Where in the Wikipedia is that answer? The XLDB at the GikiCLEF 2009 task

Nuno Cardoso[†], David Batista[†], Francisco J. Lopez-Pellicer[‡] and Mário J. Silva[†]
[†]University of Lisbon, Faculty of Sciences, LaSIGE
[‡]University of Zaragoza, Centro Politécnico Superior, Spain
{ncardoso, dsbatista}@xldb.di.fc.ul.pt, fjlopez@unizar.es, mjs@di.fc.ul.pt

Abstract

GikiCLEF focused on the evaluation of the reasoning capabilities of systems to provide right answers for geographically-challenging topics. As we did not have previous experience in question answering, we participated in GikiCLEF with the goal of understanding best practices in extracting answers from documents though a hands-on experience. We developed a prototype that used DBpedia and the Portuguese Wikipedia as raw knowledge resources, and created a new world geographic ontology, also derived from the Wikipedia, for supporting geographic reasoning. The prototype was not ready to produce automatic runs at the submission deadline, but we observed that the best results we could aspire with the devised approach would be under our initial expectations. Wikipedia and DBpedia information coverage and location revealed to be much different from what we were initially expecting.

We learned that when planning on improving a GIR system with modules aimed to reason over the query before stepping into the retrieval procedure, such modules must be specifically crafted around the used raw knowledge resources, as they will shape the extraction approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

GikiCLEF, GeoCLEF, Geographic IR, Geographic Ontology, Question Answering, Information Extraction, Wikipedia Mining, DBpedia

1 Introduction

We are researching methods, algorithms and a software architecture for geographic information retrieval (GIR) systems since 2005 [14]. As GeoCLEF debuted in 2005, we seized each evaluation task to focus on several problems within our own GIR approach. Throughout the four editions of the GeoCLEF evaluation task [4, 5, 6, 10], we have focused on:

- Identification of a geographic weighting scheme that approximates the geographic similarity between the geographic criteria in the queries (*query scopes*) and the geographic area of interest of the documents (*document scopes*), and combine such weight scheme with term weighting scores to generate a global ranking measure for the retrieval process;

- Development of a geographic knowledge base that models the geographic domain, capable of generating geographic ontologies that can be used for geographic reasoning by GIR components;
- Automatic extraction of geographic evidence from documents to compute geographic signatures of documents, that is, document surrogates that describe their geographic area of interest.
- capture geographic criteria from query strings, if they exist, and perform ontology-driven query reformulation for the geographic terms, to better define the query scopes.

GeoCLEF adopted the TREC ad-hoc retrieval evaluation methodology [15], which lead us to focus mainly on the retrieval and ranking part of the GIR process. For instance, our GeoCLEF 2008 participation included a thorough optimisation step on the BM25 weighting scheme, within our experiments over the best term/geographic index weight ratio and its impact on overall GIR performance [6].

With GikiCLEF, our evaluation goals shifted considerably: the task evaluates straight answers, not document lists, encouraging a better understanding of the topics (given as questions) and a careful reasoning of the answers. That lead us to focus on other tasks that have been somehow overlooked in previous GeoCLEF participations.

For GikiCLEF, we set the following goals:

- Devise a new question analyser module that models the key concepts on the information need as formulated in the questions into workable objects, and uses Wikipedia and DBpedia to search for more details about such key concepts, verify conditions, comprehend the geographic restrictions included in the topic, and finally reason the correct answers;
- Develop a new world geographic ontology, Wiki WGO 2009, derived from Wikipedia and organised according to an improved version of our geographic knowledge model.

The rest of the paper is organised as follows: Section 2 overviews the prototype developed for our GikiCLEF participation. Section 3 summarises the problems faced when devising answers to the provided GikiCLEF topics. Section 4 presents the post-hoc changes made on the prototype. Section 5 concludes the paper.

2 GikiCLEF approach

Figure 1 illustrates our approach for devising answers to GikiCLEF topics, consisting on a question analyser module and its knowledge resources. The GikiCLEF topics were parsed by PALAVRAS, a Portuguese PoS tagger [3], before being input to the question analyser.

