of classes. We have mapped this sets to our set of classes (see Table 8.8), which includes the most common classes and the most

important (i.e. country is not common, but important)). The gazetteer contains approximately 3.7 million of places with its geographical class. This step is similar to the (Manov et al., 2003) approach, but they used a limited number of locations (only the 50,000 most important).

- **Extraction of non-ambiguous places from our Gazetteer.** From the gazetteer we have removed all the NEs belonging to more than one class in order to reduce, as much as possible, the use of contexts corresponding to ambiguous NEs.

- **Filtering using the Alexandria Gazetteer.** We used this gazetteer (5.94 million of places, with their coordinates but without its subclassification) to filter non-ambiguous named entities. We selected all the non-ambiguous geographical places from the Alexandria Digital Library Project[5] (ADLP), and we used this NEs to filter the previous list of non-ambiguous places extracted from our gazetteer (GNIS+GNS). Finally, we obtain a list of non-ambiguous places using the information of three gazetteers (GNIS+GNS+ADLP). Up to 385,364 geographical names classified into 24 classes have been extracted in this way (with a very irregular distribution, from 166,564 cities to only 40 deserts). Table 8.8 shows the number of names per class.

- **Corpus pre-processing.** We have used the AQUAINT [6] corpus to extract the examples for learning. This corpus is a large collection of news in English (more than 3 Gbytes) extracted from the Associated Press Journal (APW), the New York Times (NYT) and the Xinhua English (XIE). A previous preprocess was carried out including POS tagging (Brants, 2000), lemmatization (using WordNet 1.7.1) and *Abionet*, a NERC system (Carreras et al., 2003a).

- **Examples extraction from Corpus**. From this corpus we have extracted the context, in this case the sentence, of each occurrence as well as the needed lexical, morphological and semantic information: (words, POS, lemmas and Named Entity tags). This procedure resulted in a total of 1,218,491 examples (see last column in Table 8.12). We removed classes without examples (such as beach).

- **Ontology mapping.** A location sub-ontology has been created using Sekine's ontology (Sekine et al., 2002) and we mapped our classes to this ontology. In this process the classes district and urban_place have been removed due to their similarity with city. The final ontology has three-levels (see below). At the first level with three nodes (facility, political_region and geological_region). The second level has 11 nodes, and the third level has 15 nodes. The location sub-ontology is shown here:

```
FACILITY
  BUILDING
  PARK
  STATION
     AIRPORT
     PORT
```

[5] **Alexandria Digital Library Project**, http://www.alexandria.ucsb.edu

[6] The corpus has been used for our participation in TREC-2004. More information about AQUAINT corpus can be obtained at http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31

Classes	Number
airport	4104
bay	3606
beach	225
building	10236
cape	4384
channel	2066
city	166564
country	318
county	13320
desert	40
district	8916
gulf	71
island	10161
lake	7446
mountain	43237
park	238
port	774
province	2348
river	62001
sea	75
sea_accident	4717
urban_place	35895
valley	3959
water_reserve	663
Total	385,364

Table 8.8: Number of geographical non-ambiguous names per class.

```
GEOLOGICAL_REGION
  ISLAND
  LAND_FORM
     DESERT
     MOUNTAIN
     VALLEY
  SEA_FORM
     BAY
     CAPE
     GULF
     SEA
     SEA_ACCIDENT
  WATER_FORM
     CHANNEL
     LAKE
     RIVER
      WATER_RESERVE
POLITICAL_REGION
  CITY
  COUNTY
  COUNTRY
  PROVINCE
```

8.2.1 Machine Learning

The Support Vector Machines (SVM) algorithm has been used to learn Geographical Name Disambiguation. SVM is a supervised Machine Learning (ML) algorithm that learns from a set of training data previously classified a decision function that accurately predicts the class of unseen examples. SVM tries to compute the hyperplane that best separates the set of training examples (the hyperplane with maximum margin) (Vapnik, 1995). Sometimes, when a set of examples is not linearly separable is possible to use kernels. The use of kernels implies a transformation of the vectors to a high-dimensional space using non-linear functions. In order to allow misclassification a positive parameter C, is used to relax the fact that all the examples must be classified correctly.

We have used SVM-light[7] to learn one binary classifier for each class. For our particular learning problem we have used the features presented in Table 8.9.

Feature type	Features
lexical information	- Bag of words of positions -5 to +5 (NE not included). - Bag of words of the left context of the sentence (NE not included). - Bag of words of the right context of the sentence (NE not included). - Bag of words of all the sentence. (NE not included) - Words in position from -1 to -5. - Words in position from +1 to +5. - All the possible combinations of tokens included in the NE. (NE not included)
morphological information	- Bag of lemmas of positions -5 to +5 (NE not included). - Bag of lemmas of the left context of the sentence (NE not included). - Bag of lemmas of the right context of the sentence (NE not included). - Bag of lemmas of all the sentence. (NE not included) - Part-of-speech (POS) in position from -1 to -5. - Part-of-speech (POS) in position from +1 to +5. - Lemmas in position from -1 to -5. - Lemmas in position from +1 to +5.
semantic information	- NE in positions from -1 to -5. - NE in positions +1 to +5

Table 8.9: Features used by SVM.

8.2.2 Experiments

A set of three experiments that consists in applying k-Fold Cross-Validation at different levels of the ontology have been designed:

[7] **SVM-light**. http://svmlight.joachims.org

1. Top level of the ontology: creating classifiers for only 3 classes: *facility, geological_region, and political_region* .

2. Middle level of the ontology: creating classifiers for 11 classes: *city, country, county, province, building, park, station, land_form, water_form, island, and sea_form* .

3. Bottom level of the ontology: creating classifiers for the final 21 classes: *airport, bay, building, cape, channel, city, country, county, desert, gulf, island, lake, mountain, park, port, province, river, sea, sea_accident, valley, and water_reserve* .

SVM has learnt a binary classifier for each class. We have used the k-Fold Cross-Validation measure to evaluate these classifiers. The k parameter means the number of sets to split the examples, k has been set to 5. We have balanced the number of examples used to learn the classifiers taking a threshold of 5000 to the terminal nodes in the ontology that have many examples (see column 2 of Table 8.12). The SVM configuration is the following: lineal kernels and parameter C is zero.

The number of examples used to train at the different levels of the ontology is shown in tables 8.10, 8.11 and 8.12 (found in pages from 120 on).

Classes	#Examp. (5CV)
facility	1,480
geological_region	20,000
political_region	36,008
Total	57,488

Table 8.10: Number of examples used in 5-CV at top level.

Classes	#Examp. (5CV)
building	25
city	5,000
country	5,000
county	5,000
island	5,000
land_form	10,002
park	1,221
province	5,000
sea_form	6,902
station	234
water_form	14,104
Total	57,488

Table 8.11: Number of examples used in 5-CV at middle level.

Classes	#Examp. (5CV)	#Examp. (total)
airport	12	12
bay	101	101
beach	0	0
building	25	25
cape	347	347
channel	367	367
city	5,000	184,804
country	5,000	205,694
county	5,000	286,065
desert	2	2
gulf	177	177
island	5,000	137,594
lake	3,791	3791
mountain	5,000	73,292
park	1,221	1,221
port	222	222
province	5,000	164,084
river	5,000	307,877
sea	1,277	1,277
sea_accident	5,000	25,407
valley	5000	6,316
water_reserve	4,946	4,946
Total	57,488	1,218,491

Table 8.12: Number of examples used in 5-CV at bottom level and total.

