Getting geographical answers from Wikipedia: the GikiP pilot at CLEF

Diana Santos[1], Nuno Cardoso[2], Paula Carvalho[2], Iustin Dornescu[3], Sven Hartrumpf[4]
Johannes Leveling[4], Yvonne Skalban[3]
[1]Linguateca, Oslo node, SINTEF ICT (Norway)
[2]University of Lisbon, DI, LasiGE, XLDB, Linguateca (Portugal)
[3]Research Group in Computational Linguistics (CLG) at the University of Wolverhampton (UK)
[4]Intelligent Information and Communication Systems (IICS),
University of Hagen (FernUniversität in Hagen) (Germany)
Diana.Santos@sintef.no, ncardoso@xldb.di.fc.ul.pt, pqfcarvalho@gmail.com
i.dornescu2@wlv.ac.uk, Sven.Hartrumpf@FernUni-Hagen.de
Johannes.Leveling@FernUni-Hagen.de, yvonne.skalban@wlv.ac.uk

Abstract

This paper reports on the GikiP pilot that took place in 2008 in GeoCLEF. After providing motivation from both organizers and participants, it presents the task description, detailing topic choice and evaluation measures. Results are reported together with assessment difficulties and issues. Each participant system is described in detail, and the paper concludes with remarks on the current venue as well as ideas for improvements for future editions of GikiP or similar evaluation contests.

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; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids, Geographical information retrieval, Cross-lingual information retrieval, Wikipedia, German, Portuguese, English, Evaluation

1 Motivation

This paper introduces GikiP[1], an evaluation contest on retrieving information from Wikipedia [30] in the form of a list of answers (corresponding to articles) that have some geographical component. We start by reporting different kinds of motivation that led us to propose, organize or participate in GikiP:

[1]http://www.linguateca.pt/GikiP/

- the first is the often voiced dissatisfaction of GeoCLEF participants (and people involved in GIR in general) with continuing querying a collection of old newspapers – this has been a constant in GeoCLEF [5, 6, 19] breakout sessions. Although we believe that the main GeoCLEF task is interesting enough, and that there are valid user models for it, there is a wealth of other sources as well as other kinds of applications where geographical information can be brought to bear. Furthermore, variation in an evaluation campaign, which after all has as one of its goals to foster innovation, is certainly beneficial for the field.

- the second is the emergence of Wikipedia as an unavoidable resource for IR and NLP (becoming soon even more used than WordNet, as could for example be appreciated in this year's LREC [18]). In fact, Wikipedia's growth does not appear to slow down at present, with considerable content (more than 50%) in languages other than English [29], and it is on the top 10 of the most visited sites on the Web, according to [1] (for all this it appeals to the IR community as much as to the NLP one – if it still makes sense to separate the two).

- the third is our own interest in finding more realistic models for evaluation tasks, also in a multilingual context. Wikipedia is a truly multilingual resource, and it is not, as most other multilingual resources are, based on machine translation. Still, it has interesting alignment properties and metadata, which for example newspaper collections do not have.

Our view is that traditional evaluation tasks tend to create artificial divides where ultimately what is at stake is the satisfaction of user needs concerning access to information and knowledge through automatic means (or automatic helpers). It is not the form of the question or of the answer that should ultimately define the limits of what information access can offer.

Clearly, CLEF [20], NTCIR [16], and TREC [28] are undeniably extremely important for bringing progress and respectability to the IR and information access communities. Key ingredients are: (i) Separation of the teams who create the evaluation data from those who develop the systems, (ii) the proposal of challenging tasks to advance the state of the art, (iii) and attempting to measure progress from edition to edition.

However, the need to capitalize on the evaluation setup already created, as well as the upsurge of scientific communities around a particular task, may, in the long run, cause separation of otherwise similar concerns and interests, as is, in our opinion, what happens with QA@CLEF and GeoCLEF (or, for the sake of the argument, also WebCLEF and Ad-Hoc CLEF).

GikiP is an attempt to bring back together the two communities/tasks, by merging – albeit in a very specific context – the two forms of information request (questions or topics) and the two kinds of expected answers (factoids or documents), which are typically the hallmark of QA and IR, respectively.

Another concern of GikiP is to encourage multilingual and cross-lingual processing. In fact, in CLEF this has – understandably – not been a priority for participants. Even though the organizers take pains to treat all languages equally well, one might call CLEF (or some of CLEF tracks, at least) a set of monolingual evaluation campaigns in disguise. This is not a criticism. Rather, it reflects an important reality: In most research groups, most if not all resources are devoted to the processing of one's language (as it should be). Nevertheless, GikiP provides a task where it might be comparatively easy to satisfy other languages' needs (and/or make use of other languages), given that it has as target one of the most genuinely multilingual resources available: Wikipedia.

Let us acknowledge that CLEF has already witnessed an interesting pilot with Wikipedia: WiQA 2006 [15, 31, 14]. WiQA was unfortunately not continued, after a promising start with seven participants in three languages (Dutch, English and Spanish). Our guess is that this happened because it was too ambitious and had a too specific user model, namely to help create new Wikipedia pages.

On the contrary, GikiP has a very straightforward and understood task (answer open list questions), and a very broad intended user community: any user/reader of Wikipedia might be interested in asking questions to it and get a list of articles. Systems that can perform successfully in this task may help harnessing the wealth of information that is included in Wikipedia. (And in fact, GikiP could be just the beginning of a contest where one would also look for images [23] – cf. the concurrent pilot at ImageCLEF with its WikipediaMM task [13].)

But let us go back to the 2008 pilot in the remainder of this paper. We first describe the task in Section 2, with Section 3 devoted to topic choice, translation and assessment. Section 4 describes actual participation and results, while Sections 5 to 7 present each participant's approach in detail. Section 8 sums up and concludes.

2 The GikiP task

We defined the particular specific task to be solved by this year's participants:

> Find Wikipedia entries (i.e. articles) that answer a particular information need which requires geographical reasoning of some sort.

In order to guarantee a common evaluation ground, participants were required to use the Wikipedia collection(s) already used in the QA@CLEF main track[2], henceforth called the GikiP 2008 collections.

