Resource Integration for Question Answering and Geographical Information Retrieval

(Integración de recursos para las tareas de búsqueda de respuestas y recuperación de información geográfica)

A research project report presented
by

Davide Buscaldi

to
The Department of Information Systems and Computation
in partial fulfillment of the requirements
for the obtention of
Diploma of Advanced Studies
(Diploma de Estudios Avanzados)
in the subject of

Pattern Recognition and Artificial Intelligence

Polytechnic University of Valencia
Valencia, Spain
September 2007

Abstract

One of the first scenarios imagined by the researchers in Artificial Intelligence was the problem of conversing with a machine in natural language. Alan Turing in 1950 proposed a test in order to check the capability of a machine to demonstrate intelligence, and that test, that carries his name, is mostly based on conversation and language understanding. Obtaining responses to questions has always been the ambition of the human being. Question Answering (QA) allows the automatic answering of questions poned by a human user. Question Answering is particularly challenging because it needs the advancements of several fields, mainly Natural Language Processing (NLP), e.g. document understanding, information extraction, language generation, question analysis, word sense disambiguation and Information Retrieval (IR), e.g. document analysis, query formulation, relevance feedback. Given this multidisciplinary scenario, it is easy to realize how difficult it is to build QA systems.

The aim of this research work was to explore these research fields in order to build a functional Question Answering system that integrated various knowledge sources, among them geographical information sources.

The main achievement of the investigations carried out has been the realization of a complete Question Answering system. This system is able to answer questions in Spanish, Italian and French with an accuracy comparable with the current state-of-the-art systems. The research work over knowledge resources (particularly WordNet) resulted also in the development of novel Geographical Information Retrieval (GIR) and Word Sense Disambiguation (WSD) methods.

Contents

Citations to previously published work

Large portions of the following chapters have been published in the following papers:

P.Rosso, F.Masulli, D.Buscaldi. Word Sense Disambiguation using Conceptual Distance, Frequency and Gloss. In: Int. Conf. on Natural Language Processing and Engineering Knowledge, IEEE Press, Beijing, China, pp. 120-125, 2003.

Gomez J.M., Buscaldi D., Bisbal E., Rosso P., Sanchis E. QUASAR: The Question Answering System of the Universidad Politecnica de Valencia. In: CLEF 2005 Proceedings. Springer Verlag, LNCS(4022), Vienna, Austria, 2006.

Buscaldi D., Rosso P., Sanchis E. Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In: CLEF 2005 Proceedings. Springer Verlag, LNCS(4022), Vienna, Austria, 2006.

Buscaldi, D., Rosso, P. Mining Knowledge from Wikipedia for the Question Answering task. Proc. of the LREC 2006, Genova, Italy, 2006.

Buscaldi, D., Rosso, P. A Naive bag-of-words approach to Wikipedia QA. CLEF 2006 Working notes, Alicante 20-22 September, C.Peters Ed, 2006.

Buscaldi, D., Rosso, P., Sanchis, E. WordNet-based Index Terms Expansion for Geographical Information Retrieval. CLEF 2006 Working notes, Alicante 20-22 September, C.Peters Ed., 2006.

Buscaldi, D., Gmez, J.M., Rosso, P., Sanchis, E. The UPV at QA@CLEF 2006. CLEF 2006 Working notes, Alicante 20-22 September, C.Peters Ed., 2006.

Sanchis, E., Buscaldi, D., Grau, S., Hurtado, L., Griol, D. Spoken QA based on a Passage Retrieval Engine. SLT 2006, Aruba 10-13 December 2006.

Rosso, P., Buscaldi, D., Iskra, M. Web-based Selection of Optimal Translations of Short Queries. SEPLN, Revista no.38 (Abril 2007) pp. 49-53 ISSN: 1135-5948, 2007.

Buscaldi, D., Rosso, P. UPV-WSD : Combining different WSD Methods by means of Fuzzy Borda Voting. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) pp. 434-437, Prague, Czech Republic, 2007.

Some portions of chapters 2 and 3 will be published in the following paper:

Buscaldi, D., Rosso, P. A conceptual density-based approach for the disambiguation of toponyms. To appear in International Journal of Geographical Information Systems.

Chapter 1

Introduction

Seeking answers to questions has always been one of the characteristic features of the human being. An answer can be described as the minimal amount of information needed in order to satisfy a user's information need. Question Answering (QA) is a task consisting in the automatic answering of questions poned by a human user. Taking into the account the previous definition of an answer, we can see QA as a specialized kind of Information Retrieval (IR), where the user expresses her/his information need in natural language, instead of reducing it to a set of relevant terms. Moreover, whilst in IR a whole document is retrieved, in QA only the needed piece of information is handed over to the user. For instance, if we are interested in knowing the name of the dynasty currently reigning in England, by the IR approach we probably would write into a web search engine the query "*current dynasty England*", or, more probably, "*England royal dynasties*" with the objective of finding a web page that includes the information we are seeking. The same search in the QA approach would be carried out by submitting the exact question: "*What is the name of the dynasty currently reigning in England?*", obtaining the string "*Windsor*" as a result.

It is straightforward to realize that Question Answering needs the advancements of several fields related to Natural Language Processing (NLP), e.g. document understanding, information extraction, language generation, question analysis, word sense disambiguation and also to information retrieval, e.g. document analysis, query formulation, relevance feedback. Given this multidisciplinary scenario, it is easy to realize how difficult it is to build QA systems.

The first QA systems date back to the 1960s and were based on a closed and limited world knowledge repository [20, 60]. Nowadays, researchers are pointing towards the Web as a resource, due to its open and dynamic nature [14, 7]. Anyway, the key feature of any QA system is constituted by the knowledge resources employed in order to find the answer. Recently, some efforts have been carried out for the integration of different knowledge sources such as the WordNet [43] ontology together with the Web [61], or Wikipedia together with other sources [27], or alone [26].

The introduction in 1999 of the QA track in the US government's Text Retrieval

and Evaluation Conference (TREC)[1], organized as a competition-based system evaluation with dozens of participants, boosted the interest in QA. The CLEF[2], an European-based evaluation conference, focused on multilingual issues and European languages has also been started in 2000. The results obtained by the best QA systems are typically between 40 and 70 percent in accuracy, depending on the language and the type of exercise. These numbers indicates the difficulty of the Question Answering task. Therefore, some efforts are being conducted in order to focus only on particular types of questions (restricted domain QA), including law, genomics and the geographical domain [15] among others.

More attention is currently being payed to the geographical domain. A study by Mark Sanderson [53] revealed that more than the 18% of the queries submitted to search engines contains a geographical reference. Such interest resulted in the creation of the GeoCLEF[3] task at the CLEF and the GIR workshops [48] at CIKM and SIGIR conferences. These efforts are centered on Geographical Information Retrieval (GIR).

Geographical Information Retrieval can be viewed as a "geographically-flavoured" kind of Information Retrieval. Purves and Jones defined GIR as "*the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope*" [48]. One of the major issues encountered in GIR is the ambiguity of place names (*toponyms*). The typical solution to this issue is to apply Word Sense Disambiguation (WSD) techniques to toponyms. WSD itself is an open problem in the field of NLP. Many techniques have been developed until now but their efficiency usually does not exceed the baseline, as evidenced by the Senseval/Semeval competitions[4]. Moreover, the impact of WSD over Geographical Information Retrieval is also an object of debate.

1.1 Overview of the research report

This research report covers three major areas: Question Answering, Geographical Information Retrieval and Word Sense Disambiguation. QA can be considered the main object of the research work, whilst GIR and WSD were investigated as complementary parts of the main research direction. Particularly, WSD addresses some problems in GIR, and GIR is related to Question Answering, especially by means of the resources that can be used in both fields.

Therefore, this document is structured as follows: in Chapter 2 we describe the general architecture of QA system and the QUASAR system we realized with the objective to participate to the CLEF and TREC exercises. We also will examine

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org
[3] http://ir.shef.ac.uk/geoclef/
[4] http://www.semeval.org

the resources that are generally used in QA. In Chapter 3 the focus will be moved over GIR and the WordNet-based method we developed, together with the results obtained in our participation to the GeoCLEF competitions. Finally, in Chapter 4 we will describe a WSD method based on Conceptual Density and its application to the geographical domain, specifically with the purpose to be used in GIR systems.

Chapter 2

Question Answering and Resources

2.1 Architecture of Question Answering Systems

The basic building blocks of QA systems are usually three modules: question classification and analysis, document or passage retrieval and answer extraction. The aim of the first module is to recognize the type or category of the expected answer (e.g. if it is a Person, Quantity, Date, etc.) from the user question. The second module obtains the passages (or pieces of text) which contain the terms of the question. Finally, the answer extraction module uses the information collected by the previous modules in order to extract the correct answer.

2.1.1 Question Analysis and Classification

Question Classification (QC) is defined as the task to assign a class (chosen from a predefined hierarchy) to each question formulated to a system. Its main goal is to apply a different answer extraction strategy for each question type in the last stage, the answer extraction phase, and to restrict the candidate answers: the way to extract the answer to "What is nitre?", which is looking for a definition, is not the same as to "Who invented the radio?", which is asking for the name of a person. It is probably the most critical step of a QA system: a study which analyzes the errors in open domain question answering systems [44] revealed that more than 36% of them are directly due to the question classification module. Most QC systems use patterns and heuristic rules [23] in order to achieve a high accuracy.

Together with QC, an analysis of the question is typically carried out by these modules. The most important elements that can be extracted from the question are the *question focus* (property or object sought by the question: for instance, *colour* in "What *colour* is a mango?") and the *question topic* (object or event that the question is about: for instance, *Mt. Erebus* "What is the height of *Mt. Erebus*?"), often a Named Entity.

2.1.2 Passage Retrieval

Document or passage retrieval is typically used as the first step in current question answering systems. In most of the QA systems classical PR methods are used [41, 4, 45]. The main problems of these QA systems derive from the use of PR methods which are adaptations of classical document retrieval systems, not specifically oriented to the QA task. These methods use the question keywords to find relevant passages. For instance, if the question is "Who is the President of Mexico?", these methods return those passages which contain the words "President" and "Mexico".

In [49, 24] it is shown that standard IR engines often fail to find the answer in the documents (or passages) when presented with natural language questions. There are other PR approaches which are based on Natural Language Procesing (NLP) in order to improve the performance of the QA task [3, 21, 39]. The main disadvantage of these approaches have is that they are very difficult to adapt to other languages or to multilingual tasks.

Another strategy is to search the obviousness of the answer in the Web. They send the user question to a Web search engine with the expectations to get a passage containing the same expression of the question or a similar one. They suppose that, due to the high redundancy of the Web, the answer will be written in several different ways, including a form close to the one used in the question.

2.1.3 Answer Extraction

This is the crucial step of every QA system. In this phase an answer is determined on the basis of the retrieved passages and the constraints retrieved during the Question Analysis. Typical QA system use in this phase some kind of knowledge, that can be a customized database [20, 60], the web [37, 42], ontologies [32] or encyclopedias [29]. The retrieved passages are further refined for enhanced precision. Passages that do not satisfy the semantic constraints specified in the question are discarded.

The search for answers within the retrieved passages is restricted to those candidates corresponding to the expected answer type. If the expected answer type is a named entity, the candidates are identified with a named entity recognizer. For instance, if the expected answer type is MONEY, the identified candidates may include $1euro$ and $USD520$. Conversely, if the answer type is a DEFINITION, the candidates are usually obtained by matching a set of answer patterns on the passages.

Each candidate answer receives a relevance score according to lexical and proximity features such as distance between keywords, or the occurrence of the candidate answer within an apposition. The candidates are sorted in decreasing order of their scores.

Finally, the system selects the candidate answers with the highest relevance scores. The final answers are either fragments of text extracted from the passages around the best candidate answers, or they are internally generated.

2.2 The QUASAR Question Answering System

The architecture of QUASAR is shown in Fig.2.1.