8.2.3 Results

The results of the three 5-fold cross-validation experiments are summarized in tables 8.13, 8.14 and 8.15 (found in pages from 122 on). These tables contain the following average evaluation measures of 5-fold test sets for each class: precision, recall, F_1[8] and the variance of F_1. As shown in these tables, experiment 1 achieves the best overall performance with an 0.8805 average measure of F_1. This results are due to the use of the only three upper classes of the ontology.

Experiment 2 has achieved an overall performance of 0.8520 in F_1. In this experiment, only one class have a poor performance (building), this is due to the low number of examples of this class. Experiment 3 has achieved an overall performance of 0.8329 in F_1. This drop is expected because we use 21 classes instead of 3 and 11 in the previous experiments. Three classes (airport, building and desert) show poor performance, also because of their few number of examples.

The results can not be compared directly with our previous ILP experiments in this field (Ferrés et al., 2004b) (see Section 8.1), because for the SVM approach we used a different set of features, context, classes and number of classes. A comparison with other approaches can not be done due to

[8] F_β is the harmonic mean of recall (ρ) and precision (π) (Rijsbergen, 1979). The F_β function formula is: $F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho}$.

Classes	Precision	Recall	F_1	varF_1
facility	0.9032	0.8628	0.8824	0.00205973
geological_region	0.8420	0.9286	0.8831	0.00056728
political_region	0.8209	0.9393	0.8761	0.00044496
Total Avg	0.8553	0.9102	0.8805	0.00102399

Table 8.13: Results of 5-fold cross validation at the upper level of the ontology (Experiment 1).

Classes	Precision	Recall	F_1	varF_1
building	0.6317	0.8800	0.7251	0.04363486
city	0.8380	0.8124	0.8246	0.00399237
country	0.8680	0.8668	0.8671	0.00155266
county	0.8713	0.8420	0.8561	0.00233113
island	0.8989	0.8748	0.8864	0.00163007
land_form	0.8305	0.8875	0.8581	0.00133861
park	0.9234	0.8726	0.8958	0.00272126
province	0.8751	0.8068	0.8383	0.00386324
sea_form	0.9868	0.8566	0.9168	0.00050535
station	0.8496	0.8296	0.8393	0.00621171
water_reserve	0.8803	0.8511	0.8654	0.00149342
Total Avg	0.8594	0.8527	0.8520	0.00629769

Table 8.14: Results of 5-fold cross validation at the middle level of the ontology (Experiment 2).

the lack of a test corpus in the Geographical Name Disambiguation field (Leidner, 2004).
We can guess that small variance of F_1 measure resulting in most of the classes and experiments is a clear indicator that we can obtain similar results by training with the whole training set and testing with a test corpus.

Classes	Precision	Recall	F_1	varF_1
airport	0.8000	0.8000	0.8000	0.16000000
bay	0.9001	0.8476	0.8630	0.00711385
building	0.6317	0.8800	0.7251	0.04363486
cape	0.9383	0.8600	0.8930	0.00474829
channel	0.9943	0.9133	0.9504	0.00172881
city	0.8380	0.8124	0.8246	0.00399237
country	0.8680	0.8668	0.8671	0.00155266
county	0.8713	0.8420	0.8561	0.00233113
desert	0.4000	0.4000	0.4000	0.24000000
gulf	0.8010	0.7667	0.7643	0.05257043
island	0.8989	0.8748	0.8864	0.00163007
lake	0.9169	0.8445	0.8783	0.00290532
mountain	0.8614	0.8318	0.8461	0.00282648
park	0.9234	0.8726	0.8958	0.00272126
port	0.8271	0.7419	0.7792	0.00767270
province	0.8751	0.8068	0.8383	0.00386324
river	0.8669	0.8098	0.8363	0.00370021
sea	0.9858	0.9249	0.9542	0.00032318
sea_accident	0.9915	0.8688	0.9259	0.00034158
valley	0.8598	0.8726	0.8651	0.00467585
water_reserve	0.9213	0.7973	0.8430	0.02052891
Total Avg	0.8557	0.8206	0.8329	0.02708862

Table 8.15: Results of 5-fold cross validation at the bottom of the ontology (Experiment 3).

Chapter 9

Work Plan

This thesis is intended to study current research lines in Restricted-Domain NLP applications (including factoid QA and IR) and explore new research lines in these fields and specially for the Geographical Domain. As Restricted Domain for QA and IR research lines in the Geographical Domain are reached during the doctoral thesis work in the following months, we will study methodologies for generic Restricted-Domain adaptability. The following research lines of Natural Language Processing will be treated from now for completing the thesis project:

1. **An ILP Approach for Machine Learning for Answer Extraction**. The first research line is to apply supervised Machine Learning (ML) techniques to obtain a Knowledge Set of Answer Extraction rules for Open-Domain and Restricted-Domain QA. First, we plan to limit the experiments with few question types and then we plan to extend its application to the whole set of question types.

   A huge amount of <question, answer> pairs is required as a corpus for learning. Sources from story comprehension such as the REMEDIA (Hirschman et al., 1999) or CBC4Kids (Breck et al., 2001) corpora and the collections of TREC [1] and CLEF evaluations (both the set of questions and the valid answers provided by the organization), the Florent Jousse's QA corpus [2](Jousse et al., 2005), and the *Bilingual Reading Comprehension Corpus (BRCC)* (Xu & Meng, 2005), will be used as a train/test corpora.

   We plan to use the Inductive Logic Programming (ILP) paradigm. The involved predicates come from the semantic representation of both questions and passages (represented in environment structures). ILP is pretended to learn a set of weighted rules for each question type. Two sets of predicates will be included in the rule, context and condition. The context comes from the Mandatory Constraints of the question while condition tries to represent the expected predicates constraining the answer.

2. **A New Geographical Toponym Resolution Approach**. Experiments with a new Toponym Resolution (TR) algorithm are planned using the annotated corpus created by (Garbin &

[1] The QA data sets of the past evaluations are publicly available from the TREC website `http://trec.nist.gov/data/qamain.html`

[2] `http://www.grappa.univ-lille3.fr/~jousse/EN/corpusQR.html`

Mani, 2005). We consider using the following heuristics: ”one sense per discourse heuristic”, geo-scope resolution algorithm based on a variation of page-rank (Martins & Silva, 2005), and Geographical Knowledge. Geographical Knowledge includes exhaustive Geographical Knowledge Bases (GKBs) such as: GeoNet Names Server, GNIS, ADL, etc. and linguistic knowledge extracted from structured GKBs like GNS (Ferrés & Rodríguez, 2006a), and the Alexandria Digital Library Feature Type Thesaurus (ADLFTT) with more than 500 geographical sub-types.

3. **Experiments with Probabilistic Models for IR and PR**. This line will explore the use of the Terrier IR platform (Ounis et al., 2006) in the ODQA and GIR evaluation contests CLEF 2007, TREC 2007 and CLEF 2008. Preliminary experiments that we performed with Terrier using geographical knowledge achieved top performance results over the GeoCLEF 2006 participants (i.e achieving a MAP of 0.3162).

4. **Geographical Information Retrieval Experiments**. We will apply the new geographical TR algorithm for the GIR task in the context of the GeoCLEF 2007 and GeoCLEF 2008 evaluations.

5. **Geographical Question Answering Experiments**. We will also apply the new geographical TR algorithm for the Geographical QA task using our own corpus developed for the MLQA06 experiments (Ferrés & Rodríguez, 2006a)

6. **Study of Generic Restricted-Domain adaptability of our AE and TR algorithms**. We will try to adapt our algorithms to other domains such as: medicine, genomics, laws, etc.