As topics are in the form of a question, the initial task performed by the *question interpreter* (QI) is to convert questions into object representations (*question objects*). From there, the *question reasoning* (QR) reasons on the best strategy to obtain the correct answers, which may include information extraction steps over the raw knowledge resources. The QR reasoning approaches were derived from manual experiments conducted in the past year, while participating in the GikiP pilot task [13]. The QR output is the answers and their justifications, which are converted into the GikiCLEF run format.

The question analyser explores three knowledge resources: i) Wikipedia tagged documents, ii) the DBpedia v3.2 dataset, and iii) the Wiki WGO 2009 geographic ontology.

The Portuguese Wikipedia documents, as given by the Portuguese piece of the GikiCLEF collection, were tagged on-demand by HENDRIX, a named entity recognition system that we have been developing. As it was not possible to tag the whole Portuguese collection in time, we used an on-the-fly tagging strategy, which considerably limited some of the question reasoning approaches.

The DBpedia v3.2 dataset (`http://wiki.dbpedia.org/Downloads32`) consists on N-Triple files with information extracted from the English Wikipedia pages of October 2008. DBpedia is a community effort to extract information from Wikipedia and make it available on the Web in a structured, machine-friendly way [1]. The version 3.2 of the DBpedia dataset includes an ontology with 170 classes and 940 properties, and over 882,000 entities classified according to this ontology (`http://wiki.dbpedia.org/Ontology`).

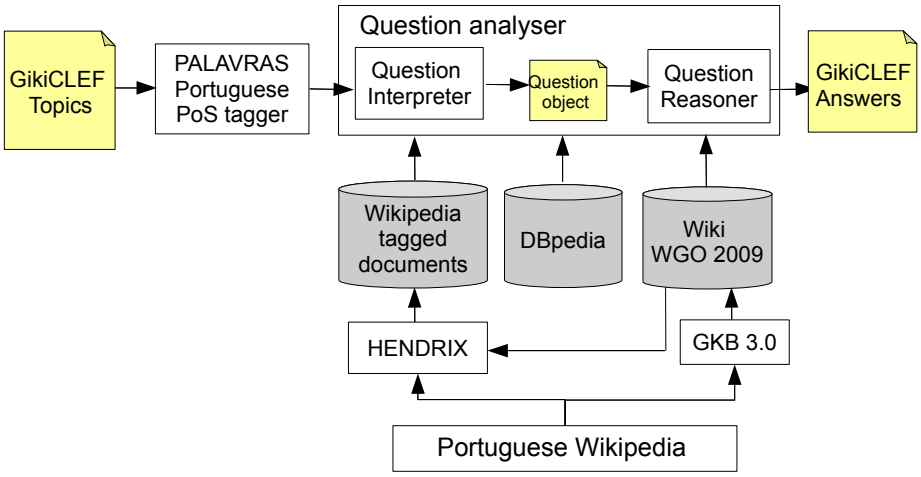

Figure 1: Overview of our approach on GikiCLEF.

The Wiki WGO 2009 geographic ontology, built with our geographic knowledge base, GKB 3.0, provided geographic information to the question analyser and to HENDRIX.

We now detail the question interpreter and question reasoner modules, the HENDRIX tagging system and the Wiki WGO 2009 generation process.

2.1 Question interpreter

The question interpreter (QI) converts a natural language question into a machine-interpretable object representing that question (the question object). We designed the QI so that the key elements included in the question object can be grounded to DBpedia resources, thus making it easier to interoperate among modules and knowledge resources.

The question object is composed of the following elements:

Subject, the entity that defines the type of expected answers. The subject can be grounded as i) a DBpedia resource that has a property `rdf:type` for a value `skos:Concept`, ii) a DBpedia ontology class, or iii) a semantic classification as defined in the HAREM categorization [12], on this preferential order.

Conditions, a list of criteria that filter the answers list. Each condition is composed of i) a DBpedia ontology property, ii) an operator and iii) a DBpedia resource.

Expected Answer Type (EAT), used to define properties that the final set of answers must have.