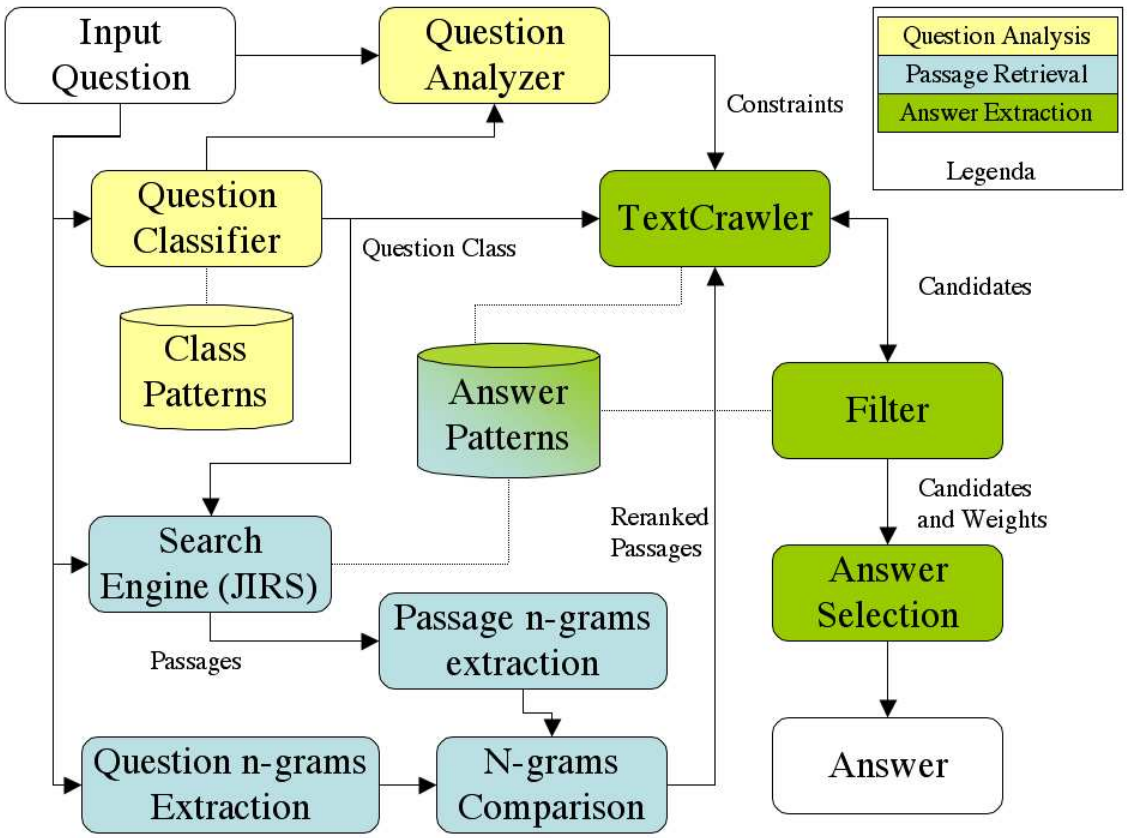

Figure 2.1: Diagram of the QA system

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

2.2.1 Question Analysis Module

This module obtains both the expected answer type (or *class*) and some constraints (i.e., focus and topic) from the question.

The different answer types that can be treated by our system are shown in Table 2.1.

Table 2.1: QC pattern classification categories.

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

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

The Question Analyzer has the purpose of identifying the constraints to be used in the AE phase. These constraints are made by sequences of words extracted from the POS-tagged query by means of POS patterns and rules. For instance, any sequence of nouns (such as *ozone hole*) is considered as a relevant pattern. The POS-taggers used were the SVMtool[1] for English and Spanish, and the TreeTagger[2] for Italian and French.

There are two classes of constraints: a *target* constraint, which is the word of the question that should appear closest to the answer string in a passage (corresponding roughly to the *focus* of the question), and zero or more *contextual* constraints, keeping the information that has to be included in the retrieved passage in order to have a

[1]http://www.lsi.upc.edu/ nlp/SVMTool/

[2]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

chance of success in extracting the correct answer (similar to the *topic* of the question). For example, in the following question: "*Dónde se celebraron los Juegos Olímpicos de Invierno de 1994?*" (*Where did the Winter Olympic games of 1994 take place?*) *celebraron* is the target constraint, while *Juegos Olímpicos de Invierno* and *1994* are the contextual constraints. There is always only one target constraint for each question, but the number of contextual constraint is not fixed. For instance, in "*Quién es Neil Armstrong?*" the target constraint is *Neil Armstrong* but there are no contextual constraints.

2.2.2 The JIRS Passage Retrieval Module

The passages containing the relevant terms are retrieved by JIRS using a classical keyword-based IR system. This year, the module was modified in order to rank better the passages which contain an answer pattern matching the question type. Therefore, this module is not as language-independent as in 2005 because it uses informations from the Question Classifier and the patterns used in the Answer Extraction phase.

Sets of unigrams, bigrams, ..., n-grams are extracted from the passages and from the user question. In both cases, n is the number of question terms. These n-gram sets are compared in order to obtain the weight of each passage, which is proportional to the size of the question n-grams found in the passage.

For instance, if the question is "*What is the capital of Croatia?*" and the system retrieves the following two passages: "*...Tudjman, the president of Croatia, met Eltsin during his visit to Moscow, the capital of Russia...*", and "*...they discussed the situation in Zagreb, the capital of Croatia...*". The second passage must have more importance because it contains the 4-gram "*the capital of Croatia*", whereas the first one contains the 3-gram "*the capital of*" and the 1-gram "*Croatia*". This example also shows the advantage of considering n-grams instead of keywords: the two passages contains the same question keywords, but only one of them contains the right answer.

In order to calculate the weight of n-grams of every passage, the greatest n-gram in the passage is identified and it is assigned a weight equal to the sum of all its term weights. Subsequently, smaller n-grams are searched. The weight of every term is determined by means of formula (2.1):

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \quad . \tag{2.1}$$

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

Sometimes a term unrelated to the question can obtain a greater weight than those assigned to the Named Entities (NEs), such as names of persons, organizations

and places, dates. The NEs are the most important terms of the question and it does not make sense to return passages which do not contain them. Therefore, NEs are given a greater weight than the other question terms, in order to force its presence in the first ranked passages. NEs are recognized by simple rules, such as capitalization, or by checking if they are a number. Once all the terms have been weighted, the sum is normalized.

JIRS can be obtained at the following URL: `http://jirs.dsic.upv.es`.

2.2.3 Answer Extraction

The input of this module is constituted by the n passages returned by the PR module and the constraints (including the expected type of the answer) obtained through the *Question Analysis* module. A *TextCrawler* is instantiated for each of the n passages with a set of patterns for the expected type of the answer and a pre-processed version of the passage text. Some patterns can be used for all languages; for instance, when looking for proper names, the pattern is the same for all languages. The pre-processing of passage text consists in separating all the punctuation characters from the words and in stripping off the annotations of the passage. It is important to keep the punctuation symbols because we observed that they usually offer important clues for the individuation of the answer: for instance, it is more frequent to observe a passage containing "*The president of Italy, Carlo Azeglio Ciampi*" than one containing "*The president of Italy IS Carlo Azeglio Ciampi*" ; moreover, movie and book titles are often put between apices.

The positions of the passages in which occur the constraints are marked before passing them to the TextCrawlers. Some spell-checking function has been added in this phase by using Levenshtein distance to compare strings. The TextCrawler begins its work by searching all the passage's substrings matching the expected answer pattern. Then a weight is assigned to each found substring s, depending on the positions of the constraints, if s does not include any of the constraint words. Let us define $w_t(s)$ and $w_c(s)$ as the weights assigned to a substring s as a function, respectively, of its distance from the target constraints (2.2) and the context constraints (2.3) in the passage.

$$w_t(s) = \max_{0<k\leq|p(t)|} close(s, p_k(t)) \tag{2.2}$$

$$w_c(s) = \frac{1}{|c|}\sum_{i=0}^{|c|} \max_{0<j\leq|p(c_i)|} near(s, p_j(c_i)) \tag{2.3}$$

Where c is the vector of contextual constraints, $p(c_i)$ is the vector of positions of the constraint c_i in the passage, t is the target constraint and $p(t)$ is the vector of

positions of the target constraint t in the passage. *Close* and *near* are two proximity function defined as:

$$close(s,p) = \exp\left(-\left(\frac{d(s,p)-1}{5}\right)^2\right) \tag{2.4}$$

$$near(s,p) = \exp\left(-\left(\frac{d(s,p)-1}{2}\right)^2\right) \tag{2.5}$$

Where p is a position in the passage and $d(s,p)$ is computed as:

$$d(s,p) = \min_{i\in\{0,|s|-1\}} \sqrt{(s_i - p)^2} \tag{2.6}$$

Where s_i indicates the position of the i-th word of the substring s. The proximity functions can roughly be seen as fuzzy membership functions, where $close(s,p)$ means that the substring s is adjacent to the word at the position p, and $near(s,p)$ means that the substring s is not far from the word at position p. The 2 and 5 values roughly indicate the range within the position p where the words are considered really "close" and "near", and have been selected after some experiments with the CLEF2003 QA Spanish test set. Finally, the weight is assigned to the substring s in the following way:

$$w(s) = \begin{cases} w_t(s) \cdot w_c(s) & \text{if } |p(t)| > 0 \wedge |c| > 0 \\ w_c(s) & \text{if } |p(t)| = 0 \wedge |c| > 0 \\ w_t(s) & \text{if } |p(t)| > 0 \wedge |c| = 0 \\ 0 & elsewhere. \end{cases} \tag{2.7}$$

This means that if in the passage have been found both the target constraint and the contextual constraints, the product of the weights obtained for every constraint will be used; otherwise, only the weight obtained for the constraints found in the passage will be used.

Usually, the type of expected answer directly affects the weighting formula. For instance, the "DEFINITION" questions (such as "Who is Jorge Amado?") usually contain only the target constraint, while "QUANTITY" questions (such as "How many inhabitants are there in Sweden?") contain both target and contextual constraints. For the other question types the target constraint is rarely found in the passage, and weight computation relies only on the contextual constraints (e.g. "From what port did the ferry Estonia leave for its last trip?", port is the target constraint but it is not mandatory in order to found the answer, since it is most common to say "The Estonia left from Tallinn", from which the reader can deduce that Tallinn is -or at least has- a port, than "Estonia left from the port of Tallinn").

The filter module takes advantage of some knowledge resources, such as a mini knowledge base or the web, in order to discard the candidate answers which do not

match with an allowed pattern or that do match with a forbidden pattern. For instance, a list of country names in the four languages has been included in the knowledge base in order to filter country names when looking for countries. When the filter rejects a candidate, the TextCrawler provide it with the next best-weighted candidate, if there is one.

Finally, when all TextCrawlers end their analysis of the text, the *Answer Selection* module selects the answer to be returned by the system. The following strategies have been developed:

- Simple voting (SV): The returned answer corresponds to the candidate that occurs most frequently as passage candidate.
- Weighted voting (WV): Each vote is multiplied for the weight assigned to the candidate by the TextCrawler and for the passage weight as returned by the PR module.
- Maximum weight (MW): The candidate with the highest weight and occurring in the best ranked passage is returned.
- Double voting (DV): As simple voting, but taking into account the second best candidates of each passage.
- Top (TOP): The candidate elected by the best weighted passage is returned.

SV is used for every "NAME" type question, while WV is used for all other types. For "NAME" questions, when two candidates obtain the same number of votes, the Answer Selection module looks at the DV answer. If there is still an ambiguity, then the WV strategy is used. For other types of question, the module use directly the MW. TOP is used only to assign the confidence score to the answer, obtained by dividing the number of strategies giving the same answer by the total number of strategies (5), multiplied for other measures depending on the number of passages returned (n_p/N, where N is the maximum number of passages that can be returned by the PR module and n_p is the number of passages actually returned) and the averaged passage weight. The weighting of NIL answers is slightly different, since is obtained as $1 - n_p/N$ if $n_p > 0$, 0 elsewhere.