7. **Open Domain Question Answering Experiments**. This year we plan to participate in CLEF 2007 and TREC 2007 for QA. There is also a possibility that we participate in CLEF 2008.

9.1 Thesis Project Scheduling

The final PhD thesis is expected to be completed in one year and half. Experiments with the research lines presented before should take a year, and then the doctoral thesis is expected to be written and concluded in 6 months. Machine Learning for Answer Extraction is the most important research line and is expected to take 8 months (from February 2007 to September 2007). An evaluation of the proposed technique should be completed for the TREC 2007 evaluation. After this period 4 months could be dedicated to study the adaptability of this approach to other domains such as: medicine, genetics, etc. The Geographical Toponym Resolution implementation and experiments should take at least 5 months. This implementation will be a framework for the new GIR and Geographical QA experiments, that should take 7 months after the final implementation of the TR is achieved. We expect the experiments with Probabilistic Models for IR and PR to be performed in 3 months.

Chapter 10

Related Publications

In this chapter, the papers published by the author in different conference proceedings are enumerated. Most of the works are related with Question Answering and Geographical Information Retrieval. There are also some works related to Multilingual Summarization, Named Entity Extraction and Word Sense Disambiguation.

10.1 Geographical Question Answering

- 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. ISBN: 2-9524532-4-1.* Trento, Italy, April 2006.

- Jordi Luque, Daniel Ferrés, Javier Hernando, José B. Mariño and Horacio Rodríguez.
  GeoVAQA: A Voice Activated Geographical Question Answering System. *In Actas de las IV Jornadas en Tecnología del Habla (4JTH).* Zaragoza, Spain, November 2006.

10.2 Geographical Information Retrieval

- Daniel Ferrés and Horacio Rodríguez.
  TALP at GeoCLEF-2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus.
  In Working Notes for the CLEF 2006 Workshop. ISBN: 2-912335-23-x, 20-22 September, Alicante.

- 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.
  6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers Series: Lecture Notes in Computer Science , Vol. 4022

10.3 Open Domain Question Answering

- Daniel Ferrés, Samir Kanaan, Alicia Ageno, Edgar Gonzàlez, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo.
  The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. *In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, CLEF, volume 3491 of Lecture Notes in Computer Science*, pages 557–568. Springer, 2004.

- Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, and Jordi Turmo.
  The TALP-QA System for Spanish at CLEF-2005.
  6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers Series: Lecture Notes in Computer Science , Vol. 4022.

- Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo.
  TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints.
  In Proceedings of the Text Retrieval Conference (TREC-2004), Gaithersburg, MD, USA, November 2005.

- Daniel Ferrés, Samir Kanaan, David Domínguez-Sal, Edgar Gonzàlez, Alicia Ageno, Maria Fuentes, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo.
  TALP-UPC at TREC 2005: Experiments Using Voting Scheme Among Three Hetereogeneous QA Systems.
  In Proceedings of the Fourteenth TREC Conference (TREC 2005)., Gaithersburg, MD, USA, November 2005.

- Marc Massot, Horacio Rodríguez, and Daniel Ferrés.
  QA UdG-UPC System at TREC-12.
  Proceedings of the Text Retrieval Conference (TREC-2003), pages 762–771, 2003.

10.4 Multidocument Summarization

- Maria Fuentes, Edgar Gonzàlez, Daniel Ferrés, and Horacio Rodríguez.
  QASUM-TALP at DUC 2005 Automatically Evaluated with a Pyramid based Metric.
  In Proceedings of the Document Understanding Conference 2005 (DUC 2005). HLT-EMNLP 2005 Workshop., Vancouver, Canada, October 2005.

- Maria Fuentes, Horacio Rodríguez, Jordi Turmo, Daniel Ferrés.
  FEMsum at DUC 2006: Semantic-based approach integrated in a Flexible Eclectic Multitask Summarizer Architecture.

In Proceedings of the Document Understanding Conference 2006 (DUC 2006). HLT-NAACL 2006 Workshop., New York City, NY, USA. June 2006.

10.5 Word Sense Disambiguation

- Lluis Màrquez, Mariona Taulé, Antonia Martí, Núria Artigas, Mar García, Francis Real, and Dani Ferrés.
  Senseval-3: The spanish lexical sample task.
  In Rada Mihalcea and Phil Edmonds, editors, Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 21–24, Barcelona, Spain, July 2004. Association for Computational Linguistics.

- Lluis Màrquez, Mariona Taulé, Antonia Martí, Mar García, Francis Real, and Dani Ferrés.
  Senseval-3: The catalan lexical sample task.
  In Rada Mihalcea and Phil Edmonds, editors, Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 147–150, Barcelona, Spain, July 2004. Association for Computational Linguistics.

10.6 Named Entity Recognition and Classification

- Daniel Ferrés, Marc Massot, Muntsa Padró, Horacio Rodríguez and Jordi Turmo.
  Automatic Building Gazetteers of Co-referring Named Entities.
  Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC 2004). Lisbon, Portugal. May 2004.

- Daniel Ferrés, Marc Massot, Muntsa Padró, Horacio Rodríguez and Jordi Turmo.
  Automatic Classification of Geographical Named Entities.
  Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC 2004). Lisbon, Portugal. May 2004.

Acknowledgments

This work has been partially supported by the European Commission (CHIL, IST-2004-506909) and the Spanish Research Dept. (ALIADO, TIC2002-04447-C02). Daniel Ferrés is supported by a UPC-Recerca grant from Universitat Politècnica de Catalunya (UPC). TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

Bibliography

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., & Vilain, M. (1995). MITRE: Description of the ALEMBIC System Used for MUC-6. *In Proceedings of the 6th Message Understanding Conference* (pp. 141–155). Columbia, Maryland.

Abney, S. (1996). Part-of-Speech Tagging and Partial Parsing. *Corpus-Based Methods in Language and Speech*. Dordrecht, Germany: Kluwer Academic Publishers.

Ahn, K., Bos, J., D.Kor, Nissim, M., Webber, B., & Curran, J. (2006). Question Answering with QED at TREC 2005. *Proceedings of the Text REtrieval Conference (TREC-14)*.

Andogah, G. (2006). GIR Experimentation. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Andre, E., Bosch, G., Herzog, G., & Rist, T. (1986). Characterizing trajectories of moving objects using natural language path descriptions. *Proceedings of the 7th ECAI* (pp. 1–8). Brighton, UK.

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., Martin, D., Myers, K., & Tyson, M. (1995). SRI International FASTUS system: MUC-6 test results and analysis. *MUC6 '95: Proceedings of the 6th conference on Message understanding* (pp. 237–248). Morristown, NJ, USA: Association for Computational Linguistics.

Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Márquez, L., Martí, M., Padró, L., Placer, R., Rodríguez, H., Taulé, M., & Turmo, J. (1998). Morphosyntactic Analisys and Parsing of Unrestricted Spanish Text. *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC* (pp. 603–610). Granada, Spain.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The meaning multilingual central repository. *Proceedings of Global WordNet Conference 2004*. Brno, Czech Republic.

Axelrod, A. (2003). On building a high performance gazetteer database . *HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 63–68). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Baron, J. R., Lewis, D., & Oard, D. (2006). Trec 2006 legal track overview. *TREC Notebook*.