Fifteen topics were made available on the 2th June 2008 from the GikiP site, where eight example topics in the three languages had already been published. Below is topic GP4 in the English version:

```
<top lang="en">
<num>GP4</num>
<title>Which Swiss cantons border Germany?</title>
<description>Find which cantons are on the border of Switzerland with Germany.
</description>
</top>
```

Participants had 10 days to return the results, in the form of a list of Wikipedia articles, by providing their results as a list including the title of Wikipedia web pages. A (reduced) example of results for the GP4 topic is:

```
GP4
de/k/a/n/Kanton_Aargau.html
de/k/a/n/Kanton_Basel-Landschaft.html
de/k/a/n/Kanton_Basel-Stadt.html
de/k/a/n/Kanton_Zürich.html
en/a/a/r/Aargau.html
en/b/a/s/Basel-Land.html
en/c/a/n/Canton_of_Zurich.html
en/t/h/u/Thurgau.html
pt/a/r/g/Argóvia_(cantão).html
pt/b/a/s/Basiléia-Campo.html
pt/b/a/s/Basiléia-Cidade.html
pt/c/a/n/Canto_de_Zurique.html
```

2.1 Task definition

Only answers / documents of the correct type were requested (and therefore assessed as correct). In other words, if a topic concerned painters or scientists, results should be (lists of) names of people (painters and scientists), and not names of boats or countries. Conversely, if the question was about countries, the type of results should be places of the country type, and not wars or kings.

The maximum number of documents returned per topic was set to 100, but systems were strongly encouraged to try to return only the right ones (which, as the organization was at pains to emphasize, were typically much less than that number). Initially, we had decided on a maximum of two runs per system, but

[2] This was also a way to diminish the threshold for QA participants, who already had to process those collections anyway. However, in the end it apparently scared away other participants who had other Wikipedia versions at their disposal, and produced some problems for the organisers as well since some runs did provide out-of-collection answers.

Table 1: Topic titles in GikiP 2008

ID	English topic title
GP1	Which waterfalls are used in the film "The Last of the Mohicans"?
GP2	Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany?
GP3	Portuguese rivers that flow through cities with more than 150,000 inhabitants
GP4	Which Swiss cantons border Germany?
GP5	Name all wars that occurred on Greek soil.
GP6	Which Australian mountains are higher than 2000 m?
GP7	African capitals with a population of two million inhabitants or more
GP8	Suspension bridges in Brazil
GP9	Composers of Renaissance music born in Germany
GP10	Polynesian islands with more than 5,000 inhabitants
GP11	Which plays of Shakespeare take place in an Italian setting?
GP12	Places where Goethe lived
GP13	Which navigable rivers in Afghanistan are longer than 1000 km?
GP14	Brazilian architects who designed buildings in Europe
GP15	French bridges which were in construction between 1980 and 1990

later on we accepted up to six runs from the one group who suggested this alternative. And we believe that this, coupled with a restriction like "as long as the total number of different topic-document pairs does not exceed 1500 (i.e. the number of assessments needed for 1 dumb run that delivers 100 documents for each of the 15 topics)", would be a better task definition for further editions.

2.2 Evaluation score

Evaluation was devised in order to emphasize diversity and multilinguality, ensuring that the systems which were able to retrieve most cases and in most languages would be considered best by the GikiP scores.

We delivered (and assessed) topics in English, German, and Portuguese, and therefore, an additional bonus was computed for multilinguality, *mult*, which is 1, 2 or 3 depending on the number of languages tried out by the systems.

System results are thus evaluated according to the following formula $mult * N * N/total$ where *mult* rewards multilinguality, N is the number of correct hits, and $N/total$ is precision. The system's final score is given by the average of the individual scores.

3 Topics: selection, translation, and assessment

Table 1 presents the titles of all topics used in this first GikiP pilot. The topic description was generally a less condensed and more verbose version of the topic meaning, but would not add crucial information.

3.1 Topic choice

Considerable care was put in obtaining a balance between topics more naturally covered in each of the three GikiP languages, as well as providing also other topics equally alien, in principle, to any of the languages or cultures (such as those on French bridges, Afghan rivers or Polynesian islands). This is not necessarily a true description of what is or can be found in the particular Wikipedia language versions available, but this was a guideline for topic choice.

Although, by definition, a geographic flavour had to be associated, we also strived to create topics quite different from those used in the GeoCLEF main task – in order to cover higher levels of the typology described in [5]. In a way, GikiP is – as already emphasized – a way to explore further the goals of GeoCLEF by changing the document collections and also the kind of geographical reasoning required. It was therefore important not to repeat the same kind of topics with just another collection. In our view, this

Table 2: Topic size of GikiP 2008, only automatic runs

Topic	Results	Correct
GP1	5	1
GP2	31	7
GP3	28	8
GP4	79	21
GP5	69	21
GP6	36	7
GP7	90	33
GP8	49	2
GP9	49	17
GP10	53	2
GP11	35	23
GP12	51	25
GP13	9	4
GP14	60	6
GP15	18	2
Total	662	179

was one weakness of the QA task in 2007 [7], which continued with the same kind of questions and just enlarged the collections (by adding Wikipedia).

3.2 Topic translation

All topics were initially devised in Portuguese, and then roughly translated into English and from there to German. Then a thorough revision of the three versions took place. Although it was not intended, it appeared that some topics in Portuguese were hard to translate into the two Germanic languages, so we note this here, as a counterpart to the usual difficulty to translate some English topics into Portuguese (reported for example in [21]). One issue was the navigability and length, which we meant as independent features of a river, but which strongly conveyed the implication in English[3] that it was the length of the navigable stretch one was measuring.

Another issue was the difficulty of expressing something that overlapped temporally with a particular decade in English – bringing the awkward formulation "whose construction started, continued or ended in or between 1980 and 1990". Likewise, the expression of "the place of the plot" in German was not direct, either, and in fact "Welche Stücke Shakespeares spielen in Italien?" can also indicate that a theater company is doing a Shakespeare play in some part of Italy.

Of course we are not saying that either of these questions is impossible or even too difficult to express in German or English. We are just pointing out that they are probably more naturally come upon and formulated by Portuguese speakers. This is the sort of information that is interesting to amass in a cross-lingual context: some questions are easier to answer (and more natural to pose) in different languages.