In our system, candidates are compared by means of a partial string match, therefore *Boris Eltsin* and *Eltsin* are considered as two votes for the same candidate. Later, the Answer Selection module returns the answer in the form occuring most frequently.

2.2.4 Cross-Language module

Various methods have been developed recently in order to minimize the error introduced by MT in IR-related fields. In particular, the idea of combining different MT systems has already been used succesfully for the cross-lingual Ad-Hoc retrieval

task [13]. The most common form of combination of different MT systems is the selection of the best translation from a set of candidates [12, 31], although there have been also proposals for the combination of fragments from different translations [1].

In the case of the Cross-language task, the Question Analysis module needs to work with an *optimal* translation of the input query in order to obtain the best accuracy. Therefore, we opted for a selection of the best translation technique. We took into the account the following web tools: Systran[3], FreeTrans[4], Linguatec[5], Promt[6] and Reverso[7].

Given a translation X of a question q, let us define w as the sequence of n words that compose the translation:

$$w = (w_1, \ldots, w_n)$$

A *trigram chain* is, therefore, defined as the set of trigrams T:

$$T = \{(w_1, w_2, w_3), (w_2, w_3, w_4), \ldots \\ \ldots, (w_{n-2}, w_{n-1}, w_n)\}$$

For instance, let us consider the following Spanish translation of the English question *"Who is the Chairman of the Norwegian Nobel Committee?"*: *"Quién es el Presidente del Comité Nobel noruego?"*. Therefore, w =("Quién", "es", "el", "Presidente", "del", "Comité", "Nobel", "noruego"), and T = {("Quién es el"), ("es el Presidente"), ("el Presidente del"), ("Presidente del Comité"), ("del Comité Nobel"), ("Comité Nobel noruego")}.

The information entropy was introduced by Shannon [54] and its general formulation is:

$$H(X) = -K \sum_{i=0}^{n} p(i) \log p(i) \tag{2.8}$$

Where K is an arbitrary constant which depends on the problem, i is a fragment of a message X of length n, and $p(i)$ is the probability of the i-th fragment. In our case, the message is represented by the translation, and if we take into account trigrams, each fragment i corresponds to the i-th trigram of the translationd t_i.

We decided to calculate the probability of each trigrams by means of web counts. Let us name $c(x)$ the function that returns the number of pages that contain the text fragment x in the web. Let us define the i-th trigram $t_i = (w_i, w_{i+1}, w_{i+2})$ and its root bigram as $b_i = (w_i, w_{i+1})$. According to [62], the probability $p(t_i)$ can be estimated as:

$$p(t_i) = \frac{c(t_i)}{c(b_i)} \tag{2.9}$$

[3]http://babelfish.altavista.com
[4]http://www.freetranslation.com
[5]http://www.linguatec.de
[6]http://www.e-promt.com
[7]http://www.reverso.net

If we substitute $p(i)$ with Formula 2.9 in Formula 2.8, we obtain:

$$H(X) = -K \sum_{i=0}^{n} \frac{c(t_i)}{c(b_i)}(c(t_i) - c(b_i)) \tag{2.10}$$

Due to the fact that in the web usually $c(b_i) >> c(t_i)$, we used the logarithmic scale for page counts, and used a linear normalization factor as K, obtaining the formula that we used to calculate the entropy of a translation X:

$$H(X) = -\frac{1}{n} \sum_{i=0}^{n} \frac{\log c(t_i)}{\log c(b_i)}(\log c(t_i) - \log c(b_i)) \tag{2.11}$$

The selection of the best translation is made on the basis of the $H(X)$ calculated by means of Formula 2.11. Given M translations of a question q, we pick the translation $\bar{m}$ such that $\bar{m} = \arg\max_{m \in M} H(m)$.

It is important to observe that this translation is not passed to the JIRS module, that works with all the translations (passages retrieved by means of the good translations will achieve a better weight), but only to the Answer Extraction module.

An evaluation of the selection procedure was carried out in [50]. Promt resulted the best translator.

2.2.5 Participation to CLEF QA exercises

Our group participated to three editions of CLEF QA: 2005, 2006 and 2007. We publish the results of 2005 and 2006 because CLEF 2007 is a forthcoming event at the moment of writing this report.

CLEF 2005

We participated in the following monolingual task: Spanish, Italian and French, and the Spanish-English and English-Spanish cross-language tasks. In Table 2.2 we show the overall accuracy obtained in all the runs.

Definition questions obtained better results than other kinds of questions, and we suppose this is due to the ease in identifying the target constraint in these cases. Moreover, the results for the Spanish monolingual tasks are better than the other ones, and we believe this is due mostly to the fact that the question classification was performed combining the results of the SVM and pattern classifiers, whereas for French and Italian the expected type of the answer was obtained only via the pattern based classifier. Another reason can be that the majority of the preliminary experiments were done over the CLEF2003 Spanish corpus, therefore resulting in the definition of more accurate patterns for the Spanish Answer Extractor.

In order to evaluate the impact of the answer types, we grouped the results obtained for the best run by the defined categories , as shown in Table 2.3. As it can be

Table 2.2: Accuracy results for the submitted runs. Overall: overall accuracy, factoid: accuracy over factoid questions; definition: accuracy over definition questions; tr: accuracy over temporally restricted questions; nil: precision over nil questions; conf: confidence-weighted score.

task	run	overall	factoid	definition	tr	nil	conf
es-es	upv_051	33.50%	26.27%	52.00%	31.25%	0.19	0.21
it-it	upv_051	25.50%	20.00%	44.00%	16.67%	0.10	0.15
fr-fr	upv_051	23.00%	17.50%	46.00%	6.67%	0.06	0.11
en-es	upv_051	22.50%	19.49%	34.00%	15.62%	0.15	0.10
es-en	upv_051	17.00%	12.40%	28.00%	17.24%	0.15	0.07

seen, the best results have been obtained for the "LOCATION.COUNTRY" category, as expected, due to the use of a customized knowledge source. The worst results have been obtained for the questions "OTHER", for which there is not a defined strategy.

Table 2.3: Accuracy results for the upv 051eses run, grouped by answer type.

category	questions	accuracy
NAME	2	0.0%
NAME.PERSON	25	28.0%
NAME.TITLE	1	0.0%
NAME.LOCATION	6	16.7%
NAME.LOCATION.COUNTRY	14	92.8%
NAME.LOCATION.CITY	2	100.0%
NAME.LOCATION.GEO	2	0.0%
DEFINITION	61	44.3%
DATE	11	36.3%
DATE.DAY	4	0.0%
DATE.YEAR	2	0.0%
QUANTITY	21	33.3%
QUANTITY.AGE	4	25.0%
TIME	4	0.0%
OTHER	41	4.8%

CLEF 2006

We submitted two runs for each of the following monolingual task: Spanish, Italian and French. The first runs (labelled *upv_061*) use the system with JIRS as PR engine, whereas for the other runs we used Lucene, adapted to the QA task with the implementation of a weighting scheme that privileges long passages and is similar to

the word-overlap scheme of the MITRE system [36]. In Table 2.4 we show the overall accuracy obtained in all the runs.

Table 2.4: Accuracy results for the submitted runs. Overall: overall accuracy, factoid: accuracy over factoid questions; definition: accuracy over definition questions; nil: precision over nil questions (correctly answered nil/times returned nil); CWS: confidence-weighted score.

task	run	overall	factoid	definition	nil	CWS
es-es	upv_061	36.84%	34.25%	47.62%	0.33	0.225
	upv_062	30.00%	27.40%	40.48%	0.32	0.148
it-it	upv_061	28.19%	28.47%	26.83%	0.23	0.123
	upv_062	28.19%	27.78%	29.27%	0.23	0.132
fr-fr	upv_061	31.58%	31.08%	33.33%	0.36	0.163
	upv_062	24.74%	26.35%	19.05%	0.18	0.108

With respect to 2005, the overall accuracy increased by $\sim$ 3% in Spanish and Italian, and by $\sim$ 7% in French. We suppose that the improvement in French is due to the fact that the target collection was larger in 2006. Spanish was still the language in which we obtained the best results.

We obtained an improvement over the 2005 system in factoid questions, but also worse results in definition ones, probably because of the introduction of the *object* definitions by the organizers in 2006.

The JIRS-based systems performed better than the Lucene-based ones in Spanish and French, whereas in Italian they obtained almost the same results. The difference in the CWS values obtained in both Spanish and French is consistent and weights in favour of JIRS. This prove that the quality of passages returned by JIRS for these two languages is considerably better.

We measured also the inter-agreement of the two systems, by counting the number of times that the two systems returned the same source document divided by the number of times that they returned the same answer (we called this measure *Agreement on Answer* or *AoA*), and when the answer was the right one (in this case we call it *AoRA*).

Table 2.5: Inter-agreement between the two systems, calculated by means of the AoA and AoRA measures.

task	AoA	AoRA	Collection size
es-es	61.33%	42.52%	1086MB
it-it	71.92%	46.68%	170MB
fr-fr	53.47%	29.81%	487MB

As it can be observed in Table 2.5, the best agreement has been obtained in Italian, as one would expect due to the smaller size of the collection.

2.2.6 Voice-QUASAR: a future development

The first attempts to build a voice-activated question answering systems date back to 1999 [55] for Japanese and 2002 [22] in English. Voice-QUASAR could represent the first example of a Spanish voice-activated QA system. Moreover, the characteristics of both JIRS and the Answer Extraction module make it particularly stable, due to the insensivity to errors in the voice recognition phase.

We carried out a preliminary study with 200 questions of the Spanish monolingual CLEF QA task. We performed some experiments, one with real input speech and others with simulated errors in the input sentences. Due to the difficulty of the task, we have considered very good conditions for the speech recognizer. In order to do that, the language model and the considered vocabulary were obtained exclusively from the sentences. In this way we had not to consider words out of vocabulary. The perplexity of the Language Model (LM) of the original questions was 8.71. In table 2.6 are shown the error rates and perplexity for the question sets.

Table 2.6: Word Error Rates (WER) and Perplexity for the question sets: the original one, the one with real speech (SR), and the two with generated errors (WER20 and WER30).

Question set	WER	LM Perplexity
Original	0%	8.71
SR	13%	11.07
WER20	20%	30.23
WER30	30%	53.89

It must be remarked that many errors in the recognition process are in some keywords, such as *Quién* (Who) that is confused with *Qué* (What). These words, that are acoustically similar in Spanish, are the keywords used by the question analysis module in order to assign the class to the question, thus determining whether the AE module will search for a person or not. Therefore, these errors are crucial for the final result of the QA process. Other kind of errors, consisting in the insertion, substitution or deletion of stopwords are not very important, since the AE module uses the constraints extracted in the question analysis phase. Therefore, it is clear that in the QA task it is very important that the speech recognizer implements a kind of specialization (or confidence measures) over the keywords used for the classification of the questions. A possible way to overcome this problem is by means of a dialog strategy that ask the user to confirm the keywords that appear in the question.

Other experiments were performed by generating errors in the written sentences. We have generated errors considering the same substitution insertion and deletion

proportions that in the real recognition but increasing the total word error rate up to 20% and 30%. We imposed the condition that the Named Entities had to be unaffected by errors. The results are given in table 2.7.

Table 2.7: Obtained results.

Question set	Precision	R answers	X answers
Original	36.5%	73	11
SR	32.0%	64	16
WER20	27.5%	55	17
WER30	19.5%	39	12

The precision was calculated as the number of correct answers divided by 200 (the total number of questions). We recall that we intend as *correct* the answers labeled R in accordance to the CLEF 2005 QA guidelines [56]. However, we decided to report also the number of X answers (partially correct, i.e., the returned answer does not satisfy completely the user's information need), since these kind of errors are directly related to the quality of the AE module and not with the format of the question.

If we compare these results with the precision obtained by our system using the set of CLEF 2005 cross-language English-Spanish questions (22.5% 2.2), where the same questions are given in English and must be automatically translated to Spanish (the language used in the document collection), we can see that the behaviour is better in the case of speech recognition (although it must be taken into account the special condition of the recognizer), even in the case of an error rate of 20%. From our point of view, this result show that spoken QA should be worth the same attention of cross-language tasks.