Bender, O., Och, F. J., & Ney, H. (2003). Maximum Entropy Models for Named Entity Recognition. *Proceedings of CoNLL-2003* (pp. 148–151). Edmonton, Canada.

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: a high-performance learning name-finder. *Proceedings of the fifth conference on Applied natural language processing* (pp. 194–201). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Bikel, D. M., Schwartz, R. L., & Weischedel, R. M. (1999). An Algorithm that Learns What's in a Name. *Machine Learning, 34*, 211–231.

Billerbeck, B., Cannane, A., Chatteraj, A., Lester, N., Webber, W., Williams, H. E., Yiannis, J., & Zobel, J. (2004). RMIT University at TREC 2004. *Proceedings Text Retrieval Conference (TREC)*. Gaithersburg, MD: National Institute of Standards and Technology Special Publication 500-261.

Bischoff, K., Mandl, T., & Womser-Hacker, C. (2006). Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Black, W., Rinaldi, F., & Mowatt, D. (1998). FACILE: Description of the NE System Used for MUC-7. *Proceedings of the MUC-7 Conference*.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H. S., & Winograd, T. (1977). Gus, a frame-driven dialog system. *Artif. Intell., 8*, 155–173.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). Description of the Mene named entity system as used in MUC-7. *Proceedings of the MUC-7 Conference*. NYU.

Bos, J. (2006). The La Sapienza Question Answering System at TREC 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The TIGER Treebank. *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. *Proceedings of the 6th Applied NLP Conference, ANLP-2000*. Seattle, WA, United States.

Breck, E., Burger, J., Ferro, L., House, D., Light, M., & Mani, I. (1999). A Sys Called Qanda. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* (pp. 499–506). NIST Special Publication, Gaithersburg, Md., November 1999.

Breck, E., Light, M., Mann, G. S., Riloff, E., Brown, B., Anand, P., Rooth, M., & Thelen, M. (2001). Looking Under the Hood: Tools for Diagnosing your Question Answering Engine. *Working Notes of the ACL Workshop on Question Answering Systems*.

Brill, E. (1992). A simple rule-based part-of-speech tagger. *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing* (pp. 152–155). Trento, IT.

Briscoe, T., Carroll, J., & Watson, R. (2006). The Second Release of the RASP System. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (pp. 77–80). Sydney, Australia: Association for Computational Linguistics.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C.-Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., & Weishedel, R. (2000). Issues, Tasks, and Program Structures to Roadmap Research in Question & Answering (Q&A). .

Buscaldi, D., Rosso, P., & Arnal, E. S. (2005). Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* (pp. 939–946). Berlin: Springer.

Buscaldi, D., Rosso, P., & Sanchis, E. (2006). WordNet-based Index Terms Expansion for Geographical Information Retrieval. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Callan, J. P., Croft, W. B., & Harding, S. M. (1992). The INQUERY Retrieval System. *DEXA* (pp. 78–83).

Carbonell, J., Harman, D., Hovy, E., Maiorano, S., Prange, J., & Sparck-Jones, K. (2000). Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization. *NIST draft paper*.

Cardoso, N., Martins, B., Chaves, M., Andrade, L., & Silva, M. J. (2005). The XLDB Group at GeoCLEF 2005. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* (pp. 997–1006). Berlin: Springer.

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC*. Lisbon, Portugal.

Carreras, X., Màrquez, L., & Padró, L. (2002). Named Entity Extraction using AdaBoost. *Proceedings of CoNLL-2002* (pp. 167–170). Taipei, Taiwan.

Carreras, X., Màrquez, L., & Padró, L. (2003a). A Simple Named Entity Extractor using AdaBoost. *Proceedings of CoNLL-2003* (pp. 152–155). Edmonton, Canada.

Carreras, X., Màrquez, L., & Padró, L. (2003b). Learning a Perceptron-Based Named Entity Chunker via Online Recognition Feedback. *Proceedings of CoNLL-2003* (pp. 156–159). Edmonton, Canada.

Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., & Vidal, D. (2006). Priberam's Question Answering System in a Cross-Language Environment. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Chieu, H. L., & Ng, H. T. (2002). Named Entity Recognition: A Maximum Entropy Approach Using Global Information. *COLING*.

Chieu, H. L., & Ng, H. T. (2003). Named Entity Recognition with a Maximum Entropy Approach. *Proceedings of CoNLL-2003* (pp. 160–163). Edmonton, Canada.

Chinchor, N., & Robinson, P. (1997). MUC-7 Named Entity Task Definition (Version 3.5). *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Chung, H., Song, Y., Han, K., Yoon, D., Lee, J., Rim, H., & Kim, S. (2004). A Practical QA System in Restricted Domains. *Proceedings of the Workshop Question Answering in Restricted Domains, within ACL-2004*.

Collins, M. (1999). *Head-driven statistical models for natural language parsing.* Doctoral dissertation, University of Pennsylvania.

Collins, M. (2002). Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (pp. 1–8). Morristown, NJ, USA: Association for Computational Linguistics.

Curran, J. R., & Clark, S. (2003). Language Independent NER using a Maximum Entropy Tagger. *Proceedings of CoNLL-2003* (pp. 164–167). Edmonton, Canada.

Diaz, J., Rubio, A., Peinado, A., Segarra, E., Prieto, N., & Casacuberta, F. (1998). Development of Task-Oriented Spanish Speech Corpora. *Procceedings of the First International Conference on Language Resources and Evaluation* (pp. 497–501). Granada, Spain.

Downey, L. L., & Tice, D. M. (1999). A usability case study using TREC and ZPRISE. *Inf. Process. Manage.*, *35*, 589–603.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. pub-MIT.

Ferrés, D., Kanaan, S., Ageno, A., González, E., Rodríguez, H., Surdeanu, M., & Turmo, J. (2004). The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. In (Peters et al., 2005), 557–568.

Ferrés, D., Kanaan, S., González, E., Ageno, A., Rodríguez, H., Surdeanu, M., & Turmo, J. (2005). TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints. *Proceedings of the Text Retrieval Conference (TREC-2004)*.

Ferrés, D., Massot, M., Padró, M., Rodríguez, H., & Turmo, J. (2004a). Automatic Building Gazetteers of Co-referring Named Entities. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC*. Lisbon, Portugal.

Ferrés, D., Massot, M., Padró, M., Rodríguez, H., & Turmo, J. (2004b). Automatic Classification of Geographical Named Entities. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC*. Lisbon, Portugal.

Ferrández, A., & Peral, J. (2000). A computational approach to zero-pronouns in Spanish. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, (ACL'2000)*.

Ferrández, O., Kozareva, Z., Toral, A., Noguera, E., Montoyo, A., Muñoz, R., & Llopis, F. (2005). University of Alicante at GeoCLEF 2005. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. (pp. 924–927). Berlin: Springer.

Ferrés, D., Ageno, A., & Rodríguez, H. (2005a). The GeoTALP-IR System at GeoCLEF 2005: Experiments Using a QA-Based IR System, Linguistic Analysis, and a Geographical Thesaurus. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. (pp. 947–955). Berlin: Springer.

Ferrés, D., Kanaan, S., Domínguez-Sal, D., Gonzàlez, E., Ageno, A., Fuentes, M., Rodríguez, H., Surdeanu, M., & Turmo, J. (2005b). TALP-UPC at TREC 2005: Experiments Using Voting Scheme Among Three Hetereogeneous QA Systems. *Proceedings of the Fourteenth TREC Conference (TREC 2005).*. Gaithersburg, MD, USA.