3.3 Topic assessment

In order to provide an idea of the assessment work, we describe briefly the pool obtained, by listing, for each topic, the number of different answers from all participants, as well as number of the correct answers, in Table 2.[4]

As far as assessment matters are concerned, as usual some decisions had to be made: As a general rule, we considered as wrong answers all cases where the human assessor was not able, to the best of her ability, to verify the truth of the answer. For example, in topic GP14 (*Brazilian architects who designed buildings in Europe*), several Brazilian architects where no mention to works in Europe could be found were deemed

[3] And possibly also in German. Here, opinions diverge.

[4] Note that the answers themselves may not necessarily be different, they just need to correspond to a different Wikipedia article.

incorrect, because we expect that non-verifiable information is not useful for any user. If the user wanted to assess himself based on the information that only he had, then he would have asked for Brazilian architects only.

Of course, other issues were more difficult to assess. For example, does the mention that a particular river was *strategic for military operations* imply that it is navigable? – we assumed a negative answer in topic GP13 (*Which navigable rivers in Afghanistan are longer than 1000 km?*).

We acknowledge that our role as human assessors may not guarantee perfect knowledge, but this is in no way different from any other evaluation contests which involve human judgements.

Rather more interesting was the problem of different language versions having different answers (for example as far as population of the Tahaa island, for GP10 (*Polynesian islands . . .*) is concerned). Here, we assumed a very liberal procedure. If a positive answer could be found in any version, all of them were deemed correct.

Also, some answers to GP7 (*African capitals . . .*) referred to places that no longer exist or have changed name or status (such as *Salisbúria* or *Abidjan*). We considered them correct since no temporal restriction was explicitly mentioned in the topic, but this is obviously an issue that has to be better dealt with in real-life information access.

In fact, the temporal interaction is even more complex: if we take a closer look at the GP7 topic, three ways to interpret it arise: (i) a user might be interested in African capitals with a given population at any time (this is the broadest possible interpretation) or (ii) only with such population when they were capitals (even if they are no longer capitals), or even (iii) only cities with such population, even though no longer capitals (provided they had been once).

Yet another interesting issue also raised in connection with the GP7 topic is the status of capital itself: in some countries, it is distributed among different cities, as is the case of South Africa, with three capitals (administrative: Pretoria, legislative: Cape Town and judicial: Bloemfontein). Any of these cities was considered a correct hit if it also satisfied the other topic requirement(s).

During assessment, we also found interesting language differences concerning cognate proper names: While *Salisbúria* in Portuguese refers unambiguously to a previous (temporal phase of a current) African capital, now named *Harare*, the "corresponding" *Salisbury* entry in the English Wikipedia points to a place in the United Kingdom, more specifically in the English county of Wiltshire. (This reminds us of the issues with last year's GeoCLEF topics on St. Paul's Cathedral and St. Andrews, whose "translation" into *São Paulo* and *Santo André* brought up different and more prominent places in Brazil and Portugal as well [19]). This shows clearly the need not to completely trust translation equivalences, which can be misleading in one of the directions.

At another level, it was not always possible to maintain the level of detail as far as specific objects or concepts were concerned: for example, *rápidos* in Portuguese is a subtype of waterfall (namely one which is navigable), which apparently has to be rendered in English as *waterfall*, which is undeniably much more generic.

Finally, some features of Wikipedia itself caused unexpected problems for assessment: While a question for a city population seems to be natural from a user's point of view, we found that for some cities there are up to three different numbers (city, urban, metropolitan)! This implies that interaction with the user would be needed to identify which of these concepts s/he had in mind. For the time being, we accepted answers as correct provided one of the numbers satisfied the restriction stated in the topic (two topics concerned city populations: GP3 and GP7).

4 Overview of participation and results

As usual, a pilot task gets more expressions of interest than actual participations, and this happened in GikiP, with 13 groups reporting interest but only three participants in the end. One probably not irrelevant consideration is that GikiP is a hybrid task between QA and GIR, but was fully deployed – and therefore only conveniently publicized – under the GeoCLEF umbrella.[5] These are the participating systems:

[5]This should by no means be read as a critic of QA@CLEF organizers. On the contrary, they gave us at once access to their data and considered the task interesting. It just happened that the task was located at GeoCLEF because we were involved in GeoCLEF organization and not in QA@CLEF at the time.

Table 3: GikiP participation in 2008

System	Runs	Type of Run	Size	Languages
GIRSA-WP	6	automatic	798 (372)	de, en, pt
RENOIR	1	semi-automatic	218	en, pt
WikipediaListQA@wlv	1	automatic	123	de, en, pt
Human	1	manual	235	de, en, pt

Table 4: GikiP results in 2008

Run	Answers	Correct	Avg. Prec	Score
GIRSA-WP (best)	79	9	0.107	0.704
GIRSA-WP (all runs merged)	372	11	0.038	0.286
RENOIR	218	122	0.554	10.946
WikipediaListQA@wlv	123	93	0.632	15.815

- **GIRSA-WP**, represented by Sven Hartrumpf and Johannes Leveling, Intelligent Information and Communication Systems (IICS) at the FernUniversität in Hagen (Germany)
- **RENOIR** (acronym for REMBRANDT's Extended NER On Interactive Retrievals), represented by Nuno Cardoso, University of Lisbon, Faculty of Sciences, LaSIGE, XLDB (Portugal)
- **WikipediaListQA@wlv**, represented by Iustin Dornescu, Research Group in Computational Linguistics (CLG) at the University of Wolverhampton (UK)

Curiously, we had one participant per country where one of the three languages is spoken, and participation was divided equally between GeoCLEFers and QA@CLEFers (given that the IICS group is known to participate in both). We have also kept non-official submission to GikiP open until later (30 June) in order for people busy with the GeoCLEF main task to be able to try GikiP, although non-officially, but we received no further submissions.

The participating systems are shortly described in Table 3, together with a fully manual run based on the current Wikipedia, which was provided by Paulo Rocha. The point of requesting this run was to be able to, later on, compare human performance to automatic answers. Also, we wanted to assess how much Wikipedia information had changed regarding the particular topics, from the official collections to the June 2008 date. Each system will be fully described in the corresponding section in the paper.

The results obtained by the systems can be found in Table 4. For the record, the maximum number of documents returned per topic was 23 (for GP7), and the minimum was zero. It is also interesting to note that the human participant was not able to find any results for topics GP2 (*Which Vienna circle members* ...) and GP15 (*French bridges* ...), contrarily the automatic systems, which together managed to find 7 and 5 correct hits, respectively.