2.3 Resources for QA

2.3.1 Wikipedia as a Resource for QA

Encyclopedic knowledge is valuable for many Natural Language Processing (NLP) applications, and in particular for the Question Answering (QA) task. Recently, the availability of a large, open domain encyclopedia, such as the Wikipedia[8], has captured the attention of some researchers [38, 3] in the Question Answering field. Until now, the focus of these works was on the use of the encyclopedia in order to look for the answer to the questions. However, the results did not fulfill the expectations.

We investigated the use of Wikipedia in some slightly different aspects of the Question Answering task: answer validation and generation of answer patterns. In the first case, the problem consists in, given a possible answer, saying wether it is the right one or not. Previous work on answer validation has been carried out

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

by exploiting the redundancy of the web [42], giving good results. In the case of encyclopedias, redundancy is not an option, because usually each topic is covered by no more than one article. Therefore, the quality of the information extracted from the question is crucial to find the related article.

In the second case, the problem consists in building a regular expression pattern that (possibly) match the right answer. When a question pertains to a specific class, usually deduced by the structure of the question (for instance, the right answer for a question starting with the word *where* will be *at least* some kind of location) patterns can be built by hand, usually together with a custom-built ontology [32]. However, when the question cannot be classified using a given taxonomy, the semantic class can be deduced by the question itself, such as in "Which *fruit* contains vitamine C?": in this case, the class is "fruit", and we want to find a suitable answer string for that class. In order to do that, we exploited the categorization of articles in Wikipedia. For instance, the article corresponding to the category *http://en.wikipedia.org/wiki/Category:Fruit* contains a list of fruits. It can be observed that the "category" entries constitute a sort of Wikipedia ontology, since some categories contain also subcategories.

We carried out some experiments with the set of 200 Spanish monolingual questions from the CLEF 2005 Question Answering track. The objective was to find in which cases Wikipedia was helpful. The manual evaluation, imagining a "perfect" passage retrieval and answer extraction system, found that the potential improvement with the help of Wikipedia was of 29 questions, corresponding to a $14,5\%$ gain in recall and coverage. See Table 2.8 for details.

Question type	Answers (tot)	Pot. Rec. gain
All	29 (200)	14,5%
Definition	8 (50)	16,0%
Name	7 (42)	16,6%
Generic	5 (25)	20,0%

Table 2.8: Potential recall gain (i.e., questions where Wikipedia could be useful but was not possible to use its information), grouped by question type.

Another interesting feature discovered by error analysis is that when Wikipedia proved to be useless, it is usually due to one of the following reasons:

- The question is about facts unrelated with the Spanish world. That is, the answer could be present in another localization of Wikipedia. For instance, the answer to the question *Who is Giulio Andreotti?* could be find in the Italian or English versions of Wikipedia.
- The question is about facts too specific to be taken into account into the

Wikipedia. For instance, *Who discovered the galleon San Diego?*, or *Who is Rolf Ekeus?*.

More less significative failure reasons were the ambiguity of some categories (for instance, "Qué *plataforma* estaba acampada en el Paseo de la Castellana de Madrid ?" - "Which platform was camped at Paseo de la Castellana in Madrid?"), or the fact that the category was imaginary (for instance, "Para qué *periódico* trabajaba Clark Kent?", "For which newspaper does Clark Kent work?": the category *newspapers* (periódicos) exists, but does not contain the *Daily Planet*).

Although the results obtained showed that Wikipedia can be actually used to improve the performance of our Question Answering system, especially for "Generic" questions, they are well below the potential. This is due mainly to the following three reasons: the performance of passage retrieval and answer extraction systems, the localization of Wikipedia editions, and the fact that knowledge related to small-scale events or less known people usually is not included into the Wikipedia. In this last case, no action can be taken, since it is a feature of a massive distributed project like Wikipedia; however, we can work to improve the passage retrieval system and answer extraction subsystem, obtaining better passages and candidate answers. Another interesting work direction should be a multilingual approach that could take into account the various localizations of Wikipedia in the other languages, preferably those containing many articles.

Participation to WiQA

WiQA was a task aimed at helping the readers/authors of Wikipedia rather than finding answers to user questions. In the words of the organizers[9], the purpose of WiQA was "to see how IR and NLP techniques can be effectively used to help readers and authors of Wikipages get access to information spread thoughout Wikipedia rather than stored locally on the pages". An author of a given Wikipage can be interested in collecting information about the topic of the page that is not yet included in the text, but is relevant and important for the topic, so that it can be used to update the content of the Wikipage. Therefore, an automatic system will provide the author with information snippets extracted from Wikipedia with the following characteristics:

- *unseen*: not already included in the given source page;
- *new*: providing new information (not outdated);
- *relevant*: worth the inclusion in the source page.

[9] http://ilps.science.uva.nl/WiQA/Task/index.html

Our approach to WiQA exploited a simple bag-of-words method based on the supposed behaviour of a typical user (i.e. a WIkipedia author/editor). The user behaviours emulated by our system are the following:

1. The user searches for pages containing the title of the page he is willing to expand;

2. The user analyzes the snippets, discarding the ones being too similar to the source page.

In the first case, passing the title as phrase to Lucene is enough for most topics. We observed that some topics needed a better analysis of title contents, such as the 7th of the English monolingual test set:"*Minister of Health (Canada)*", however these cases were not so many with respect to the total number of topics.

In the second case, the notion of *similarity* between a snippet and a page is the key for obtaining unseen (and relevant) snippets. Note that we decided to bound together relevance and the fact of not being already present in the page; we did not implement any method in order to determine whether the snippets contain outdated informations or not. In our system, the similarity between a snippet and the page is calculated by taking into account the number of terms they share. Therefore, we define the similarity $f_{sim}(p, s)$ between a page p and a snippet s as:

$$f_{sim}(p, s) = \frac{|p \cap s|}{|s|} \tag{2.12}$$

Where $|p \cap s|$ is the number of terms contained in both p and s, and $|s|$ is the number of terms contained in the snippet. This measure is used to rearrange the ranking of snippets by penalizing those being too similar to the source page. If w_l is the standard $tf \cdot idf$ weight returned by Lucene, then the final weight w of the snippet s with respect to page p is:

$$w(s, p) = w_l \cdot (1 - f_{sim}(p, s)) \tag{2.13}$$

The snippets are then ranked according to the values of w.

The obtained results are shown in Table 2.9. The most important measure is the average yield, that is, the average number of "good" snippets per topic among top 10 snippets returned. The best average yield of the systems participating to the task was up to 3.38 for the English monolingual subtask [**?**]. We remark that our system ranked as the best one in Spanish, although the number of participants was smaller than for the English task.

The *Average yield* is calculated as the total number of important novel non-repeated snippets for all topics, divided by the number of topics. The *MRR* is the Mean Reciprocal Rank or the first important non-repeated snippet. The precision is calculated as the number of important novel non-repeated snippets, divided by the total number of snippets per topic. Our system returned always a number of snippets ≤ 10.

Table 2.9: Results obtained by our system.

Language	Average yield	MRR	Precision
English	2.6615	0.4831	0.2850
Spanish	1.8225	0.3661	0.2273

2.3.2 WordNet

WordNet is an ontology of English, developed at the University of Princeton under the direction of G.Miller ([43]). Its last version (3.0) contains 155,327 words grouped into 117,597 *synsets*. A synset (*set* of *syn*onyms) is a group of words that are considered semantically equivalent. An example of synset for a geographical location is the following: `(London, Greater London, British capital, capital of the United Kingdom)`. Each synset is associated to an unique id and a *gloss*, i.e. the definition of the concept (in the case of London: `the capital and largest city of England; located on the Thames in southeastern England; financial and industrial and cultural center`). Moreover, the most important feature of WordNet is that it also provides a set of semantic relationships which connect different synsets.

Probably, the most important relationship provided by WordNet is the *hypernymy* (or *is-a*) relationship. This relationship connects two concepts where one is more general than the other, such as '*clock*' and '*cuckoo clock*'. The inverse relationship (from a more specific concept to a more general one) is called *hyponymy*. The *meronymy*, or *part-of*, relationship connects concepts that are part of another one and vice versa (in the latter case it is named *holonymy*). In the example of Figure **??**, '*England*' is holonym of '*London*'. Finally, the *instance* relationship connects abstract concepts to real world instances, such as '*clock*' and '*Big Ben*'. Most relationships connect words of the same lexical category, also known as Part-of-Speech (POS) category, such as those named here, which connect only noun concepts.

WordNet has been widely used in NLP, mainly because of its role as sense inventory. It was also employed also to semantically annotate the Brown corpus ([28]), obtaining the SemCor (*Sem*antic *Cor*respondance) corpus ([30]). In SemCor every word belonging to the noun, verb, adjective and adverb POS categories has been labelled with a WordNet sense. It is often used as a training corpus for supervised word sense disambiguation methods. Its use as a resource for QA systems has been extensively discussed in [47].

WordNet contains valuable lexico-semantic knowledge that can be exploited in all modules of a state-of-the-art QA system. In question processing, the identification of the kind of answer expected requires extensive semantic information. For example, given the question "What flowers did Van Gogh paint?", we need to know that the answer type is a kind of flower. WordNet encodes 470 hyponyms of flowers that can be searched in the retrieved passages. WordNet can also help with synonyms for reformulations of the questions [58].

Chapter 3

Geographical Information Retrieval

Nowadays, many documents in the web or in digital libraries contain some kind of geographical information. News stories often contain a reference that indicates the place where an event took place. Nevertheless, the correct identification of the locations to which a document refers to is not a trivial task. Explicit information about areas including the cited geographical entities is usually missing from texts (e.g. usually *France* is not named in a news related to *Paris*). Moreover, using text strings in order to identify a geographical entity creates problems related to ambiguity, synonymy and names changing over time.

Ambiguity and synonymy are well-known problems in the field of Information Retrieval. The use of semantic knowledge may help to solve these problems, even if no strong experimental results are yet available in support of this hypothesis. Some results [6] show improvements by the use of semantic knowledge; others do not [51]. The most common approaches make use of standard keyword-based techniques, improved through the use of additional mechanisms such as document structure analysis and automatic query expansion.

We investigated the use of automatic query expansion by means of WordNet [43] meronyms and synonyms in our 2005 to the GeoCLEF, but the obtained results were below the average of participants [19, 11]. Although there are some effective query expansion techniques [16] that can be applied to the geographical domain, we think that the expansion of the queries with synonyms and meronyms does not fit with the characteristics of the GeoCLEF task. Other methods using thesauri with synonyms for general domain IR also did not achieve promising results [59].

In our work for GeoCLEF 2006 we focused on the use of WordNet in the indexing phase, specifically for the expansion of index terms by means of synonyms and holonyms. We used the subset of the WordNet ontology related to the geographical domain. It is quite difficult to calculate the number of geographical entities stored in WordNet, due to the lack of an explicit annotation of the synsets, however we retrieved some figures by means the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains.

Geographical resources like gazetteers usually contains a much greater quantity of information. For instance, the Geonet Names Server[1] (GNS) contains more than 5 million of place names.

In the following section we describe in detail the techniques developed for our participations to the GeoCLEF 2005 and 2006 tasks.

3.1 WordNet-based GIR methods

There can be many different ways to refer to a geographical entity. This may occur particularly for foreign names, where spelling variations are frequent (e.g. *Rome* can be indicated also with its original italian name, *Roma*), acronyms (e.g. *U.K.* or *G.B.* used instead of the extended form *United Kingdom of Great Britain and Northern Ireland*), or even some popular names (for instance, *Paris* is also known as the *ville lumiére*, i.e., the city of light). Each one of these cases can be reduced to the *synonymy* problem. Moreover, sometimes the rhetoric figure of *metonymy* (i.e., the substitution of one word for another with which it is associated) is used to indicate a greater geographical entity (e.g. *Washington* for *U.S.A.*), or the indication of the including entity is omitted because it is supposed to be well-known to the readers (e.g. *Paris* and *France*).