Ferrés, D., Kanaan, S., González, E., Ageno, A., Rodríguez, H., & Turmo., J. (2005c). The TALP-QA System for Spanish at CLEF-2005. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum (CLEF-2005), Revised Selected Papers* (pp. 947–955). Vienna, Austria.

Ferrés, D., & Rodríguez, H. (2006a). Experiments Adapting an Open-Domain Question Answering System to the Geographical Domain Using Scope-Based Resources. *Proceedings of the Multilingual Question Answering Workshop of the EACL 2006*. Trento, Italy.

Ferrés, D., & Rodríguez, H. (2006b). TALP at GeoCLEF-2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Fleischman, M. (2001). Automated Subcategorization of Named Entities. *39th Annual Meeting of the Association for Computational Linguistics, Student Research Workshop. ACL (Companion Volume)*. (pp. 25–30). Toulouse, France.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). Named Entity Recognition through Classifier Combination. *Proceedings of CoNLL-2003* (pp. 168–171). Edmonton, Canada.

Fonseca, F., Egenhofer, M., Agouris, P., & Camara, G. (2002). Using ontologies for integrated geographic information systems. *Transactions in Geographic Information Systems, 6*.

Frew, J., Freeston, M., Freitas, N., Hill, L. L., Janee, G., Lovette, K., Nideffer, R., Smith, T. R., & Zheng, Q. (1998). The Alexandria Digital Library Architecture. *ECDL* (pp. 61–73). Springer.

G. R. Krupka, K. H. (1998). IsoQuest Inc.: Description of the NetOwlTM Extractor System as Used for MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., & Wilks, Y. (1995). University of sheffield: Description of the lasie system as used for muc. *MUC6 '95: Proceedings of the 6th conference on Message understanding*.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). One sense per discourse. *HLT '91: Proceedings of the Workshop on Speech and Natural Language* (pp. 233–237). Morristown, NJ, USA: Association for Computational Linguistics.

Garbin, E., & Mani, I. (2005). Disambiguating Toponyms in News. *HLT/EMNLP*. The Association for Computational Linguistics.

García-Vega, M., García-Cumbreras, M. A., Ureña-López, L. A., & Perea-Ortega, J. M. (2006a). SINAI at GeoCLEF 2006: Expanding the Topics with Geographical Information and Thesaurus. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

García-Vega, M., García-Cumbreras, M. A., Ureña-López, L. A., Perea-Ortega, J. M., Ariza-López, F. J., Ferrández, O., Toral, A., Kozareva, Z., Noguera, E., Montoyo, A., Muñoz, R., Buscaldi, D., & Rosso, P. (2006b). R2D2 at GeoCLEF 2006: a Mixed Approach. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Nunzio, G. M. D., & Ferro, N. (2006a). GeoCLEF 2006: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. *CLEF*. Alicante, Spain.

Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Nunzio, G. M. D., & Ferro, N. (2006b). GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., & Petras, V. (2005). GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In (Peters et al., 2006), 908–919.

Giménez, J., & Màrquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of the 4th LREC*.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve Text Retrieval. *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP* (pp. 38–44). Montreal, Canada.

Graesser, A. C., Person, N., & Huber, J. (1992). *Mechanisms that Generate Questions*, chapter 9, 167–187. Lawrence Erlbaum Associates.

Green, B., Wolf, A., Chomksy, C., & Laughery, K. (1963). Baseball: An automatic question answerer. *Computers and Thought*, 207–216.

Greenwood, M. A. (2004). Using Pertainyms to Improve Passage Retrieval for Questions Requesting Information About a Location. *Proceedings of the Workshop on Information Retrieval for Question Answering (SIGIR 2004)*.

Guillén, R. (2005). CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* (pp. 987–996). Berlin: Springer.

Guillén, R. (2006). Monolingual and Bilingual Experiments in GeoCLEF2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Guttman, A. (1984). R-Trees: A Dynamic Index Structure for Spatial Searching. *SIGMOD Conference* (pp. 47–57). ACM Press.

Harabagiu, S., Hickl, A., Williams, J., Bensley, J., Roberts, K., Shi, Y., & Rink, B. (2006). Question Answering with LCC's CHAUCER at TREC 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A., & Wang, P. (2005). Employing Two Question Answering Systems in TREC 2005. *Proceedings of the Text Retrieval Conference (TREC-2005)*.

Harabagiu, S. M., Moldovan, D. I., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R. C., Girju, R., Rus, V., & Morarescu, P. (2000). FALCON: Boosting Knowledge for Answer Engines. *TREC*.

Hauff, C., Trieschnigg, D., & Rode, H. (2006). University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Hendrickx, I., & van den Bosch, A. (2003). Memory-based one-step named-entity recognition: Effects of seed list features, classifier stacking, and unannotated data. *Proceedings of CoNLL-2003* (pp. 176–179). Edmonton, Canada.

Hendrix, G. G. (1977). Lifer: A Natural Language Interface Facility. *Berkeley Workshop* (pp. 196–).

Herrera, J., Peñas, A., & Verdejo, F. (2004). Question Answering Pilot Task at CLEF 2004. In (Peters et al., 2005), 557–568.

Hersh, W., Cohen, A. M., Roberts, P., & Rekapalli, H. K. (2006). Trec 2006 genomics track overview. *TREC Notebook*.

Herzog, O., & Rollinger, C.-R. (Eds.). (1991). *Text understanding in lilog, integrating computational linguistics and artificial intelligence, final report on the ibm germany lilog-project*, vol. 546 of *Lecture Notes in Computer Science*. Springer.

Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries* (pp. 280–290). London, UK: Springer-Verlag.

Hirschman, L., Light, M., Breck, E., & Burger, J. D. (1999). Deep Read: a reading comprehension system. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 325–332). Morristown, NJ, USA: Association for Computational Linguistics.

Hovy, E. H., Gerber, L., Hermjakob, U., Junk, M., & Lin, C.-Y. (2000). Question Answering in Webclopedia. *TREC*.

Hu, Y.-H., & Ge, L. (2006). UNSW at GeoCLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., & Wilks, Y. (1998). University of sheffield: Description of the lasie-ii system as used for muc-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Ittycheriah, A., Franz, M., & Roukos, S. (2001). IBM's Statistical Question Answering System - TREC-10. *TREC*.

Jacquemin, C., Klavans, J. L., & Tzoukermann, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. *Proceedings of the 35th annual meeting on Association for Computational Linguistics* (pp. 24–31). Morristown, NJ, USA: Association for Computational Linguistics.

Jijkoun, V., Mishne, G., de Rijke, M., Schlobach, S., Ahn, D., & Müller, K. (2004). The University of Amsterdam at QA@CLEF 2004. *Results of the CLEF 2004 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2004 Workshop* (pp. 321–324). Bath, England.

Jones, C., & Purves, R. (2005). GIR '05: Proceedings of the 2005 Workshop on Geographic Information Retrieval.

Jones, C., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., & Weibel, R. (2002). Spatial information retrieval and geographical ontologies – an overview of the spirit project. *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002).*.

Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., & Vaid, S. (2004). The spirit spatial search engine: Architecture, ontologies and spatial indexing. *Proceedings of the Geographic Information Science: Third International Conference, GIScience 2004*. Springer Berlin / Heidelberg.

Jorg Tiedemann (2004). A comparison of o -the-shelf IR engines for question answering. *Poster presentation at CLIN 2004*. Leiden, The Netherlands.