The question object is generated by applying a set of pattern rules over the PoS tags and the terms of the questions. Take for instance the question "Which Romanian writers were born in Bucharest?", the QI would perform as follows (illustrated in English):

Ground the subject: the first set of pattern rules detect the terms that define the subject; in the given example, the rule "`Which <[name]+ [adjective]*> [verb]+`" captures the terms "Romanian writers", which were previously tagged by the PoS tagger as name and adjective, right after the 'Which" term. Afterwards, these terms are mapped to `http://dbpedia.org/resource/Category:Romanian_writers`, which is a DBpedia resource derived from the corresponding Wikipedia's category page, and has a property `rdf:type` with the value `skos:Concept`.

Ground the EAT: after the subject is grounded, pattern rules determine the EAT according to the question type and the subject type. For "Which X" questions as the given example, the EAT is assigned to the subject, that is, the answers must have a `skos:subject` property with the subject's resource as the value,

meaning that all the answers must be about Romanian writers. Note that, for instance, another rule, "How many X", grounds the EAT as a number instead, thus telling to the QR that the answer should be the size of the final answer list, or a statement about a quantity property.

If the subject cannot be mapped to a DBpedia resource, it is mapped to a DBpedia ontology class or a HAREM category, which may be assigned to the EAT. Suppose that the QI failed to ground "Romanian writers" to a DBpedia resource; using a simple term mapping hash, the term "writers" triggers a mapping of the EAT to the `http://dbpedia.org/ontology/Writer` class. Lastly, if the QI cannot map to a DBpedia ontology class, the generic category/type `PERSON/INDIVIDUAL` is used.

Ground conditions: in the given example, there is a condition that filters the correct answers from an initial list of Romanian writers to those who were born in the city of Bucharest. Pattern rules triggered by the terms "were born in Bucarest" generate a new condition with a property grounded to the `dbpedia-owl:birthPlace` property (`http://dbpedia.org/ontology/birthplace`, an operator `IS` (the default operator), and a referent entity grounded to the DBpedia resource `http://dbpedia.org/resource/Bucharest`.

For questions where the condition consists on a numeric clause over a property value, as in "mountains higher than 2000 meters", the condition operator can be grounded to different values, such as `GREATER`, `LESSER`, `BETWEEN`, so that the QR step can perform a numeric test instead. This question object model can encompass more than one condition, as in "Which Romanian writers were born in Bucharest in the 20th century", requiring only the proper pattern rules to capture both conditions.

2.2 Question reasoner

The question reasoner (QR) processes the question object. Depending on the elements present in the question object, the QR task is to decide on the best strategy to get the answers. The QR strategy consists on a pipeline of SPARQL queries made to the knowledge resources, to obtain and validate answers and their justifications.

In the given example, the question object has a EAT given by the DBpedia resource `Category:Romanian_writers`, and a single condition described by the DBpedia property `dbpedia-owl:birthPlace`, an operator `IS` and a referent entity given by DBpedia resource `http://dbpedia.org/resource/Bucharest`. For this type of question objects, the QR strategy consists on issuing the following SPARQL query to the DBpedia dataset:

```
SELECT ?RomanianWriters WHERE {
 ?RomanianWriters skos:subject <http://dbpedia.org/resource/Category:Romanian_writers> .
 ?RomanianWriters dbpedia-owl:birthplace <http://dbpedia.org/resource/Bucharest>
}
```

Using DBpedia's SPARQL endpoint in `http://dbpedia.org/sparql`, for the current DBpedia 3.3 dataset, there are two answers, `http://dbpedia.org/resource/Eugen_Filotti` and `http://dbpedia.org/resource/Mihail_Fărcăşanu`.

2.3 Wiki WGO 2009 ontology and GKB 3.0

Our work around geographic knowledge bases dates from 2005, when we released the first version of geographic ontologies produced with GKB [7]. The new version of GKB, 3.0, was developed for this participation in GikiCLEF and included a major review of its metamodel. In summary, GKB 3.0 allows the generation of ontologies using linked data (for instance, SKOS properties [11]), and loosening the rigid use of feature and feature types as resource descriptors, embracing a much expressive set of properties as given, for instance, by Wikipedia categories.

We created a geographic ontology in RDF/OWL format – the Wiki WGO 2009 – using the Portuguese Wikipedia as the sole information source, as a part of the validation procedure for the GKB 3.0 model. The Wiki WGO 2009 ontology was loaded in a triple-store server, so that the QR module could issue SPARQL queries regarding the geographic similarity between answers and query scopes.