WordNet can help in solving these problems. In fact, WordNet provides synonyms (for instance, {*U.S., U.S.A., United States of America, America, United States, US, USA* } is the synset corresponding to the "*North American republic containing 50 states*"), and meronyms (e.g. *France* has *Paris* among its meronyms), i.e., concepts associated through the "part of" relationship.

Therefore, it is straightforward to employ WordNet as a resource for Geographical Information Retrieval.

3.1.1 Query Expansion

Our first approach was to develop a query expansion method in order to take advantage from the synonymy and meronymy relationships. The query is first tagged with POS labels. After this step, the query expansion is done in accordance to the following algorithm:

1. Select from the query the next word (w) tagged as proper noun.
2. Check in WordNet if w has the {*country, state, land*} synset among its hypernyms; if not, return to 1, else add to the query all the synonyms, with the exception of stopwords and the word w, if present; then go to 3.

[1] http://earth-info.nga.mil/gns/html/index.html

3. Retrieve the meronyms of w and add to the query all the words in the synset containing the word *capital* in its gloss or synset, except the word *capital* itself. If there are more words in the query, return to 1, else end.

For example, the query: *Shark Attacks off Australia and California* is POS-tagged as follows: NN/shark, NNS/attacks, PRP/off, NNP/Australia CC/and NNP/California. Since "Shark" and "Attacks" do not have the {*country, state, land*} synset among their hypernyms, therefore Australia is selected as the next w. The corresponding WordNet synset is {*Australia, Commonwealth of Australia*}, with the result of adding "*Commonwealth of Australia*" to the expanded query. Moreover, the following meronym contains the word "capital" in synset or gloss: "*Canberra, Australian capital, capital of Australia - (the capital of Australia; located in southeastern Australia)*", therefore *Canberra* is also included in the expanded query. The next w is *California*. In this case the WordNet synset is {*California, Golden State, CA, Calif.*}, and the words added to the query are "*Golden State*", "*CA*" and "*Calif.*". The following two meronyms contain the word "capital":

- *Los Angeles, City of the Angels - (a city in southern California; motion picture capital of the world; most populous city of California and second largest in the United States)*
- *Sacramento, capital of California - (a city in north central California 75 miles northeast of San Francisco on the Sacramento River; capital of California)*

Moreover, during the POS tagging phase, the system looks for word pairs of the kind "adjective noun" or "noun noun". The aim of this step was to imitate the search strategy that a human would attempt. Stopwords are also removed from the query during this phase. Therefore, the expanded query that is handed over to the search engine is: *"shark attacks" Australia California "Commonwealth of Australia" Canberra "Golden State" CA Calif. "Los Angeles" "City of the Angels" Sacramento.*

Experimental Results

For every query the top 1000 ranked documents were returned by the system. We performed two runs, one with the unexpanded queries, the other one with expansion. For both runs we plotted the precision/recall graph (see Fig. 3.1) which displays the precision values obtained at each of the 10 standard recall levels.

The obtained results show that our system was the worst among the participants to the exercise [19]. The query expansion technique proved effective only in a few topics (particularly the topic number 16: "Oil prospecting and ecological problems in Siberia and the Caspian Sea"). The worst results were obtained for topic number 5 ("Japanese Rice Imports").

There are two main explanations for the obtained results: the first is that the keyword grouping heuristic was too simple: for instance, in topic number 5 the words

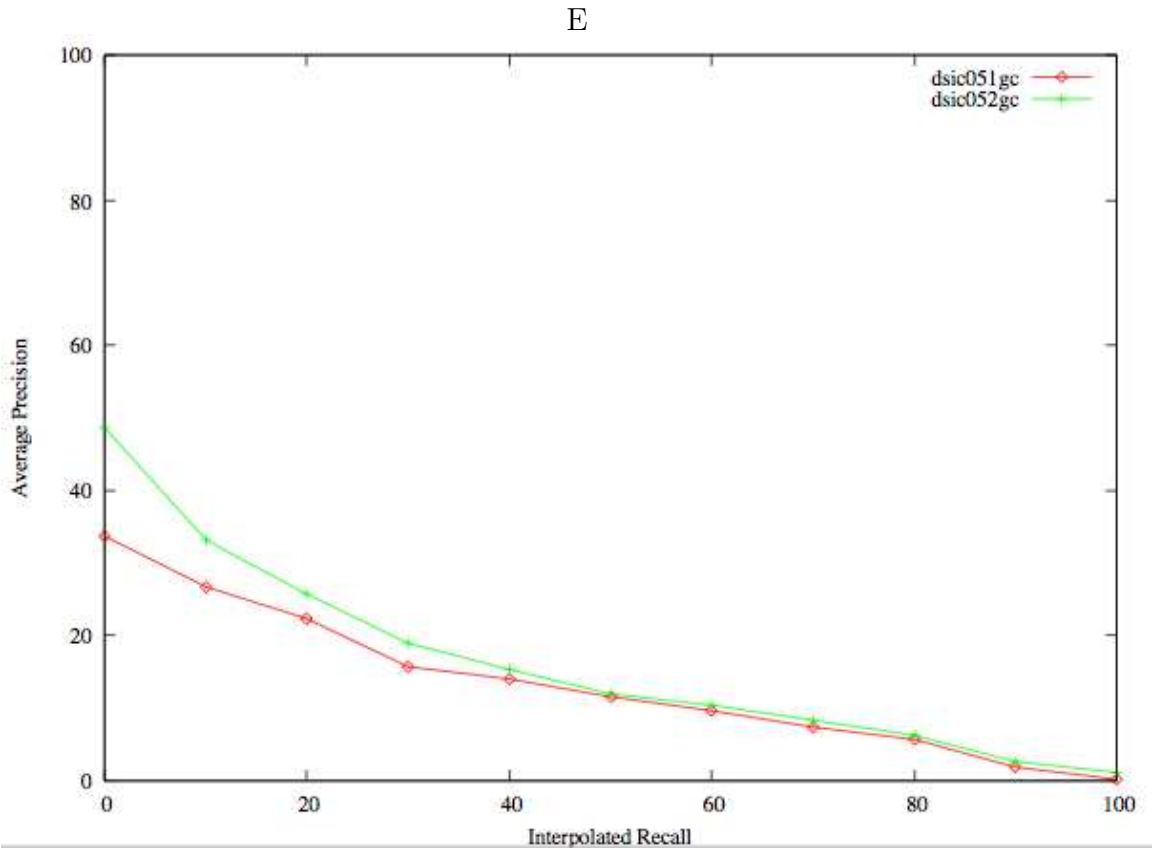

Figure 3.1: Interpolated precision/recall graph for the two system runs: *dsic051gc*, using only the topic title and description fields, and *dsic052gc*, using also the "concept" and "location" fields.

are grouped as: "Japanese Rice" and "Imports". The correct grouping should be "Japanese" and "Rice Imports". The second reason is that the expansion may introduce unnecessary information. For example, if the user is asking about "shark attacks in California", we have seen that *Sacramento* is added to the query. Therefore, documents containing "shark attacks" and "Sacramento" will obtain an higher rank, with the result that documents that contain "shark attacks" but not "Sacramento" are placed lower in the ranking. Since it is unlikely to observe a shark attack in Sacramento, the result is that the number of documents in the top positions will be reduced with respect to the one obtained with the unexpanded query, with the consequence of achieving a smaller precision.

In order to better understand the obtained results, we compared them with two baselines, the first obtained by submitting to the Lucene search engine the query without the synonyms and meronyms, and the latter by using only the tokenized fields from the topic. For instance, the query *"shark attacks" Australia California*

"Commonwealth of Australia" Canberra "Golden State" CA Calif. "Los Angeles" "City of the Angels" Sacramento would be *"shark attacks" Australia California* for the first baseline (without WN) and *shark attacks Australia California* in the second case.

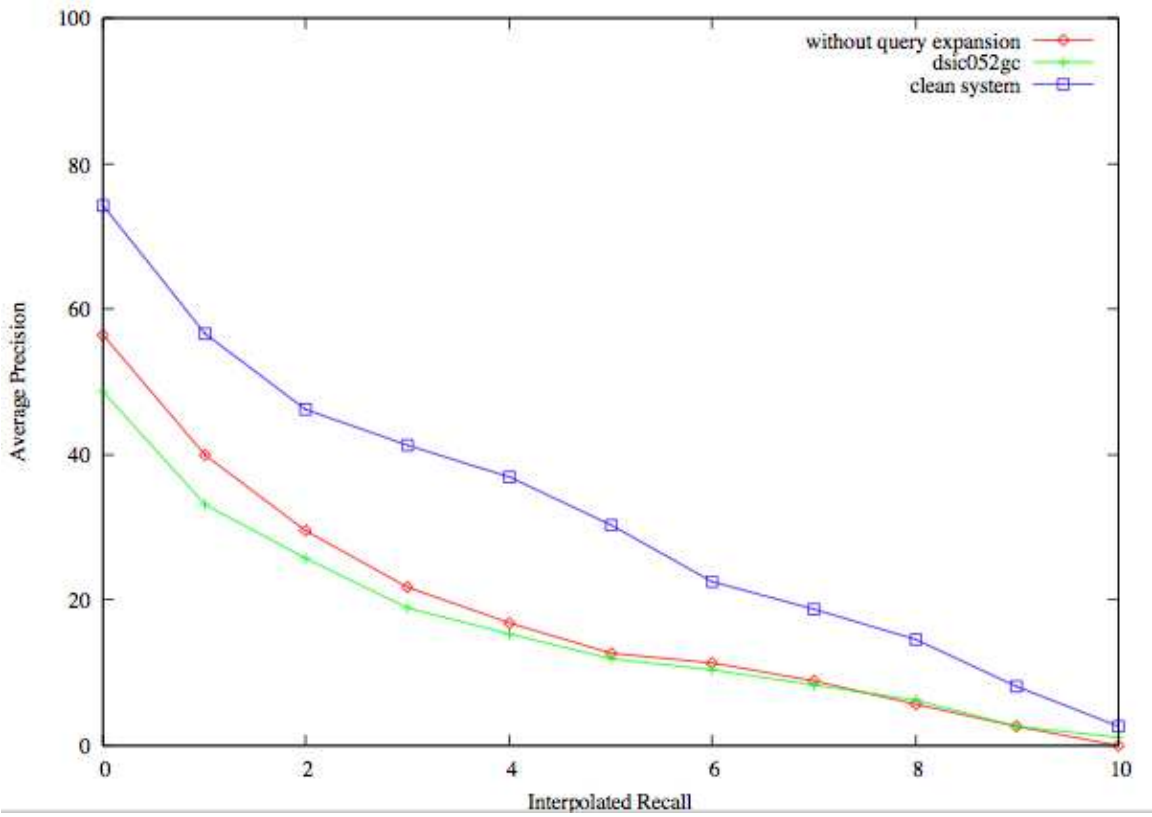

Figure 3.2: Comparison of our best run (dsic052gc) with the "without query expansion" baseline and the clean system (neither query expansion nor keyword grouping).

The interpolated precision/recall graph in Fig. 3.2 demonstrates that both of our explanations for the obtained results are correct: in fact, the system using keyword grouping but not query expansion performs better than the system that uses both; however, this system is still worse than the one that do not use neither the query expansion nor keyword grouping.

3.1.2 Index Term Expansion

The expansion of index terms is a method that exploits the *holonymy* relationship of the WordNet ontology. A concept A is *holonym* of another concept B if A contains B, or viceversa B is part of A (B is also said to be *meronym* of A). Therefore, our idea is to add to the geographical index terms the informations about their holonyms, such

that a user looking information about *Spain* will find documents containing *Valencia*, *Madrid* or *Barcelona* even if the document itself does not contain any reference to Spain.