Jousse, F., Tellier, I., Tommasi, M., & Marty, P. (2005). Learning to Extract Answers in Question Answering: Experimental Studies. *Actes de la 2ème Conférence en Recherche d'Information et Applications (CORIA'05)* (pp. p.85–100). Grenoble, France: HERMES.

Juárez-Gonzalez, A., Téllez-Valero, A., Denicia-Carral, C., y Gómez, M. M., & Villaseñor-Pineda, L. (2006). INAOE at CLEF 2006: Experiments in Spanish Question Answering. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Kaisser, M., Scheible, S., & Webber, B. (2006). Experiments at the University of Edinburgh for the TREC 2006 QA Track (draft). *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Katz, B., Lin, J. J., & Felshin, S. (2002). The START Multimedia Information System: Current Technology and Future Directions. *Multimedia Information Systems* (pp. 117–123). Arizona State University.

Klein, D., & Manning, C. D. (2003a). Accurate Unlexicalized Parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Klein, D., & Manning, C. D. (2003b). Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.

Klein, D., Smarr, J., Nguyen, H., & Manning, C. D. (2003). Named Entity Recognition with Character-Level Models. *Proceedings of CoNLL-2003* (pp. 180–183). Edmonton, Canada.

Kornai, A. (2005). Evaluating Geographic Information Retrieval. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* (pp. 928–938). Berlin: Springer.

Kosseim, L., Beaudoin, A., Keighbadi, A., & Razmara, M. (2006). Concordia University at the TREC 15 QA Track. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Krupka, G. R. (1995). SRA: description of the SRA system as used for MUC-6. *MUC6 '95: Proceedings of the 6th conference on Message understanding* (pp. 221–235). Morristown, NJ, USA: Association for Computational Linguistics.

Kupiec, J. (1993). MURAX: A Robust Linguistic Approach for Question Answering Using an On-Line Encyclopedia. *Research and Development in Information Retrieval* (pp. 181–190).

Lana-Serrano, S., Goñi-Menoyo, J. M., & González-Cristóbal, J. C. (2005). MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.* (pp. 920–923). Berlin: Springer.

Lana-Serrano, S., Goñi-Menoyo, J. M., & González-Cristóbal, J. C. (2006). Report of MIRACLE Team for Geographical IR in CLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Laurent, D., Séguéla, P., & Nègre, S. (2006). Cross Lingual Question Answering using QRISTAL for CLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Lehnert, W. G. (1978). *The Process of Question Answering*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Leidner, J. L. (2004). Towards a Reference Corpus for Automatic Toponym Resolution Evaluation. *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR 2004)*. Sheffield, UK.

Leidner, J. L. (2005). Experiments with Geo-Filtering Predicates for IR. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers*. (pp. 987–996). Berlin: Springer.

Leidner, J. L. (2006). *Toponym Resolution: A First Large-Scale Comparative Evaluation*Research Report EDI–INF–RR–0839). School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK.

Leidner, J. L., Bos, J., Dalmas, T., Curran, J. R., Clark, S., Bannard, C. J., Steedman, M., & Webber, B. (2004). The QED Open-Domain Answer Retrieval System for TREC 2003. *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)* (pp. 595–599). Gaithersburg, MD.

Leidner, J. L., Sinclair, G., & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering . *HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 31–38). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Lenat, D. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*.

Leveling, J., Hartrumpf, S., & Veiel, D. (2005). Using Semantic Networks for Geographic Information Retrieval. In (Peters et al., 2006), 977–986.

Leveling, J., & Veiel, D. (2006). University of Hagen at GeoCLEF2006: Experiments with metonymy recognition in documents. *CLEF*. Alicante, Spain.

Li, H., Srihari, K. R., Niu, C., & Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction . *HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 39–44). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Li, H., Srihari, R. K., Niu, C., & Li, W. (2002). Location Normalization for Information Extraction. *COLING*.

Li, X., & Roth, D. (2002). Learning Question Classifiers. *Proceedings of the 19th International Conference on Computational Linguistics, 2002*.

Li, X., & Roth, D. (2004). Learning Question Classifiers: The Role of Semantic Information. *Natural Language Engineering, 1*.

Li, Y., Bontcheva, K., & Cunningham, H. (2005). SVM Based Learning System For Information Extraction. *Proceedings of Sheffield Machine Learning Workshop*. Springer Verlag.

Li, Z., Wang, C., Xie, X., & Ma, W.-Y. (2006). MSRA Columbus at GeoCLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Lin, C.-Y., & Hovy, E. (2000). The Automated Acquisition of Topic Signatures for Text Summarization. *COLING* (pp. 495–501). Morgan Kaufmann.

Lin, D. (1998). Dependency-Based Evaluation of MINIPAR. *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*. Granada, Spain.

Lin, J., & Katz, B. (2003). Question answering from the web using knowledge annotation and knowledge mining techniques. *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management* (pp. 116–123). New York, NY, USA: ACM Press.

Liu, S., Liu, F., Yu, C., & Meng, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 266–272). New York, NY, USA: ACM Press.

Lo, K. K., & Lam, W. (2006). Using Semantic Relations with World Knowledge for Question Answering. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Luque, J., Ferrés, D., Hernando, J., Mariño, J. B., & Rodríguez, H. (2006). GeoVAQA: A Voice Activated Geographical Question Answering System. *Actas de las IV Jornadas en Tecnología del Habla (4JTH)*.. Zaragoza, Spain.

Magnini, B., & Cavagliá, G. (2000). Integrating subject field codes into WordNet. *Proceedings of LREC-2000* (pp. 1413–1418). Athens, Greece.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K. I., & Sutcliffe, R. F. E. (2004). Overview of the CLEF 2004 Multilingual Question Answering Track. In (Peters et al., 2005), 371–391.

Manov, D., Kiryakov, A., Popov, B., Bontcheva, K., Maynard, D., & Cunningham, H. (2003). Experiments with geographic knowledge for information extraction . *HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 1–9). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

Martins, B., Cardoso, N., Chaves, M. S., Andrade, L., & Silva, M. J. (2006). The University of Lisbon at GeoCLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Martins, B., & Silva, M. J. (2005). A Graph-Ranking Algorithm for Geo-Referencing Documents. *ICDM* (pp. 741–744). IEEE Computer Society.

Martins, B., Silva, M. J., & Andrade, L. (2005). Indexing and ranking in Geo-IR systems. *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval* (pp. 31–34). New York, NY, USA: ACM Press.

Massot, M., Rodríguez, H., & Ferrés, D. (2003). QA UdG-UPC System at TREC-12. *Proceedings of the Text Retrieval Conference (TREC-2003)* (pp. 762–771).

Mayfield, J., McNamee, P., & Piatko, C. (2003). Named Entity Recognition using Hundreds of Thousands of Features. *Proceedings of CoNLL-2003* (pp. 184–187). Edmonton, Canada.

McCallum, A., & Li, W. (2003). Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of CoNLL-2003* (pp. 188–191). Edmonton, Canada.

Merchant, R., & Okurowski, M. E. (1996). The Multilingual Entity Task (MET) Overview. *Proceedings of TIPSTER Text Program (Phase II)*.

Michele Banko and Eric Brill and Susan Dumais and Jimmy Lin (2002). AskMSR: Question answering using the Worldwide Web. *Proceedings EMNLP 2002*..

Mihalcea, R., & Moldovan, D. (2000). Semantic indexing using WordNet senses. *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval* (pp. 35–45). Morristown, NJ, USA: Association for Computational Linguistics.