To illustrate the role of Wiki WGO 2009 in our overall GikiCLEF approach, consider the GikiCLEF topic nr. 1, "List the Italian places where Ernest Hemingway visited during his life." The EAT is grounded to the DBpedia ontology class `http://dbpedia.org/ontology/Place`, and the conditions list includes a condition that the answered features must be part of the country of Italy (let's focus on this condition alone). The QR can test this conditions on candidate answers by querying the WGO Wiki 2009 ontology.

For instance, if a candidate answer is "Rome", we can obtain a list of geographic feature types with the following SPARQL query:

```
SELECT ?featTypeLavel WHERE {
    ?feat skos:prefLabel "roma"@pt .
    ?feat rdf:type ?featype .
    ?featype  skos:prefLabel ?featTypeLabel
}
```

For this SPARQL example, Wiki WGO 2009 returns the following results: "catholic pilgrimage sites", "lugares de peregrinação cristã", "capital of a political entity", "italy geography stubs" and "itália", meaning that it's a capital, and somehow is related to Italy.

The WGO Wiki 2009 uses the `http://purl.org/dc/terms/isPartOf` property to represent the `PartOf` relationship between geographic concepts. If there is uncertainty about the relationship between *Rome* and *Italy*, we can issue a SPARQL query to list all properties that relate "Rome" to "Italy" in the ontology (we omit the grounding clauses from labels to ontology resources, to unclutter the example):

```
SELECT ?featType WHERE {
    ?feat skos:prefLabel "roma"@pt .
    ?feat2 skos:prefLabel "itália"@pt .
    ?feat ?featType ?feat2
}
```

If the response includes the `http://purl.org/dc/terms/isPartOf` value, then there is a direct `partOf` relationship, and the candidate answer "Rome" passes the condition.

2.4 HENDRIX

HENDRIX (**H**endrix is an **E**ntity **N**ame **D**esambiguator and **R**ecognizer for **I**nformation E**x**traction) is a named entity recognition (NER) system developed in-house, based on Minorthird [8]. It makes use of Minorthird's Conditional Random Fields (CRF) implementation, a supervised machine learning technique for text tagging and information extraction [9]. HENDRIX uses the Wiki WGO 2009 ontology to detect relations between named entities tagged as geographic locations.

HENDRIX was trained to recognise places, organisations, events and people. The training phase used the HAREM's Golden Collections (GC), which are manually-tagged collections of Portuguese documents used in HAREM's NER evaluation contests [12]. There are three GCs available; two were used for the training phase and one to evaluate HENDRIX's entity recognition performance. The results of the evaluation are shown in Table 1.

HENDRIX has only one CRF model for extracting NEs within the four different types of entities described above. The performance of HENDRIX as used in the GikiCLEF evaluation is very unsatisfactory. We observed that there is a significative amount of named entities that are correctly extracted but incorrectly

Entity	Precision	Recall	F-Measure
PERSON	0.5915	0.4095	0.4840
PLACE	0.4590	0.5006	0.4789
EVENT	0.3281	0.2515	0.2847
ORGANIZATION	0.4464	0.4783	0.4618

Table 1: HENDRIX's NER performance results

categorised, which prompt us to evaluate how a simpler CRF model trained separately for each type of entities would perform.

The HENDRIX trained model would then be used to extract named entities from Portuguese Wikipedia articles. All the extracted entities tagged as a `PLACE` were afterwards used for detection of semantic relationships among the entities, using the Wiki WGO 2009 ontology. HENDRIX outputs a summary of each Wikipedia article with the tagged information and detected relationships.

3 GikiCLEF results

The prototype was not finished in time to be able to generate unsupervised runs before the GikiCLEF submission deadline, so we submitted a single run generated with strong manual supervision, using the reasoning strategies described above. Although the score of the run does not measure the prototype's performance, it points to many weaknesses of the overall approach, and how well does it cover the necessary reasoning steps to answer the evaluation topics.