We used the well-known Lucene[2] search engine. Two indices are generated for each text during the indexing phase: a *geo* index, containing all the geographical terms included in the text and also those obtained through WordNet, and a *text* index, containing the stems of text words that are not related to geographical entities. Thanks to the separation of the indices, a document containing "John Houston" will not be retrieved if the query contains "Houston", the city in Texas. The adopted weighting scheme is the usual $tf \cdot idf$. The geographical terms in the text are identified by means of a Named Entity (NE) recognizer based on maximum entropy[3], and put into the *geo* index, together with all its synonyms and holonyms obtained from WordNet.

For instance, consider the following text:

"*A federal judge in Detroit struck down the National Security Agency's domestic surveillance program yesterday, calling it unconstitutional and an illegal abuse of presidential power.*"

The NE recognizer identifies *Detroit* as a geographical entity. A search for Detroit synonyms in WordNet returns {*Detroit, Motor city, Motown*}, while its holonyms are:

```
-> Michigan, Wolverine State, Great Lakes State, MI
   -> Midwest, middle west, midwestern United States
       -> United States, United States of America, U.S.A., USA,
            U.S., America
           -> North America
             -> northern hemisphere
             -> western hemisphere, occident, New World
             -> America
```

Therefore, the following index terms are put into the *geo* index: { *Michigan, Wolverine State, Great Lakes State, MI, Midwest, middle west, midwestern United States, United States, United States of America, U.S.A., USA, U.S., America, North America, northern hemisphere, western hemisphere, occident, New World* }.

Experimental Results

In table 3.1 we show the recall and average precision values obtained by our system at GeoCLEF 2006. Recall has been calculated for each run as the number of relevant

[2]http://lucene.jakarta.org

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

documents retrieved divided by the number of relevant documents in the collection (378). The average precision is the non-interpolated average precision calculated for all relevant documents, averaged over queries.

The results obtained in term of precision show that non-WordNet runs are better than the other ones, particularly for the all-fields run *rfiaUPV02*. However, as we expected, we obtained an improvement in recall for the WordNet-based system, although the improvement was not so significant as we hoped (about 1%).

Table 3.1: Average precision and recall values obtained for the four runs. WN: tells whether the run uses WordNet or not.

run	WN	avg. precision	recall
rfiaUPV01	no	25.07%	78.83%
rfiaUPV02	no	**27.35%**	80.15%
rfiaUPV03	yes	23.35%	79.89%
rfiaUPV04	yes	26.60%	**81.21%**

In order to better understand the obtained results, we analyzed the topics in which the two systems differ more (in terms of recall). Topics 40 and 48 resulted the worst ones for the WordNet based system. The explication is that topic 40 does not contain any name of geographical place ("*Cities near active volcanoes*"); topic 48 contains references to places (*Greenland* and *Newfoundland*) for which WordNet provides little information.

On the other hand, the system based on index term expansion performed particularly well for topics 27, 37 and 44. These topics contain references to countries and regions (*Western Germany* for topic 27, *Middle East* in the case of 37 and *Yugoslavia* for 44) for which WordNet provides a rich information in terms of meronyms.

3.1.3 The Problem of toponym ambiguity

A great portion of the information currently available in digital format is constituted by textual, unstructured documents. The continuous growth of this kind of information and the increasing number of users that can access it constitutes a challenge to the developers of Information Retrieval (IR) systems. One of the most challenging problems is the *ambiguity* of human language. When searching for specific keywords, it is desirable to eliminate occurrences in documents where the word or words are used in an inappropriate sense ([25]). Ambiguity can be of various types: proper names may identify different class of named entities (for instance, 'London' may identify the writer 'Jack London' or a city in the UK), or may be used as a name for different instances of a same class; e.g. 'London' is also a city in Canada. The task of assigning the most appropriate sense to a word within its context is named *Word*

Sense Disambiguation (WSD). Notably, this is still an open problem in the field of Natural Language Processing (NLP).

In [8] we studied the application of a knowledge-based WSD method in the geographical domain, specifically to the disambiguation of toponyms. The method we propose is based on the one ([52]) we developed for the disambiguation of nouns, which implemented a variation of the *Conceptual Density* formula by [2].

Toponym disambiguation is a relatively new field. From a NLP perspective, it is merely the application of WSD to place names. Its most direct application should be the improvement of the searches both in the Web or in large news collections, due to the fact that it is very common to find geographical information in web pages or news stories (e.g. '*Elections in Italy*', '*Plane crash in Teheran*'). A growing interest in the field of Geographical Information Retrieval (GIR) is testified by the recent creation of the GeoCLEF exercise and the increment of the attendance at the GIR workshops[4] held at the last SIGIR events. The lack of a reference corpus has long been an obstacle to the evaluation of algorithms for toponym resolution ([33]). Recently some corpora have been compiled ([17, 34]), but the lack of a mapping between WordNet and the locations IDs used in these corpora prevented us from evaluating our method with these resources. We overcome this problem by selecting the geographical entities in the SemCor[5] corpus that was originally developed for the WSD task.

[4]http://www.geo.unizh.ch/~rsp/gir06/

[5]http://www.cs.unt.edu/~rada/downloads.html#semcor

Chapter 4

Word Sense Disambiguation

WSD is a long standing problem in computational linguistics. The most extended approach is to attempt to use the context of the word to be disambiguated together together with information about each of its word senses to solve this problem. The WordNet ontology, based on synsets (sets of synonyms), is the extemal lexical rcsource which is used to perform the WSD task. When the initial input source of information (i.e., the word and its context) is only processed together with the lexical knowledge source, a fully automatic method which do not require any kind of training process is needed to perform the word sense disambiguation.

In the following sections of this chapter we will present a new method based on Conceptual Density [2] and its improvements. This method is an high precision method that can be applied both in Question Answering and Geographical Information Retrieval in order to solve the ambiguity problems that can be encountered in both tasks.

4.1 Conceptual Density - based WSD

Conceptual Density (CD) was introduced by [2] as a measure of the correlation between the sense of a given word and its context. It is computed on WordNet subhierarchies, determined by the hypernymy relationship. The disambiguation algorithm by means of CD consists of the following steps:

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

Our formulation of the Conceptual Density of a WordNet subhierarchy s is ([52]):

$$CD(m, f, n) = m^{\alpha} \left(\frac{m}{n}\right)^{\log f}, \tag{4.1}$$

where m are the *relevant* synsets in the subhierarchy, n is the total number of synsets in the subhierarchy, and f is the rank of frequency of the word sense related to the subhierarchy (e.g. 1 for the most frequent sense, 2 for the second one, etc.). The inclusion of the frequency rank means that less frequent senses are selected only when $m/n \geq 1$. The relevant synsets are both the synsets of the word to disambiguate and those of the context words. Our formulation allows to solve some problems with the original CD due to the higher granularity of newer WordNet versions.

The WSD system based on this formula obtained 81.5% in precision over the nouns in the SemCor (baseline: 75.5%, calculated by assigning to each noun its most frequent sense), and participated at the Senseval-3 competition as the CIAOSENSO system ([10]), obtaining 75.3% in precision over nouns in the all-words task (baseline: 70.1%). These results were obtained with a context window of only two nouns, the one preceding and the one following the word to disambiguate.

4.2 Conceptual-density based Toponym Disambiguation

When we considered adapting this algorithm to the disambiguation of toponyms, we realised that the hypernymy relation was not suitable. For instance *Cambridge(1)* and *Cambridge(2)* are both instances of the '*city*' concept and therefore, they share the same hypernyms. The result is that the subhierachies are composed only by the synsets of the two senses of '*Cambridge*', and they are left undisambiguated because their density is the same (in both cases it is 1).

Our idea is to consider the *holonymy* relationship instead of hypernymy. With this relationship it is possible to create subhierarchies that allow to discern different locations (having the same name) in a more effective way. For instance, the last three holonyms for '*Cambridge*' are:

(1) `Cambridge` → `England` → `UK`

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

The best choice for context words is represented by other place names, because holonymy is always defined through them and because they constitute the actual 'geographical' context of the toponym we are disambiguating. In Figure 4.1 we show an example of a holonym tree obtained for the disambiguation of '*Georgia*' with the context '*Atlanta*', '*Savannah*' and '*Texas*', from the following fragment of text extracted from the `br-a01` file of SemCor:

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

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

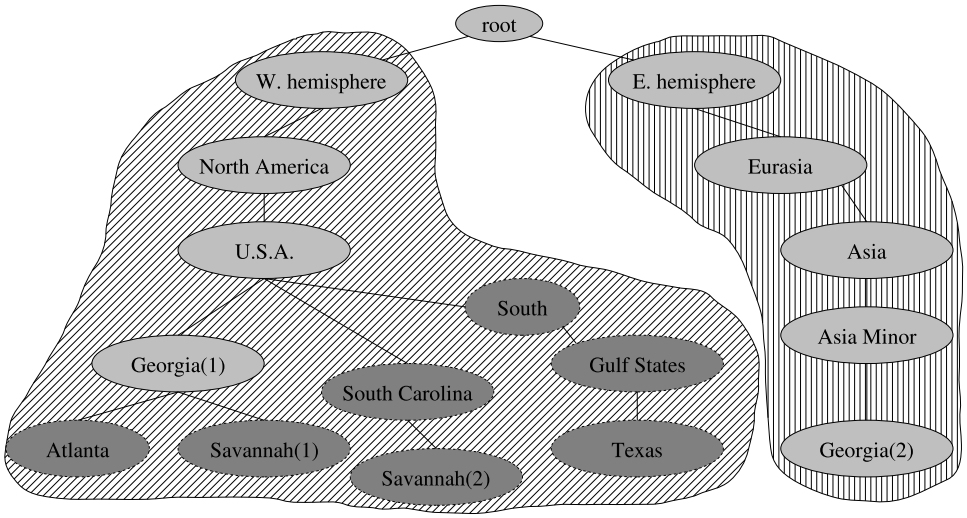

Figure 4.1: Example of holonym hierarchy for the disambiguation of *Georgia*, with context: $\{Atlanta, Savannah, Texas\}$ from the file `br-a01` of SemCor. Nodes are synsets, dark grey nodes are synset of context words.

As one would expect, the holonyms of the context words populate exclusively the subhierarchy related to the first sense (the area filled with a diagonal hatching in Figure 4.1); this is reflected in the CD formula, which returns a CD value 4.29 for the first sense ($m = 8$, $n = 11$, $f = 1$) and 0.33 for the second one ($m = 1$, $n = 5$, $f = 2$). In this work, we considered as relevant also those synsets which belong to the paths of the context words that fall into a subhierarchy of the toponym to disambiguate.

4.3 Integration of WSD methods: a Fuzzy Borda approach

One of the lessons learned from our previous experience at Senseval-3[1] [9, 57] is that the integration of different systems usually works better than a standalone system. In our opinion this reflects the reality where humans do not apply always the same rule in order to disambiguate the same ambigue word; for instance, if we consider the sentences "*He hit a home run*" and "*The thermometer hit 100 degrees*", in the first case the sport domain helps in determining the right sense for *hit*, whereas in the latter the disambiguation is carried out mostly depending on the fact that the subject of the sentence is an object.

The combination of distinct methods represents itself a major problem. If the methods return different answers, how can we select the best one? In this sense the available choices are the following:

- *Rule-based* selection: a set of rules that can be both hand-made or automatically learned from examples;
- *Probability-based*: the output of the methods is normalized in the range $[0, 1]$ and is considered as a probability. Then the values are multiplied in order to obtain the sense with a maximum probability.
- *Vote-based*: the output of the methods is considered as a weighted vote. Then a voting scheme is used in order to obtain the most voted sense.

Working with probabilities can be problematic due to the null probabilities that make necessary the adoption of smoothing techniques. Therefore, we opted for a voting scheme, in this case the fuzzy Borda [46, 18], one of the best known methods in the field of collective decision making. With this scheme the disambiguation methods are considered as experts providing a preference ranking over the sense of the word.

The methods we choose as experts are the sense probability calculated over SemCor, the Conceptual Density algorithm by [52], the extended Lesk by [5], and an algorithm that takes into account the domains of the word to be disambiguated and the context words. In the following sections we describe in detail the fuzzy Borda scheme and each WSD expert.