Mikheev, A., Grover, C., & Moens, M. (1998). Description of the LTG system used for MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, Virginia.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *Int J Lexicography*, *3*, 235–244.

Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., & the Annotation Group (1998). Algorithms that learn to extract information–BBN: Description of the SIFT system as used for MUC. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Moldovan, D., Bowden, M., & Tatu, M. (2006). A Temporally-Enhanced PowerAnswer in TREC 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Moldovan, D., Clark, C., Harabagiu, S., & Maiorano, S. (2003). COGEX: a logic prover for question answering. *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 87–93). Morristown, NJ, USA: Association for Computational Linguistics.

Moldovan, D., Harabagiu, S., Clark, C., Bowden, M., Lehmann, J., & Williams, J. (2004). Experiments and Analysis of LCC's two QA Systems Over TREC 2004. *TREC 2004 Conference Note Book* (pp. 21–30).

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Gîrju, R., & Rus, V. (1999). LASSO: A tool for surfing the answer net. *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*.

Molla, D., & Vicedo, J. (2005). *AAAI-05 Workshop on Question Answering in Restricted Domains*. AAAI Press. to appear.

Monz, C. (2003). *From document retrieval to question answering.* Doctoral dissertation, University of Amsterdam.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*..

Nyberg, E., Mitamura, T., Callan, J. P., Carbonell, J. G., Frederking, R. E., Collins-Thompson, K., Hiyakumoto, L., Huang, Y., Huttenhower, C., Judy, S., Ko, J., Kupsc, A., Lita, L. V., Pedro, V., Svoboda, D., & Durme, B. V. (2003). The JAVELIN Question-Answering System at TREC 2003: A Multi-Strategh Approach with Dynamic Planning. *Proceedings of the TREC-12*.

Ogilvie, P., & Callan, J. P. (2001). Experiments Using the Lemur Toolkit. *Proceedings of the Text REtrieval Conference (TREC-10)*.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*. Seattle, Washington, USA.

Overell, S., Magalhães, J., & Rüger, S. (2006). Place disambiguation with co-occurrence models. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Padró, L. (1996). Pos tagging using relaxation labelling. *COLING* (pp. 877–882).

Palomar, M., Civit, M., Díaz, A., Moreno, L., Bisbal, E., Aranzabe, M., Ageno, A., Martí, M. A., & Navarro, B. (2004). 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. *Procesamiento del Lenguaje Natural*, *33*.

Pasca, M., & Harabagiu, S. M. (2001a). Answer mining from on-line documents. *Proceedings of the workshop on ARABIC language processing* (pp. 1–8). Morristown, NJ, USA: Association for Computational Linguistics.

Pasca, M., & Harabagiu, S. M. (2001b). High Performance Question/Answering. *Research and Development in Information Retrieval* (pp. 366–374).

Pascual, F. L. (2002). *Ir-n un sistema de recuperación de información basado en pasajes.* Doctoral dissertation, Universidad de Alicante.

Peters, C., & Braschler, M. (2001). European research letter: Cross-language system evaluation: The CLEF campaigns. *Journal of the American Society for Information Science and Technology (JASIST), 52*, 1067–1072.

Peters, C., Clough, P., Gonzalo, J., Jones, G. J. F., Kluck, M., & Magnini, B. (Eds.). (2005). *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*, vol. 3491 of *Lecture Notes in Computer Science*. Springer.

Peters, C., Gey, F. C., Gonzalo, J., J.F.Jones, G., Kluck, M., Magnini, B., Müller, H., & de Rijke, M. (Eds.). (2006). *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.*, vol. 4022 of *Lecture Notes in Computer Science*. Berlin: Springer.

Porter, M. F. (1997). *An algorithm for suffix stripping*, 313–316. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J., & Jurafsky, D. (2004). Shallow Semantic Parsing Using Support Vector Machines. *Proceedings of the Human Language Technology Conference/North American chapter of the Association of Computational Linguistics (HLT/NAACL)*. Boston, MA.

Prager, J., Brown, E., Coden, A., & Radev, D. (2000). Question-answering by predictive annotation. *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 184–191). New York, NY, USA: ACM Press.

Purves, R., & Jones, C. (2004). Workshop on geographic information retrieval at SIGIR 2004. *SIGIR Forum, 38*, 53–56.

Quinlan, J. R. (1990). Learning Logical Definitions from Relations. *Machine Learning, 5*, 239–266.

Quinlan, J. R., & Cameron-Jones, R. M. (1993). FOIL: A Midterm Report. *ECML* (pp. 3–20). Springer.

Radev, D., Fan, W., Qi, H., Wu, H., & Grewal, A. (2002). Probabilistic question answering on the web. *WWW '02: Proceedings of the 11th international conference on World Wide Web* (pp. 408–419). New York, NY, USA: ACM Press.

Rauch, E., Bukatin, M., & Baker, K. (2003). A confidence-based framework for disambiguating geographic terms . *HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 50–54). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Ravichandran, D., & Hovy, E. H. (2002). Learning surface text patterns for a Question Answering System. *ACL* (pp. 41–47).

Rees, T. (2003). C-squares, a new spatial indexing system and its applicability to the description of oceanographic data. *Ocenography*, *16*, 11–19.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann.

Roberts, I., & Gaizauskas, R. J. (2004). Evaluating Passage Retrieval Approaches for Question Answering. *ECIR* (pp. 72–84). Springer.

Robertson, S. E., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *SIGIR* (pp. 232–241). ACM/Springer.

Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertanga, F., & Roventini, A. (1998). The Top-Down Strategy for Bulding EuroWordNet: Vocabulary Coverage, Base Conceps and Top Ontology. *Computer and Humanities 32*. Kluwer Academic Publishers.

Ruiz, M. E., Shapiro, S., Abbas, J., Southwick, S. B., & Mark, D. (2006). UB at GeoCLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Sacaleanu, B., & Neumann, G. (2006). DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Saiz, M. (2002). *Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español*. Doctoral dissertation, Universidad de Alicante.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, *24*, 513–523.

Salton, G., & Lesk, M. E. (1965). The SMART automatic document retrieval systems an illustration. *Communications of the ACM*, *8*, 391–398.

Sanderson, M., & Kohler, J. (2004). Analyzing Geographic Queries. *Proceedings of the Geographic Information Retrieval Workshop of the SIGIR 2004*.

Schlaefer, N., Gieselman, P., & Sautter, G. (2006). The Ephyra QA System at TREC 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Schone, P., Ciany, G., Cutts, R., McNamee, P., Mayfield, J., & Smith., T. (2006). QACTIS Enhancements in TREC QA 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Sekine, S. (1998). Nyu: Description Of The Japanese Ne System Used For Met-2. *Proceedings of the MUC-7 Conference*.

Sekine, S., & Eriguchi, Y. (2000). Japanese named entity extraction evaluation: analysis of results. *Proceedings of the 18th conference on Computational linguistics* (pp. 1106–1110). Morristown, NJ, USA: Association for Computational Linguistics.

Sekine, S., Sudo, K., & Nobata, C. (2002). Extended Named Entity Hierarchy. *Proceedings of Thirth International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

Shen, D., Leidner, J. L., Merkel, A., & Klakow, D. (2006). The Alyssa System at TREC 2006: A Statistically-Inspired Question Answering System. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Simmons, R. F. (1965). Answering English questions by computer: a survey. *Commun. ACM, 8*, 53–70.

Smith, D. A., & Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. *ECDL* (pp. 127–136). Springer.