To better understand the causes of failure in the interaction between the question analyser modules and the knowledge resources, we took all the correct and justified answers from GikiCLEF assessments and browsed the current Wikipedia articles to search for the locations in Wikipedia pages where the answer can be found, and in what languages (we assume that the changes made on Wikipedia, from the June 2008 snapshots used for the GikiCLEF collection, to the June 2009 articles are not significantly relevant.) We observed the following issues:

DBpedia's limited coverage of answers: most of the information on DBpedia datasets is extracted from Wikipedia's infoboxes, that is, the template tables used in most Wikipedia articles to summarise relevant properties for the entity. While questions regarding people's birthplaces, country's presidents or mountain heights are likely to be found in infoboxes, for questions regarding places that someone visited, or countries who published a certain book, the answers are more likely to be found only in the body of the Wikipedia articles.

For the English Wikipedia, which had answers and justifications for 45 topics, we observed that the Wikipedia infoboxes were useful in finding answers in only 9 of those topics (nr. 11, 12, 25, 29, 31, 34, 38, 42 and 50). For the Portuguese Wikipedia, there are 6 topics that could be correctly answered using infobox information. Consequently, all the QR strategies that relied on SPARQL queries over the DBpedia information, were unsuccessful in most of the topics.

Heterogeneous answer distribution over Wikipedia languages: as we are mostly interested in the Portuguese language, we wondered how many topics could have been answered with at least one correct answer using solely the Portuguese Wikipedia (and without resorting to any language links). From the 47 GikiCLEF topics that had at least one correct answer, the Portuguese Wikipedia was self-sufficient to answer 25 topics (53%).

If we consider language links, that is, if we had an information extraction (IE) system that could start from Portuguese pages, be able to process other language's equivalent pages and then return Portuguese Wikipedia pages as answers justified in other language's pages, the number of topics that could be answered with Portuguese pages would be 37. Nonetheless, there are 10 topics that do not have Portuguese Wikipedia pages that represent correct answers. If we count answers instead of topics, and within a universe of 243 correct answers assessed by GikiCLEF organisers, 111 of them had no Portuguese Wikipedia page equivalent (45.7%), 65 of them had an answer page but no justification was found on the Portuguese Wikipedia (26.7%), leaving only 67 answers – that's 27.6% – as the maximum recall we could achieve in using only the Portuguese Wikipedia.

4 Post-GikiCLEF work

We were planning on performing an on-the-fly NER tagging on Wikipedia article's text, but we soon realized that it was too slow for any reasoning step, and that we should work on having a fully-tagged

Wikipedia corpus to work on IE approaches over Wikipedia. The GikiCLEF evaluation showed that we depended too much on DBpedia, and should better explore the text of Wikipedia, which could not be fully tagged in time to be of service to the question analyser during the GikiCLEF run submission period. This limited considerably our reasoning strategy options; for instance in the topic nr. 1, "List the Italian places where Ernest Hemingway visited during his life.", one strategy is capturing NE entities of type PLACE on the Ernest Hemingway's Wikipedia and test them for its geographic relationships against Italy, which require one tagged Wikipedia page; another strategy is to search for references of Ernest Hemingway on Wikipedia pages about Italian places, which requires that those pages were able to be readily selected and fully tagged. For the six correct answers found on English Wikipedia for this topic, only one (*Acciaroli*) was explicitly justified on Ernest Hemingway's page; the other answers (*Fossalta di Piave*, *Stresa*, *Torcello*, *Harry's Bar* and *Aveto Valley*) were justified on their own text.

This heterogeneously distribution of answers and justifications throughout the Wikipedia documents requires the development of a comprehensive knowledge repository to store all the information extracted from Wikipedia article texts.While the advantages and limitations on using SPARQL as the common knowledge interface and main actuator of our reasoning processes it is still unclear, the QR module could greatly benefit if the information extracted from Wikipedia article texts was made as readily accessible as DBpedia is.

Modeling such knowledge repository is only the initial step. Extracting such information in an automatic way within an acceptable accuracy ratio, represent it in RDF/N-Triple format, validate and curate the data and populate the knowledge repository is a whole different subject. Nevertheless, such knowledge database seems crucial if we want our GikiCLEF prototype to be able to successfully answer the topics in an unsupervised way.