4.3.1 Fuzzy Borda voting

The original Borda vote-counting scheme was introduced in 1770 by Jean Charles de Borda, and adopted by the French Academy of Sciences with the purpose of selecting its members. In the classical Borda count each expert gives a mark to each

[1] http://www.senseval.org

alternative, according to the number of alternatives worse than it. The fuzzy variant [46, 18] is a natural extension that allows the experts to show numerically how much some alternatives are preferred to the others, evaluating their preference intensities from 0 to 1.

Let $R^1, R^2, \ldots, R^m$ be the fuzzy preference relations of m experts over n alternatives $x_1, x_2, \ldots, x_n$. For each expert k we obtain a matrix of preference intensities:

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

where each $r_{ij}^k = \mu_{R^k}(x_i, x_j)$, with $\mu_{R^k} : X \times X \rightarrow [0, 1]$ being the membership function of R^k. The number $r_{ij}^k \in [0, 1]$ is considered as the degree of confidence with which the expert k prefers x_i to x_j. The final value assigned by the expert k to each alternative x_i is:

$$r_k(x_i) = \sum_{j=1, r_{ij}^k > 0.5}^{n} r_{ij}^k \tag{4.2}$$

which coincides with the sum of the entries greater than 0.5 in the i-th row in the preference matrix. The threshold 0.5 ensure the relation $R^k$to be an ordinary preference relation [18].

Therefore, the definitive fuzzy Borda count for an alternative x_i is obtained as the sum of the values assigned by each expert:

$$\mathbf{r}(x_i) = \sum_{k=1}^{m} r_k(x_i) \tag{4.3}$$

In order to fill the preference matrix with the correct confidence values, the output weights $w_1, w_2, \ldots, w_n$ of each expert k are transformed to fuzzy confidence values by means of the following transformation:

$$r_{ij}^k = \frac{w_i}{w_i + w_j} \tag{4.4}$$

An example of how fuzzy Borda is used to combine the votes in order to obtain the right sense of the target word is shown in Section 4.3.1.

We considered five experts in order to carry out the disambiguation process. Sense probability and the extended lesk were available for every word, while the Conceptual Density was calculated only for nouns. Therefore, all the experts were available only for the nouns. For each expert different contexts were taken into account, depending on the specific characteristics of each expert.

Sense Probability

This expert is the simplest one: its votes are calculated using only the frequency count in SemCor of the WordNet senses of the word. The transformation of the frequency counts to the preference ranking is done according to Formula (4.4). Zero frequency are normalized to 1.

A second CD-based expert exploits the *holonymy*, or *part-of* relationship instead of *hyperonymy*. This expert uses as context all the nouns in the sentence of the word to be disambiguated.

Extended Lesk

This expert is based on the algorithm by [5], a WordNet-enhanced version of the well-known dictionary-based algorithm proposed by [35]. The original Lesk was based on the comparison of the gloss of the word to be disambiguated with the context words and their glosses. This enhancement consists in taking into account also the glosses of concepts related to the word to be disambiguated by means of various WordNet relationships. Then similarity between a sense of the word and the context is calculated by means of *overlaps*. The word is assigned the sense obtaining the best overlap match with the glosses of the context words and their related synsets.

The weights used as input for Formula (4.4) are the similarity values between the senses of the world and the context words. The context for this expert consists of 4 WordNet words (disregarding their Part-Of-Speech) located in the same sentence of the word to be disambiguated, i.e., words with POS noun, verb, adjective or adverb that can be found in WordNet.

WordNet Domains

This expert uses WordNet Domains [40] in order to provide the system with domain-awareness. All WordNet words in the same sentence of the target word are used as context. The weight for each sense is obtained by counting the number of times the same domain of the sense appears in the context (all senses of context words are considered). We decided to not take into account the "factotum" domain.

Example

In this example we will consider only the sense probability and extended Lesk experts for simplicity.

Let us consider the following phrase: "*And he has kept mum on how his decision might affect a bid for United Airlines , which includes a big stake by British Airways PLC.*" with *affect* as target word. We can observe that in WordNet the verb *affect* has 5 senses. The sense count values are 43 for the first sense, 11 for the second, 4 for both the third and the fourth one, and 0 for the last one. We decided to normalize the

cases with 0 occurrences to 1. After applying the transformation (4.4) to the sense counts, we obtain the following preference matrix for the sense probability expert:

$$\begin{pmatrix} 0.5 & 0.80 & 0.91 & 0.91 & 0.98 \\ 0.20 & 0.5 & 0.73 & 0.73 & 0.92 \\ 0.09 & 0.27 & 0.5 & 0.5 & 0.8 \\ 0.09 & 0.27 & 0.5 & 0.5 & 0.8 \\ 0.02 & 0.08 & 0.2 & 0.2 & 0.5 \end{pmatrix}$$

Therefore, the final fuzzy Borda counts by the sense probability expert are 3.60 for `affect(1)`, 2.38 for `affect(2)`, 0.8 for `affect(3)` and `affect(4)`, and 0 for `affect(5)`, obtained from the sum of the rows where the value is greater than 0.5.

The extended Lesk expert calculates the following similarity scores for thesenses of *affect*, with context words *decision*, *might*, *bid* and *include*: respectively 107, 70, 35, 63 and 71 for senses 1 to 5. After applying the transformation (4.4) to the weights, we obtain the preference matrix for this expert:

$$\begin{pmatrix} 0.5 & 0.60 & 0.75 & 0.63 & 0.60 \\ 0.40 & 0.5 & 0.67 & 0.53 & 0.49 \\ 0.25 & 0.33 & 0.5 & 0.36 & 0.33 \\ 0.37 & 0.47 & 0.64 & 0.5 & 0.47 \\ 0.40 & 0.51 & 0.67 & 0.53 & 0.5 \end{pmatrix}$$

In this case the final fuzzy Borda counts are 2.58 for the first sense, 1.2 for sense 2, 0 for sense 3, 0.64 and 1.71 for senses 4 and 5 respectively.

Finally, the sum of Borda counts of every expert for each sense (see Table 4.3.1) are used to disambiguate the word.

sense no:	1	2	3	4	5
expert 1	3.60	2.38	0.80	0.80	0
expert 2	2.58	1.20	0	0.64	1.71
total:	**6.18**	3.58	0.80	1.44	1.71

Table 4.1: Borda Count for the verb *affect* in the example phrase.

The system was not tested before SemEval. Our participation was limited to the All-Word and Coarse-Grained tasks (without the sense inventory provided by the organizers). The results are compared to the best system and the MFS (Most Frequent Sense) baseline. The results of the coarse grained task demonstrates that our system obtained better accuracy than the supervised systems in two narrow-domain documents, one about computer science, and one about art.

Chapter 5

Conclusions

The investigations carried out led to the following contributions:

- A complete Question Answering system has been developed. This system is able to answer questions in Spanish, Italian and French with an accuracy comparable with the current state-of-the-art systems, as demonstrated by the results obtained at recent CLEF competitions. The system demonstrated also its fault tolerance capabilities when the input was constituted by speech transcription or a low quality translation of the original question.
- We were able to assess that the translation errors affect the accuracy more than speech transcription errors in Question Answering.
- A comparison of off-the-shelf web translation systems was carried out. We also developed a technique based on web counts and the information entropy in order to select the best translation among multiple options. The obtained results show that this technique has to be improved in order to present an advantage with respect to the selection of a specific translator.
- Novel index expansion techniques based on the WordNet ontology for Geographical Information Retrieval have been investigated and applied in actual GIR tasks, obtaining good results in the case of Index Term Expansion.
- The use of Wikipedia in the Question Answering task was investigated, discovering that it is necessary to combine different editions (intended as language variants) of Wikipedia in order to obtain good coverage for the questions. In other words, a question in Spanish about an event of Italian politics will be found in the Italian edition of Wikipedia. This is an important result that demonstrates the need of developing cross/multi-language IR and QA systems.
- A novel WSD approach based on Conceptual Density was developed, and subsequently succesfully applied to the geographical domain for the disambiguation of toponyms.

- We investigated the use of Fuzzy Borda voting for the integration of different WSD techniques. Our investigations showed that the integration of knowledge-based WSD techniques can obtain better results than supervised methods in narrow-domain texts.

5.1 Research Papers produced

Journals:

Buscaldi, D., Rosso, P. A conceptual density-based approach for the disambiguation of toponyms. To appear in International Journal of Geographical Information Systems (late 2007: electronic edition, beginning of 2008: print).

Rosso P., Masulli F., Buscaldi D. Un Mtodo Automtico para la Desambiguacin Lxica de Nombres. In: Revista Colombiana de Computacin, UNAB, pp. 57-64.

Papers published in Springer (LNCS or LNAI):

Gomez J.M., Buscaldi D., Bisbal E., Rosso P., Sanchis E. QUASAR: The Question Answering System of the Universidad Politecnica de Valencia. In: CLEF 2005 Proceedings. Springer Verlag, LNCS(4022), Vienna, Austria, 2006.

Buscaldi D., Rosso P., Sanchis E. Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In: CLEF 2005 Proceedings. Springer Verlag, LNCS(4022), Vienna, Austria, 2006.

Buscaldi D., Juan A., Rosso, P., Alexandrov, M. Sense Cluster-based Categorization and Clustering of Abstracts. Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2006, Springer Verlag, LNCS(3878), pp.547-550, Mexico City, Mexico, 2006.

Buscaldi D., Rosso P., Pla F., Segarra, E., Sanchis E. Verb Sense Disambiguation using Support Vector Machines: impact of WordNet-extracted Features. Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2006, Springer Verlag, LNCS(3878), pp.192-195, Mexico City, Mexico, 2006.

Rosso P., Montes M., Buscaldi D., Pancardo A., Villasenor A. Two Web-based Approaches for Noun Sense Disambiguation. In: Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2005, Springer Verlag, LNCS (3406), Mexico D.F., Mexico, pp. 261-273, 2005.

Buscaldi D., Rosso P., Montes M. Context Expansion with Global Keywords for a Conceptual Density-Based WSD. In: Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2005, Springer Verlag, LNCS (3406), Mexico D.F., Mexico, pp. 257-260, 2005.

Buscaldi D., Rosso P., Masulli F. Integrating Conceptual Density with WordNet Domains and CALD Glosses for Noun Sense Disambiguation. In: España for Natural Language Processing, ESTAL-2005, Springer Verlag, LNAI (3230), Alicante, Spain, pp. 267-276.

Rosso P., Masulli F., Buscaldi D., Pla F., Molina A. Automatic Noun Sense Disambiguation. Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2003, Springer Verlag, LNCS (2588), pp. 273-276.

Conference proceedings:

Buscaldi, D., Rosso, P. A comparison of methods for the Automatic Identification of Locations in Wikipedia. Proc. of GIR 2007 Worskhop, CIKM 2007, Lisboa, Portugal, 2007.

Rosso, P., Buscaldi, D., Iskra, M. Web-based Selection of Optimal Translations of Short Queries. SEPLN, Revista no.38 (Abril 2007) pp. 49-53 ISSN: 1135-5948, 2007.

Buscaldi, D., Rosso, P. UPV-WSD : Combining different WSD Methods by means of Fuzzy Borda Voting. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) pp. 434-437, Prague, Czech Republic, 2007.

Buscaldi, D., Rosso, P., Peris, P. Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval. Proc. of the 3rd GIR Workshop, SIGIR'06, Seattle, WA, U.S.A., 2006.

Buscaldi D., Rosso P., Sanchis E. WordNet as a Geographical Information Resource. In: 3rd Global WordNet (GWN 2006) conference proceedings, Cheju, S.Korea, 2006.

Buscaldi, D., Rosso, P. Mining Knowledge from Wikipedia for the Question Answering task. Proc. of the LREC 2006, Genova, Italy, 2006.

Buscaldi, D., Rosso, P. A Naive bag-of-words approach to Wikipedia QA. CLEF 2006 Working notes, Alicante 20-22 September, C.Peters Ed, 2006.