Solorio, T., Pérez-Coutiño, M., y Gómez, M. M., Villaseñor-Pineda, L., & López-López, A. (2004). A Language Independent Method for Question Classification. *COLING-2004* (pp. 1374–1380).

Soriano, J. M. G., y Gómez, M. M., Arnal, E. S., & Rosso, P. (2005). A Passage Retrieval System for Multilingual Question Answering. *TSD* (pp. 443–450). Springer.

Soubbotin, M. M. (2001). Patterns of Potential Answer Expressions as Clues to the Right Answers. *Proceedings of the Text Retrieval Conference TREC-2001*.

Strzalkowski, T., Small, S., Hardy, H., Yamrom, B., Liu, T., Kantor, P., Ng, K., & Wacholder., N. (2005). HITIQA: A Question Answering Analytical Tool. *Proceedings of International Conf. On Intelligence Analysis*. McLean, VA.

Sundheim, B. (1995a). MUC6 named entity task definition, Version 2.1. *Proceedings of the Sixth Message Understanding Conference (MUC6)*. Columbia, MD, USA: Morgan Kaufmann.

Sundheim, B. (1995b). Overview of results of the MUC-6 evaluation. *Proceedings of the Sixth Message Understanding Conference (MUC6)*. Columbia, MD, USA: Morgan Kaufmann.

Surdeanu, M., Turmo, J., & Comelles, E. (2005). Named Entity Recognition from Spontaneous Open-Domain Speech. *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*. Lisbon, Portugal.

Suzuki, J., Taira, H., Sasaki, Y., & Maeda, E. (2003). Question classification using HDAG kernel. *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering* (pp. 61–68). Morristown, NJ, USA: Association for Computational Linguistics.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2002* (pp. 155–158). Taipei, Taiwan.

Tjong Kim Sang, E. F., & Buchholz, S. (2000). Introduction to the CoNLL-2000 Shared Task: Chunking. *Proceedings of CoNLL-2000 and LLL-2000* (pp. 127–132). Lisbon, Portugal.

Tjong Kim Sang, E. F., & De Meulder, F. (2003a). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2003* (pp. 142–147). Edmonton, Canada.

Tjong Kim Sang, E. F., & De Meulder, F. (2003b). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2003* (pp. 142–147). Edmonton, Canada.

Toral, A., Ferrández, O., Noguera, E., Kozareva, Z., Montoyo, A., & Muñoz, R. (2006). Geographic IR Helped by Structured Geospatial Knowledge Resources. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics.

Vallin, A., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., & Sutcliffe, R. (2005). Overview of the CLEF 2005 Multilingual Question Answering Track. In (Peters et al., 2006).

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Vicedo, J. L. (2002). *Semqa: Un modelo semántico aplicado a los sistemas de búsqueda de la respuesta*. Doctoral dissertation, Universidad de Alicante.

Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. *Proceedings of the Text Retrieval Conference (TREC-8)*.

Voorhees, E. M. (2001). Overview of the TREC 2001 Question Answering Track. *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Voorhees, E. M. (2002). Overview of the TREC 2002 Question Answering Track. *Proceedings of the Tenth Text REtrieval Conference (TREC 2002)*.

Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. *Proceedings of the Text Retrieval Conference (TREC-2003)* (pp. 54–68).

Voorhees, E. M. (2004). Overview of the TREC 2004 Question Answering Track. *Proceedings of the Text Retrieval Conference (TREC-2004)*.

Voorhees, E. M., & Tice, D. M. (1999). The TREC-8 Question Answering Track Evaluation. *TREC*.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. *Proceedings of the DELOS workshop on Cross-language Information Retrieval*. Zurich, Switzerland.

Voutilainen, A. (1997). Designing a (finite-state) parsing grammar. In E. Roche and Y. Schabes (Eds.), *Finite-state language processing*, 283–310. MIT Press, Cambridge.

Voutilainen, A., & Padró, L. (1997). Developing a hybrid NP parser. *Proceedings of the fifth conference on Applied natural language processing* (pp. 80–87). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Warren, D. H. D., & Pereira, F. C. N. (1982). An Efficient Easily Adaptable System for Interpreting Natural Language Queries. *American Journal of Computational Linguistics*, *8*, 110–122.

Weischedel, R. (1995). BBN*: Description of the* PLUM *system as used for muc-6*, 55–69. Proceedings of the 6th Messsage Understanding Conference. Columbia, Maryland: Morgan Kaufmann Publishers, Inc.

Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*, 36–45.

Whitelaw, C., & Patrick, J. (2003). Named Entity Recognition Using a Character-based Probabilistic Approach. *Proceedings of CoNLL-2003* (pp. 196–199). Edmonton, Canada.

Whittaker, E., Novak, J., Chatain, P., & Furui, S. (2006). TREC 2006 Question Answering Experiments at Tokyo Institute of Technology (Draft). *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Wilensky, R., Chin, D. N., Luria, M., Martin, J. H., Mayfield, J., & Wu, D. (2000). The Berkeley UNIX Consultant Project. *Artificial Intelligence Review*, *14*, 43–88.

Winograd, T. (1972). *Understanding Natural Language*. Orlando, FL, USA: Academic Press, Inc.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images, second edition*. Morgan Kaufmann.

Woods, W. (1977). Lunar rocks in natural English: Explorations in natural language question answering. *Linguistic Structures Processing*, 521–569.

Wu, M., & Strzalkowski, T. (2006). ILQUA at TREC 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Xu, K., & Meng, H. (2005). Design and Development of a Bilingual Reading Comprehension Corpus. *International Journal of Computational Linguistics & Chinese Language Processing*. ACLCLP.

Yi Li, N. S., Cavedon, L., & Moffat, A. (2006). NICTA I2D2 Group at GeoCLEF 2006. *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*. Alicante, Spain.

Yu, S., Bai, S., & Wu, P. (1998). Description of the Kent Ridge Labs System used for MUC-7. *Proceedings of the MUC-7 Conference*.

Yutaka Sasaki, Hsin-Hsi Chen, K.-h. C., & Lin, C.-J. (2005). Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1). *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (pp. 54–68).

Zelle, J. M. (1995). *Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers*. Doctoral dissertation, Department of Computer Sciences, University of Texas, Austin, TX. Also appears as Artificial Intelligence Laboratory Technical Report AI 96-249.

Zelle, J. M., & Mooney, R. J. (1996). Learning to Parse Database Queries Using Inductive Logic Programming. *AAAI/IAAI, Vol. 2* (pp. 1050–1055).

Zhang, D., & Lee, W. S. (2003). Question classification using support vector machines. *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 26–32). New York, NY, USA: ACM Press.

Zhao, Y., Xu, Z.-M., Li, P., & Guan, Y. (2006). InsunQA06 on QA Track of TREC 2006. *"The Fifteenth Text REtrieval Conference (TREC 2006) Notebook"*. Gaithersburg, MD, USA: NIST.

Zheng, Z. (2002). AnswerBus Question Answering System. *Proceedings of the Human Language Technology Conference (HLT 2002)*. San Diego, CA.

Zhou, G., & Su, J. (2001). Named entity recognition using an HMM-based chunk tagger. *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 473–480). Morristown, NJ, USA: Association for Computational Linguistics.

Zhu, J., Uren, V. S., & Motta, E. (2005). ESpotter: Adaptive Named Entity Recognition for Web Browsing. *Wissensmanagement* (pp. 505–510). DFKI, Kaiserslautern.