5 GIRSA-WP participation

GIRSA-WP (GIRSA for Wikipedia) is a fully-automatic, hybrid system combining methods from question answering (QA) and geographic information retrieval (GIR). In particular, it merges results from InSicht, an open-domain QA system [9], and GIRSA, a system for textual GIR [17].

5.1 System description

In comparison with the two underlying basic systems, GIRSA-WP applies a semantic filter on the article titles (which are encoded in the answers in GikiP) to increase precision. This semantic filter ensures that the expected answer type (EAT) of the topic and the title of a Wikipedia article are compatible. This technique is widely known from QA for typical answer types such as PERSON, ORGANIZATION, or LOCATION. In our system, a concept (a disambiguated word) corresponding to the EAT is extracted from the topic

title or description. Then, this concept and the title of a candidate article are parsed by WOCADI [8], a syntactico-semantic parser for German text. The semantic representations (more specifically, the ontological sort and the semantic features, see [12] for details) of the semantic heads are unified. If this unification succeeds, the candidate article is kept; otherwise it is discarded. For example, from topic GP4 (*Which Swiss cantons border Germany?*), the extracted concept is *canton*, which is an artificial geographical entity denoting a kind of regional institution.

The major differences to InSicht and GIRSA are that GIRSA-WP does not merge streams of answers and does not include a logical answer validation. In contrast to GIRSA, the retrieval is based on documents indexed on a per-sentence basis of Wikipedia articles. In addition, the documents from Wikipedia had not been geographically annotated at all.

For the GikiP experiments, the topic title and description were analyzed and sent to GIRSA and InSicht. In GIRSA, the top 1000 results were retrieved and scores were normalized in the interval from 0 to 1. For results returned by both GIRSA and InSicht, the maximum score was chosen. Results whose score was below a given threshold were discarded and the semantic filter was applied to the remaining results. To obtain multilingual results, the German article names were translated to English and Portuguese using the Wikipedia linking between languages. Note that this linking was the only non-textual information we used from Wikipedia; for example, categories and inter-article links were completely ignored.

In InSicht, the semantic representation of the query and the semantic representations of document sentences are compared. To go beyond perfect matches, InSicht uses many techniques, for example intratextual coreference resolution, query expansion by inference rules and lexicosemantic relations, and splitting the query semantic network at certain semantic relations. InSicht employed a special technique called *query decomposition* (first tried in GeoCLEF-2007, [17]) or *question decomposition* in the context of QA [10]. Among the different decomposition classes described in the latter paper, only meronymy decomposition and description decomposition are promising for current queries in GikiP. They led to successful decompositions, e.g. topic GP4 (*Which Swiss cantons border Germany?*) is decomposed into subquestions like *Name a canton in Switzerland.* (with subanswers *Aargau*, *Basel*, ...) and revised questions like *Does Aargau border Germany?* (examples translated from German). We tried a strategy to answer subquestions also in the CLEF-News corpus used in QA@CLEF but not in GikiP. But, the overall results were equal to the ones obtained using only Wikipedia for subquestions. InSicht achieved a higher precision than GIRSA-WP as a whole (0.144 compared to 0.107), but recall is still problematic as already seen in similar evaluations, e.g. GeoCLEF.

5.2 Experiments

Due to time constraints, the Wikipedia articles had not been fully processed and some methods have been applied to the topics only although they should have been applied to the documents, too. We performed six runs with the following experiment settings:

- run 1: topics and documents (Wikipedia sentences) are processed with full word forms (no stemming and no stopword removal); results are filtered by applying a threshold score of 0.01
- run 2: same setting as in run 1; location names are identified and normalized (for topics only)
- run 3: same setting as in run 2; German noun compounds are identified and split into their constituents (for topics only)
- run 4–6: same settings as in run 1–3; results are filtered by applying a threshold score of 0.03

5.3 Evaluation and discussion

Several reasons account for the relatively low performance of GIRSA-WP:

- In comparison with topics from the QA@CLEF task at CLEF, GikiP topics are more difficult to answer and aim at a wider range of EATs.

- In comparison with topics from the GeoCLEF task, topics are at least as difficult. They include complex geographic relations (GP2: *outside*, GP4: *on the border*), restrictions on measurable properties (GP3: *more than*, GP13: *longer than*), and temporal constraints (GP9: *Renaissance*, GP15: *between 1980 and 1990*).
- Indexing sentences instead of complete Wikipedia articles was meant to ensure a high precision. However, the fallback strategy (GIRSA) does not work well when applied on document sentences.
- Geographic entities were not annotated at all in the documents for GIRSA. Thus, the high precision and recall observed with GIRSA for the news collection could not be observed with the Wikipedia articles.
- For InSicht, the main problems were (1) that important information is given in tables (like inhabitant numbers), but the syntactico-semantic parser ignores these parts of articles and (2) that the semantic matching approach forming the basis of QA is still too strict for the IR oriented parts of GikiP queries (similar problems occurred for GeoCLEF experiments).
- The system's multilingual approach is too simple because it relies only on the Wikipedia of one language (German) and adds results by following title translation links to other languages. Therefore for questions that have no or few articles in German, relevant articles from English or Portuguese cannot be found.

5.4 Future work

Future work will include tackling some of the problems discussed in Section 5.3, enabling the annotation of geographic entities and geo-inferences, and preferring special regions of Wikipedia articles (for example, the introductory sentences).

6 RENOIR participation

The goal of RENOIR participation in GikiP is to explore new ways of doing GIR, specially for those kinds of geographic queries that cannot be correctly handled by just naïvely expanding the query terms, and hoping that the IR system with some sort of geographic reasoning capabilities captures the full meaning of the topic at stake, as we do for GeoCLEF.

As such, we chose to participate with one semi-automatic run using *query procedures* as retrieval input, instead of query terms. We define query procedures as a group of pipelined actions that express each GikiP topic. The generation of query procedures was entirely manual, and the execution of each action varies from automatic, semi-automatic and manual.

6.1 What is RENOIR

RENOIR is an interactive tool where query procedures are executed, generating partial and final results for each GikiP topic. RENOIR makes extensive use of REMBRANDT [4], a named entity recognition module which explores the Wikipedia document structure, links and categories, to identify and classify named entities (NEs) in texts written in Portuguese and English.