Buscaldi, D., Rosso, P., Sanchis, E. WordNet-based Index Terms Expansion for Geographical Information Retrieval. CLEF 2006 Working notes,

Alicante 20-22 September, C.Peters Ed., 2006.

Buscaldi, D., Gomez, J.M., Rosso, P., Sanchis, E. The UPV at QA@CLEF 2006. CLEF 2006 Working notes, Alicante 20-22 September, C.Peters Ed., 2006.

Sanchis, E., Buscaldi, D., Grau, S., Hurtado, L., Griol, D. Spoken QA based on a Passage Retrieval Engine. SLT 2006, Aruba 10-13 December 2006.

Gomez J.M., Bisbal E., Buscaldi D., Rosso P., Sanchis E. Monolingual and Cross-language QA using a QA-oriented Passage Retrieval System. In: CLEF 2005 Working Notes. 21-23 September, Vienna, Austria C. Peters (Ed.), 2005

Buscaldi D., Rosso P., Sanchis E. A WordNet-based Query Expansion method for Geographical Information Retrieval. In: CLEF 2005 Working Notes. 21-23 September, Vienna, Austria C. Peters (Ed.), 2005.

Gomez J.M., Buscaldi D., Bisbal E., Sanchis E., Rosso P. A Multilingual Question Answering System using an n-grams based Passage Retrieval Engine. In: 2nd Indian Int. Conf. on Artificial Intelligence, Phrasad Ed., Hyderabad, India.

Buscaldi D., Rosso P., Masulli F. The upv-unige-CIAOSENSO WSD System. In: Workshop Senseval-3, Int. Conf. Association of Computational Linguistics (ACL), Barcelona, Spain, pp. 77-82, 2004.

Buscaldi D., Montes M., Rosso P. Web-based WSD using Adjective-Noun pairs. In: Workshop on Lexical resources and the Web for Word Sense Disambiguation, Int. Conf. IBERAMIA, Puebla, Mxico, pp. 89-96, 2004.

Pancardo A., Montes M., Rosso P., Buscaldi D., Villasenor L. Desambiguacin Lxica de Sustantivos usando la Web. In: Workshop on Lexical resources and the Web for Word Sense Disambiguation, Int. Conf. IBERAMIA, Puebla, Mxico, pp. 118-122, 2004.

Calcagno L., Buscaldi D., Rosso P., Gmez J.M., Masulli F., Rovetta S. Comparison of Indexing Techniques based on Stems, Synsets, Lemmas and Term Frequency. In: Workshop Red Temtica en Tecnologa del Habla, Valencia, Spain, pp. 171-176, 2004.

P. Rosso, F.Masulli, D.Buscaldi. Word Sense Disambiguation using Conceptual Distance, Frequency and Gloss. In: Int. Conf. on Natural Language Processing and Engineering Knowledge, IEEE Press, Beijing, China, pp. 120-125, 2003.

Buscaldi D., Guerrini G., Mesiti M., Rosso P. Tag Semantics for the Retrieval of XML Documents. In: Workshop on Conceptual information retrieval and clustering of documents, International Symposium on Information and Communication Technologies, ACM Conf. Series, Dublin, Ireland, pp. 280-285.

Rosso P., Masulli F., Buscaldi D. Word Sense Disambiguation with and without Supervision. In: XI Int. Congress of Computation, Mexico D.F., Mexico, pp. 531-540.

Bibliography

[1] R. M. Aceves-Pérez, L. Villaseñor-Pineda, and M. Montes. Using N-gram Models to Combine Query Translations in Cross-Language Question Answering. *Lecture Notes in Computer Science, CiCLing 2006 Proceedings*, 3878:453–457, 2006.

[2] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *16th conference on computational linguistics (COLING '96)*, pages 16–22, Copenhaghen, Denmark, 1996.

[3] R. Ahn, B. Alex, J. Bos, T. Dalmas, J. L. Leidner, and M. B. Smillie. Cross-lingual question answering with qed. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004)*, Bath, UK, 2004.

[4] L. Aunimo, R. Kuuskoski, and J. Makkonen. Cross-language question answering at the university of helsinki. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004)*, Bath, UK, 2004.

[5] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of CICLing 2002*, pages 136–145, London, UK, 2002. Springer-Verlag.

[6] K. Bo-Yeong, K. Hae-Jung, and L. Sang-Lo. Performance analysis of semantic indexing in text retrieval. In *CICLing 2004, Lecture Notes in Computer Science, Vol. 2945*, Mexico City, Mexico, 2004.

[7] E. Brill, J. Lin, M. Banko, S. T. Dumais, and A. Y. Ng. Data-intensive question answering. In *Text REtrieval Conference*, 2001.

[8] D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 2007. Accepted, to be published.

[9] D. Buscaldi, P. Rosso, and F. Masulli. The upv-unige-CIAOSENSO WSD System. In *Proc. of Senseval-3 Workshop*, Barcelona (Spain), July 2004. ACL.

[10] D. Buscaldi, P. Rosso, and F. Masulli. The upv-unige-ciaosenso wsd system. In *Proceedings of the Senseval-3 Workshop*, pages 77–82, Barcelona, Spain, 2004. The Association for Computational Linguistics.

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

[12] C. Callison-Burch and R. Flournoy. A program for automatically selecting the best output from multiple translation engines. In *Proc. of the VIII Machine Translation Summit*, Santiago de Compostela, Spain, 2001.

[13] G. Di Nunzio, N. Ferro, G. J. Jones, and C. Peters. Ad hoc track overview. In *CLEF 2005 Working Notes*, Vienna, Austria, 2005.

[14] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, New York, NY, USA, 2002. ACM Press.

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

[16] G. Fu, C. Jones, and A. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *Proceedings of the ODBASE 2005 conference*, 2005.

[17] E. Garbin and I. Mani. Disambiguating toponyms in news. In *conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT05)*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[18] J. García Lapresta and M. Martínez Panero. Borda Count Versus Approval Voting: A Fuzzy Approach. *Public Choice*, 112(1-2):167–184, 2002.

[19] F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough. Geoclef: the clef 2005 cross-language geographic information retrieval track. In *Working notes for the CLEF 2005 Workshop (C.Peters Ed.)*, Vienna, Austria, 2005.

[20] B. Green and A. Wolf. Baseball:an automatic question-answerer. Technical report, MIT, Massachusetts Institute of Technology, 1960. AD0257778.

[21] M. A. Greenwood. Using pertainyms to improve passage retrieval for questions requesting information about a location. In *SIGIR*, 2004.

[22] S. Harabagiu. Open-domain voice-activated question answering. In *COLING2002*, 2002.

[23] U. Hermjakob. Parsing and question classification for question answering. In *Proceedings of the ACL 2001 Workshop on Open-Domain Question Answering*, pages 17–22, 2001.

[24] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Lin. Question answering in webclopedia. In *The Ninth Text REtrieval Conference*, 2000.

[25] N. Ide and J. Véronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[26] V. Jijkoun, G. Mishne, M. de Rijke, S. Schlobach, D. Ahn, and K. Mueller. The university of amsterdam at qa@ clef 2004. In C. Peters, editor, *CLEF 2004 Working Notes*, Bath, UK, 2004.

[27] B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner, and A. Wilcox. External knowledge sources for question answering. In *Text REtrieval Conference 2005*, 2005.

[28] H. Kucera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.

[29] J. Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Research and Development in Information Retrieval*, pages 181–190, 1993.

[30] S. Landes, C. Leacock, and R. Tengi. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press, Cambridge, MA, 1998.

[31] S. Larosa, M. Montes-y Gómez, P. Rosso, and S. Rovetta. Best Translation for an Italian-Spanish Question Answering System. In *Proc. Of Information Communication Technologies Int. Symposium (ICTIS)*, Tetuan, Morocco, 2005.

[32] D. Laurent, P. Séguéla, and S. Nègre. Cross lingual question answering using qristal for clef 2005. In *CLEF 2005 Working notes*, 2005.

[33] J. L. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval, SIGIR 2004*, Sheffield, UK, 2004.

[34] J. L. Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417, July 2006.

[35] M. Lesk. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *ACM SIGDOC Conference*, pages 24–26. ACM Press, 1986.

[36] M. Light, G. S. Mann, E. Riloff, and E. Breck. Analyses for elucidating current question answering technology. *Nat. Lang. Eng.*, 7(4):325–342, 2001.

[37] J. Lin. The web as a resource for question answering: Perspectives and challenges. In *Language Resource and Evaluation Conference (LREC 2002)*, Las Palmas, Spain, 2002.

[38] L. V. Lita, W. A.Hunt, and E. Nyberg. Resource analysis for question answering. In *ACL 2004 Proceedings*. Association of Computational Linguistics, July 2004.

[39] X. Liu and W. Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, 2002.

[40] B. Magnini and G. Cavaglià. Integrating Subject Field Codes into WordNet. In *Proc. of the 2nd LREC Conference*, pages 1413–1418, Athens, Greece, 2000.

[41] B. Magnini, M. Negri, R. Prevete, and H. Tanev. Multilingual question/answering: the DIOGENE system. In *The 10th Text REtrieval Conference*, 2001.

[42] B. Magnini, M. Negri, R. Prevete, and H. Tanev. Is it the right answer? exploiting web redundancy for answer validation. In *ACL*, pages 425–432, 2002.

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

[44] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, New York, USA, 2003.

[45] G. Neumann and B. Sacaleanu. Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004)*, Bath, UK, 2004.

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

[47] M. Pasca and S. Harabagiu. The informative role of wordnet in open-domain question answering. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 138–143, Pittsburgh, PA, USA, 2001. Carnegie Mellon University.

[48] R. Purves and C. B. Jones. *Computers, Environment and Urban Systems*, volume 30, chapter Geographic Information Retrieval (GIR), pages 375–377. Elsevier, July 2006.

[49] I. Roberts and R. J. Gaizauskas. Data-intensive question answering. In *ECIR*, volume 2997 of *Lecture Notes in Computer Science*. Springer, 2004.

[50] P. Rosso, D. Buscaldi, and M. Iskra. Web-based selection of optimal translations of short queries. *Sociedad Espaola para el Procesamiento del Lenguaje Natural (SEPLN)*, 38:49–52, 2007.

[51] P. Rosso, E. Ferretti, D. Jiménez, and V. Vidal. Text categorization and information retrieval using wordnet senses. In *CICLing 2004, Lecture Notes in Computer Science, Vol. 2945*, Mexico City, Mexico, 2004.

[52] P. Rosso, F. Masulli, D. Buscaldi, F. Pla, and A. Molina. Automatic noun sense disambiguation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 4th International Conference*, volume 2588 of *Lecture Notes in Computer Science*, pages 273–276. Springer, Berlin, 2003.

[53] M. Sanderson and J. Kohler. Analyzing geographic queries. In *proceedings of Workshop on Geographic Information Retrieval (GIR04)*, 2004.

[54] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.

[55] Y. Takehide, S. Munehiko, and K. Yasuyuki. Trial production of a voice input question answering system using the bth parser. *Joho Shori Gakkai Shinpojiumu Ronbunshu*, 99(4):63–64, 1999. In Japanese.

[56] A. Vallin, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the clef 2005 multilingual question answering track. In *CLEF 2005 Proceedings*, 2005.

[57] S. Vazquez, R. Romero, A. Suarez, A. Montoyo, M. García, M. Martin, M. García, A. Ureña, D. Buscaldi, P. Rosso, A. Molina, F. Pla, and E. Segarra. The R2D2 Team at SENSEVAL-3. In *Proc. of Senseval-3 Workshop*, 2004.

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

[59] E. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the ACM SIGIR 1994*, 1994.

[60] W. A. Woods. Lunar rocks in natural english: Explorations in natural language question answering. *Lingusitic Structures Processing*, 5:521–569, 1977.

[61] H. Yang and T. Chua. The integration of lexical knowledge and external resources for question answering. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2002)*, 2002.

[62] X. Zhu and R. Rosenfeld. Improving trigram language modeling with the World Wide Web. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.