4.1 Knowledge repository model

Right now, our post-GikiCLEF work is centered on developing a knowledge repository model with the following specifications:

Distinction between concepts and concept representations: a concept (for instance, the country of Portugal) can be designated with several names (*Portugal*, *República Portuguesa*, *Portuguese Republic*, etc), and a single name "Portugal" can designate several concepts (a country, a government, a football team, etc). The knowledge repository model, in a similar way as DBpedia and GKB, will store knowledge as a group of properties associated to grounded concepts, regardless of the names used to represent such information / concepts. Concepts will have URI identifiers (DBpedia URLs, when available), following the Linked Data recommendations [2].

Relations mapped to concepts: relations will be grounded to DBpedia ontology properties, and will be associated to pairs or concept identifiers. For instance, saying that the writer José Saramago was born in the year of 1922, implies that the relation is mapped to the `dbpedia-owl:birthYear` property, to the concepts of "José Saramago" as a person, and 1922 as a year.

Knowledge audit capability: as the knowledge repository will store mostly information extracted from Wikipedia article texts automatically, it will also have mechanisms to facilitate the validation and curation of such information, such as source documents, extraction pattern applied or confidence scores.

Linked data to Wiki WGO: the geographic concepts will be mapped to concepts in the Wiki WGO ontology, so that geographic and non-geographic reasoning can be performed seamlessly.

Allow SPARQL query endpoints: the model design should take in consideration that access to its information will be made from SPARQL queries, thus it must be able to be exported into N-Triple formats, loaded into triple-store servers, and accessed through HTTP requests.

5 Conclusions

Our participation in GikiCLEF 2009 was overall more enriching than disappointing. While we did not manage to build a working prototype that could generate unsupervised runs in time, we now have a clearer idea on what we need to have a GIR question analyser that can reason over user queries before stepping into the retrieval phase.

We focused our work on developing a question analyser module that used SPARQL queries over DBpedia and the Wiki WGO 2009, our geographic ontology, as a means to get answers to GikiCLEF topics, leaving the Wikipedia text as a backup resource that could be explored with shallow extraction methods, such as named entity recognition. We found out that mining Wikipedia for the answers is much more demanding than we initially expected, and that DBpedia's coverage on answering to GikiCLEF topics was much weaker than we thought.

Acknowledges

We would like to thank all GikiCLEF organisers. This work is partially supported by FCT (Portuguese research funding agency) for its LASIGE Multi-annual support, GREASE-II project (grant PTDC/EIA/73614/2006) and Nuno Cardoso's scholarship (grant SFRH/BD/45480/2008).

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Proceedings*, number 4825 in LNCS, pages 722–735. Springer, 2007.

[2] Tim Berners-Lee. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2009.

[3] Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, University of Aarhus, Aarhus, Denmark, November 2000.

[4] Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivian Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Languag e Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *LNCS*, pages 802–810. Springer, 2008.

[5] Nuno Cardoso, Bruno Martins, Leonardo Andrade, Marcirio Chaves, and Mário J. Silva. The XLDB Group at GeoCLEF 2005. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck a nd Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation F orum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, volume 4022 of *LNCS*, pages 997–1006. Springer, 2006.

[6] Nuno Cardoso, Patrícia Sousa, and Mário J. Silva. Experiments with Geographic Evidence Extracted from Documents. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thoma s Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross- Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, 2009.

[7] Marcirio Chaves, Mário J. Silva, and Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. In Carlos A. Heuser, editor, *20 Simpósio Brasileiro de Bancos de Dados (SBBD'2005)*, pages 40–54, Uberlândia, MG, Brazil, October 3-7 2005.

[8] William W. Cohen. Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data, http://minorthird.sourceforge.net. 2004.

[9] John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[10] Bruno Martins, Nuno Cardoso, Marcirio Chaves, Leon ardo Andrade, and Mário J. Silva. The University of Lisbon at GeoCLEF 2006. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-L anguage Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *LNCS*, pages 986–994. Springer, Setember 2007.

[11] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. SKOS Core: Simple Knowledge Organisation for the Web. In *DCMI '05: Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications*, pages 1–9. Dublin Core Metadata Initiative, 2005.

[12] Cristina Mota and Diana Santos. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2009.

[13] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, and Johannes Leveling an d Yvonne Skalban. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thoma s Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross- Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, 2009.

[14] Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding Geographic Scopes to Web Resources. *CEUS - Computers Enviroment and Urban Systems*, 30(4):378–399, 2006.

[15] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.