REMBRANDT classifies NEs according to the following 9 main categories: `PERSON`, `ORGANIZATION`, `LOCAL`, `TIMESTAMP`, `VALUE`, `ABSTRACTION`, `THING`, `MASTERPIECE` and `EVENT`. REMBRANDT participated on the second HAREM [24], a named entity recognition evaluation contest for the Portuguese language [25, 22], obtaining F-measure values of 0.57 for the full NER task.

We indexed the GikiP 2008 collection with MG4J [3], and it was used for basic document retrieval. For retrievals involving Wikipedia categories and links, we preferred to use different snapshots of Wikipedia (namely the Portuguese and English static SQL dumps dated from April 2008, onwards referred to as the Wikipedia dumps to avoid confusion), because the information regarding Wikipedia categories, redirections and page links was already available in SQL databases, and thus we did not need to preprocess the GikiP

2008 collection. As such, RENOIR also allowed us to perform basic actions to match the documents from the Wikipedia dumps to its corresponding GikiP 2008 documents, to cope with GikiP's submission format.

The RENOIR actions used for the query procedures are described as follows. The actions are labelled as automatic (RENOIR performs the action alone), semi-automatic (the action is supervised), or manual (the action is made manually).

1. Retrieval actions		
SEARCH TERM	Automatic	Performs a simple term query search in the GikiP 2008 collection, and returns a list of Wikipedia documents.
SEARCH CATEGORY	Automatic	Searches the Wikipedia dumps for documents with the given Wikipedia category, and returns a list of Wikipedia documents.
SEARCH INLINKS	Automatic	Searches the Wikipedia dumps for documents that link to a given Wikipedia document.
2. Mapping actions		
MAP DOC	Semi-automatic	Maps a document from the Wikipedia dump to its counterpart in the GikiP 2008 collection.
MAP NE	Semi-automatic	Maps a NE to its corresponding document in the GikiP 2008 collection.
3. Annotation actions		
REMBRANDT	Automatic	Annotates selected Wikipedia document(s) with REMBRANDT, generating lists of NEs for each document.
REMB. DOC TO NE	Automatic	Invokes REMBRANDT to classify the title of a given Wikipedia document, generating the respective NE.
4. Filtering actions		
FILT. NE BY TYPE	Automatic	Filters a list of NEs of a given classification category, generating a subset of NEs.
FILT. DOC BY TERM	Automatic	Filters a list of Wikipedia documents by having (or not) a given term/pattern
FILT. DOC BY EVAL	Manual	Filters a list of Wikipedia document by evaluating a condition for a given subset of NEs. For instance, if the document has a number NE greater than 1000, or if it has a place name NE within Europe.

6.2 Generating the query procedures

The query procedures were formulated in a simple modular and pipelined approach. This allowed us to "divide and conquer" the complex task of translating the GikiP topics into a machine-understandable way. So, the actions that could be made automatically were therefore implemented, while the more complex actions performed in GikiP with human intervention (so far) were also kept simple in order to be possible to extend RENOIR to perform them automatically in the future. Anyway, while RENOIR can not perform (yet) all actions automatically, the methodology devised to handle the GikiP task can already be tested, and we can get acquainted with the difficulties that will face us during further development of RENOIR.

Table 5 presents the query procedures used for the submitted runs for English. The query procedures for Portuguese had the same actions, with small exceptions discussed later. The query procedures are best explained by following the example topic GP7 (*Capitals of Africa* ...), which has a query procedure pattern similar to other topics.

SEARCH CATEGORY "Capitals of Africa" $\rightarrow$ $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$;
FILT. NE BY TYPE $NE_1 \in$ VALUE/QUANTITY $\rightarrow NE'_1$;
FILT. DOC BY EVAL NE'_1 >2,000,000 $\rightarrow$ $Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.

1. SEARCH CATEGORY ''Capitals of Africa'' $\rightarrow$ $Docs_1$ - The expected answers for this topic is a filtered list of Wikipedia pages that are likely to belong to a same category, "Capitals of Africa".

Table 5: RENOIR's query procedures for the GikiP topics. $Docs_n$ represents lists of Wikipedia documents, NE_n represents lists of NEs.

GP	Query Procedure
1	SEARCH TERM "last mohicans film" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/WATERCOURSE → NE'_1; MAP NE $NE'_1 \rightarrow Docs_2$.
2	SEARCH TERM "Vienna Circle" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ PERSON/INDIVIDUAL → NE'_1; MAP NE $NE'_1 \rightarrow Docs_2$; REMBRANDT $Docs_2 \rightarrow Docs'_2, NE_2$; FILT. NE BY TYPE $NE_2 \in$ LOCAL/COUNTRY → NE'_2; FILT. DOC BY EVAL $NE'_2 \notin$ {'Austria', 'Hungary', 'Germany'} → $Docs''_2$.
3	SEARCH CATEGORY "Rivers of Portugal" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/DIVISION → NE'_1; MAP NE $NE'_1 \rightarrow Docs_2$; REMBRANDT $Docs_2 \rightarrow Docs'_2, NE_2$; FILT. NE BY TYPE $NE_2 \in$ VALUE/QUANTITY → NE'_2; FILT. DOC BY EVAL $NE'_2 >$150,000 → $Docs''_2$; MAP DOC $Docs''_2 \rightarrow Docs'''_2$.
4	SEARCH CATEGORY "Cantons of Switzerland" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/COUNTRY → NE'_1; FILT. DOC BY EVAL $NE'_1 \in$ {'Germany'} → $Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.
5	SEARCH TERM "Greece" → $Docs_1$; SEARCH INLINKS $Docs_1 \rightarrow Docs_2$; REMB. DOC TO NE $Docs_2 \rightarrow NE_2$; FILT. NE BY TYPE $NE_2 \in$ EVENT → NE'_2; FILT. DOC BY EVAL NE'_2 contains 'war' → NE''_2; MAP NE $NE''_2 \rightarrow Docs_3$.
6	SEARCH CATEGORY "Mountains of Australia" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ VALUE/QUANTITY → NE'_1; FILT. DOC BY EVAL $NE'_1 >$2000 m → $Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.
7	SEARCH CATEGORY "Capitals of Africa" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ VALUE/QUANTITY → NE'_1; FILT. DOC BY EVAL $NE'_1 >$2,000,000 → $Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.
8	SEARCH CATEGORY "Bridges of Brazil" → $Docs_1$; FILT. DOC BY TERM "suspen*" → $Docs'_1$; MAP DOC $Docs'_1 \rightarrow Docs''_1$.
9	SEARCH CATEGORY "German composers" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ DATE → NE'_1; FILT. DOC BY EVAL $NE'_1 \geq 1380 \wedge NE'_1 \leq 1640 \rightarrow Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.
10	SEARCH TERM "Polynesia" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/ISLAND → NE'_1; MAP NE $NE'_1 \rightarrow Docs_2$; REMBRANDT $Docs_2 \rightarrow Docs'_2, NE_2$; FILT. NE BY TYPE $NE_2 \in$ VALUE/QUANTITY → NE'_2; FILT. DOC BY EVAL $NE'_2 >$5,000 → $Docs_3$.
11	SEARCH CATEGORY "Shakespearean plays" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/DIVISION,COUNTRY → NE'_1; FILT. DOC BY EVAL $NE'_1 \in$ Italy → $Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.
12	SEARCH TERMS "Goethe" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/DIVISION → NE'_1; MAP NE $NE'_1 \rightarrow Docs_1$.
13	SEARCH CATEGORY "Rivers of Afghanistan" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ VALUE/QUANTITY → NE'_1; FILT. DOC BY EVAL $NE'_1 >$1,000 km → $Docs_2$. MAP DOC $Docs_2 \rightarrow Docs'_2$.
14	SEARCH CATEGORY "Brazilian architects" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ LOCAL/DIVISION → NE'_1; FILT. DOC BY EVAL $NE'_1 \in$ Europe → $Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.
15	SEARCH CATEGORY "Bridges in France" → $Docs_1$; REMBRANDT $Docs_1 \rightarrow Docs'_1, NE_1$; FILT. NE BY TYPE $NE_1 \in$ DATE → NE'_1; FILT. DOC BY EVAL $NE'_1 \geq 1980 \wedge NE'_1 \leq 1990 \rightarrow Docs_2$; MAP DOC $Docs_2 \rightarrow Docs'_2$.

So, we first search for all Wikipedia pages that contain that category, and obtain the list of Wikipedia documents $Docs_1$. This step is done automatically.

2. `REMBRANDT` $Docs_1 \rightarrow Docs'_1$, NE_1 - Afterwards, REMBRANDT annotates the documents in $Docs_1$, generating a tagged version of these same documents, $Docs'_1$, and their corresponding named entities, NE_1. This step is done automatically.

3. `FILT. NE BY TYPE` $NE_1 \in$ `VALUE/QUANTITY` $\rightarrow NE'_1$ - We narrow the NE list to the entities that were classified by REMBRANDT as VALUE/QUANTITY, generating the subset NE'_1. This step is done automatically.

4. `FILT. DOC BY EVAL` NE'_1 `>2,000,000` $\rightarrow Docs_2$ - We now evaluate the NE subset NE'_1, searching for values greater than 2,000,000. When found, the respective document is therefore added to $Docs_2$. This step is done manually.

5. `MAP DOC` $Docs_2 \rightarrow Docs'_2$ - Finally, since we started with a pool of Wikipedia documents from the SQL dumps because of the `SEARCH CATEGORY` action, we need to map the documents in $Docs_2$ to their corresponding GikiP 2008 documents, generating $Docs'_2$, which is the final result. The URLs of $Docs'_2$ are then added to the submission file. The mapping step is done semi-automatically.

6.3 Results and discussion

Our approach for GikiP has some obvious faults and over-simplifications that are clearly compensated by human supervision. For instance, knowing which NE of type `VALUE/QUANTITY` corresponds to a population count is easy for a human, but it is not an easy task for an automated system. Yet, one of the best features of Wikipedia (from an IE system developer's point of view) is the clustering of such important features into infoboxes, which are easily machine-interpretable and thus this problem can be mitigated by developing a robust Wikipedia infobox parsing module.

Other over-simplifications involve, for instance, topic GP14 (*Brazilian architects* ...), where we assume that a Wikipedia page that contains at least one NE referring to a place in Europe is enough to consider that the targeted Brazilian architect page is relevant, or in topic GP11 (*Shakespearean plays* ...) where reference to a place in Italy is enough to consider that a given Shakespearean play actually happens in Italy. Last but not least, knowing whether the place names were Italian or European was manually done. Later on, we will try to access geographic ontologies in order to implement such restrictions, to fully verify the feasibility of the approach.

Also worth mentioning is the bypass we made to the problem of selecting the correct categories for the first pool of Wikipedia documents, as done in the query procedures that started with the `SEARCH CATEGORY` action, by doing it manually. For instance, in topic GP9 (*German composers* ...), it turned out that the best category to start with was "German composers", while in topic GP15 (*French bridges*...) the best category is "Bridges in France". We tentatively concluded that the first form is typically used to cluster persons by country and occupation, and the second is used for non-person entities.

Nonetheless, one of the requirements for an automated query procedure builder module is to figure out that "Composers of Germany" and "French bridges" are not Wikipedia categories, and thus a domain identification step as performed by WikipediaListQA@wlv (see Section 7) is quite important. Another problem arises for topics that could be started with a category, but there is no such category, as in topic GP5 (*Name all wars* ...). We expected a category such as "Wars of Greece" to exist, but it did not, so we had to choose a different query procedure that involved heavy processing: we selected Wikipedia pages with outlinks to Greece, resulting in an initial pool of over 20,000 documents linking to the Wikipedia page of Greece. Then all titles were classified by REMBRANDT, and those who had an `event` plus the term "War" in the title were finally selected.

Lastly, we report the differences between the Portuguese and English query procedures, which were related to the Wikipedia categories used in both snapshots. Take for instance the topic GP6 (*Which Australian mountains* ...). It is well bootstrapped in Portuguese by the category "Montanhas da Austrália", but, in the English Wikipedia snapshot, the category "Mountains of Australia" has a bundle of Wikipedia pages pointing to mountains and also subcategories that group mountains in Australian regions, such as

New South Wales, South Australia, Northern Territory or Tasmania. Clearly, the two topic formulations, although apparently similar, may refer in English to a whole continent while in Portuguese only to a country (the continent being named *Oceânia*).

On the other hand, in Wikipedia there are some categories which have subcategories before we arrive to the nodes that are real answers. To handle this we included new queries for the subcategories. For example, for topic GP11 (*Shakespearean plays* ...), the English Wikipedia category "Shakespearean plays" is divided into several subcategories that group the adaptations, comedies, tragedies, histories and apocrypha. So all these subcategories had to be visited. In this case, there was a difference between the English and the Portuguese Wikipedia since the latter had the subcategories but these were not yet filled, so the procedure had to be different.

6.4 Conclusions and future work

The GikiP experiments allowed us to take a first glance at the difficulties that expect us when dealing with more complex queries with specific geographic criteria. As our goal is to research retrieval approaches that can profit from a comprehensive semantic layer over queries and documents, we found the GikiP exercise to be interesting and innovative.

The next obvious step is to implement the automatic generator of query procedures, dealing with the problems that were mitigated by using human reasoning. At the same time, future work includes the improvement of the Wikipedia mining approaches, namely extracting information from infoboxes.

7 WikipediaListQA@wlv participation

The participation in this pilot task was motivated by our interest in using Wikipedia as a backbone in QA. In addition, the way the task is setup required us to rely on the information inherent in the Wikipedia article link graph and the relation between entities, rather than developing accurate textual answer extractors.

For example, if we try to find out information about the cities that the Douro river flows through, we can extract all the links present in the article describing the Douro river, creating a list of entities that are related to the river. If we only select the articles from this list that describe a settlement, a town or a city, the filtered list would likely contain localities that the river is passing through. By examining the infobox of each candidate, we can determine the population of each locality. We may apply a selection function to this list. We can compute the average population size, number of towns with population larger than 150,000 inhabitants, etc. discovering new facts.

For example, in solving the GP3 topic, the system identifies the Wikipedia category *Category:Rivers of Portugal*. All the articles directly linked to this category are likely to be rivers, thus the list of candidate answers is created. As previously described, for each river the system extracts the list of links to articles and attempts to extract the population size from each one. The maximum value encountered is computed. If this value is smaller than the threshold, the river is discarded.

In order to navigate the Wikipedia link graph we had to transform the Wikipedia SQL dump and index it with Lucene [11] because loading the data in MySQL would have taken too long (more than 200 million links in English Wikipedia alone). The index already contained a cleaned version of Wikipedia (April 2008). Thus our results had to be mapped to the November 2006 version.

7.1 Overview of the system

We propose a simple model for topic interpretation. It exploits relationships between entities that may not be expressed in the article text, but are implied by the links between the articles.

Our system starts by identifying a domain category that comprises candidate articles, and then filters out the ones that do not correspond to the topic filter. Thus two parts are identified in each topic: a) the *domain* of the candidate answers, and b) the *filters* to apply in order to select the correct ones.

Domain identification We had to identify a Wikipedia category that would describe the candidate articles. We used the Connexor FDG parser [27] in order to extract noun phrases from the topics. The first

noun phrase was matched with a category, by querying the Lucene [11] index. We used lexical rules in order to achieve a good accuracy (e.g. `"portuguese rivers"` ⇒ `("rivers in portugal" "rivers of portugal" "portugal rivers"~ 2 )^5 ("portuguese rivers"^5 "portuguese rivers"~ 2 )^5 portugal rivers portuguese`). This simple method succeeded in identifying good categories in most cases. However at times it only matched a very general category (*waterfalls* in GP1, *wars* in GP5).

Candidate filtering For the purpose of this pilot only very simple filters were implemented. There are two main filter categories: **entity filters** and **factoid filters**.

The entity filters match documents that mention or have a link to a given entity. 8 of the 15 topics had such a constraint. In 2 cases there was a list of entities that should have matched. Most notably the temporal restriction identified in topic GP15 *between 1980 and 1990* was expanded to the list of 10 years in the range. This behaviour was adapted from our QA system.

The factoid filters match documents in which the identified fact can be extracted, and the value corresponds to the selection criterion. The facts were extracted using components from our question answering system (infobox look-up, regex patterns). The facts were: *population* (3 topics), *nationality* (2 topics), *height* (1 topic) and *length* (1 topic). Articles from which the fact could not be extracted were dismissed. The selection criterion was applied to the extracted fact: *greaterThan* (5 topics - numeric facts), *inList* (1 topic), *not_inList* (1 topic).

Multilinguality The method proposed has the advantage that it can easily be adapted to a cross-lingual task. Only the English topics were analyzed. The results of the analysis can be directly mapped to any language version of Wikipedia. Firstly the corresponding category describing the domain of the candidate articles has to be identified by using inter-wiki links. Secondly the filter has to be "translated". This either means translating the entities, or having the necessary language dependent fact extractors. Cross-wiki answers can be combined and re-ranked, aggregating the results from all the languages by using the inter-wiki links. This allows an English-speaking user, for example, to exploit the fact that the Portuguese Wikipedia has much richer content regarding Brazil and Portugal since most Portuguese Wikipedia contributors live in these two countries. This year we only searched the English Wikipedia and the results were mapped to all the three languages by using the inter-wiki links because we did not have time to create the fact extractors for German and Portuguese.

7.2 Results and error analysis

The accuracy of the system is limited due to the ambiguity of links. Not all articles that pertain to *Category:Abidjan* refer to an African capital. Category relations were not classified: hypernymy vs. meronymy vs. similarity. When searching all the articles that have a certain hyper-category, due to link type ambiguity, very large article sets might be extracted (the system did not return any results for three topics, because the list of domain articles was too large). This can be avoided by using resources that map Wikipedia articles to WordNet (e.g. Yago [26] and DBpedia [2]) and disambiguate the type of the entity described in each article. Also, filters can be implemented using tools of the Semantic Web (e.g. SPARQL queries).

The current system is a simple model which has proven to have a good precision of finding the answer. Its main advantage is that – using a small set of filters – very complex data can be queried from Wikipedia. Its greatest disadvantage lies in the complexity of correctly identifying (combined) filters in natural language questions. Given an appropriate user interface, this method can become an alternative to Wikipedia search, allowing users to access information that is not textually present in the encyclopedia.

8 Concluding remarks

We think that the results presented in this pilot are encouraging, both for the possibility of automating the particular task, and for its general interest as another way of reaching the information in Wikipedia. GikiP has shown that there are interesting kinds of non-trivial questions that have a retrievable answer in Wikipedia, and which can be quickly assessed by users, often without even having to visit the page.

8.1 Summing up the approaches

Interestingly, participating systems took a wide range of different approaches, as well as different main collections (GIRSA-WP used German and WikipediaListQA2wlv English as main answering sources; while RENOIR did a parallel process for English and Portuguese). We intend, in a following publication, to study results per language and per topic to see whether results might have been heavily influenced by this choice of main collection to investigate. While one participant took a preliminary semi-automatic approach, although making use of several automated procedures, the two others (including the winner) used fully automatic systems from start to end.

8.2 Remaining work

Much work remains to be done for issues of redundancy removal, choice of which language / answer to present first (or only), as well as how to present a compound set of pages to justify a particular answer.

Also, the task needs a more reflected and precise definition.

- For example, the fact that natural language is fundamentally vague may lead to a set of answers at different levels, which, in order to satisfy a human user, should be presented in a more appropriate way: Consider the case of topic GP12 (*Places where Goethe lived*), where also names of countries were considered correct. Obviously, a more adequate answer would structure or make use of the hierarchical relations between the answers and not present all of them alphabetically as a flat list.

- Most information needs require a lot of conceptualization. This is illustrated by topic GP5 (*Name all wars that occurred on Greek soil*). After starting assessment, we found that the relation of battles or sieges (which occur in particular places) as belonging to wars may require considerable theorizing and historical knowledge. Wars, especially perhaps when one is considering ancient wars, are more difficult to pin down to particular places. So, assessment turned out to be quite tricky, and we fear that the topic did not illustrate a realistic user need.

- Finally, the *mult* factor, which was assigned 2 to RENOIR and 3 to the other two systems, was a fairly gross attempt to incentivize results in the three languages. However, different topics, as discussed in Section 3.1, would be easier or more productive in different languages and in fact answers which could only be found in one vs. two or three languages might be considered more difficult and therefore rewarded with another kind of multilinguality-related factor. This is something that must be investigated in the future. Namely, how to increase the score of particularly difficult topics.

We hope that further editions of GikiP and similar tasks will help research and development along those lines.

Acknowledgements We are grateful to Ross Purves for checking the English rendering, to Sven Hartrumpf for analysing and debugging the German and English versions with the WOCADI parser, to Anselmo Peñas for making the Wikipedia collections available to GikiP participants, and most particularly to Paulo Rocha for providing a fully manual run in two days.

The organization work was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

The development of the WikipediaListQA@wlv system was partly supported by the EU funded project QALL-ME (FP6 IST-033860).

The development of RENOIR was supported by grant SFRH/BD/29817/2006 from FCT (Portugal).

References

[1] Alexa top 500 sites. `http:www.alexa.com/site/ds/top_sites?ts_mode=global`.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007. Proceedings*, pages 722–735. Springer, 2008.

[3] Paolo Boldi and Sebastiano Vigna. MG4J at TREC 2005. In *Proceedings of the 14th Text REtrieval Conference (TREC 2005)*, 2005.

[4] Nuno Cardoso. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*, 7 September 2008.

[5] Frederic Gey, Ray Larson, Mark Sanderson, Hideo Joho, and Paul Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised Selected papers*, volume 4022 of *LNCS*, pages 908–919. Springer, Berlin, 2006.

[6] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bishoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Di Nunzio, and Nicola Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. 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-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *LNCS*, pages 852–876. Springer, Berlin, 2007.

[7] Danilo Giampiccolo, Pamela Forner, Anselmo Peñas, Christelle Ayache, Dan Cristea, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe. Overview of the CLEF 2007 Multilingual Question Answering Track. In Alessandro Nardi and Carol Peters, editors, *Working Notes for the CLEF 2007 Workshop*, 19–21 September 2007.

[8] Sven Hartrumpf. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany, 2003.

[9] Sven Hartrumpf. Question answering using sentence parsing and semantic network matching. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *LNCS*, pages 512–521. Springer, Berlin, 2005.

[10] Sven Hartrumpf. Semantic decomposition for question answering. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikos Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI)*, pages 313–317, Patras, Greece, July 2008.

[11] Erik Hatcher and Otis Gospodnetic. *Lucene in Action (In Action series)*. Manning, Greenwich, CT, USA, 2004.

[12] Hermann Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, 2006.

[13] ImageCLEF's WikipediaMM task, 2008. `http://imageclef.org/2008/wikipedia`.

[14] Valentin Jijkoun and Maarten de Rijke. Overview of the WiQA Task at CLEF 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-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *LNCS*, pages 265–274. Springer, Berlin, 2007.

[15] Valentin Jijkoun and Maarten de Rijke. WiQA: Evaluating Multi-lingual Focused Access to Wikipedia. In *The First International Workshop on Evaluating Information Access (EVIA)*, Tokyo, Japan, May 15 2007.

[16] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR Tasks at the first NTCIR Workshop. In *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44, Tokyo, Japan, August 1999.

[17] Johannes Leveling and Sven Hartrumpf. Inferring location names for geographic information retrieval. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *LNCS*, pages 773–780, Berlin, 2008. Springer.

[18] *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 28–30th May 2008. ELRA.

[19] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In Alessandro Nardi and Carol Peters, editors, *Working Notes for the CLEF 2007 Workshop*, 19–21th September 2007.

[20] Carol Peters and Martin Braschler. Cross-Language System Evaluation: the CLEF campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072, 2001.

[21] Diana Santos and Nuno Cardoso. Portuguese at CLEF. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised selected papers*, volume 4022 of *LNCS*, pages 1007–1010, Berlin, 2006. Springer.

[22] Diana Santos and Nuno Cardoso, editors. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007.

[23] Diana Santos and Nuno Cardoso. GikiP: Evaluating geographical answers from Wikipedia. In *GIR 2008*, Napa Valley, CA, USA, 1 November 2008. submitted.

[24] Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira, and Paula Carvalho. Second HAREM: new challenges and old wisdom. In António Teixeira et al., editor, *PROPOR 2008*, volume 5190 of *LNAI*, pages 212–215. Springer, 8–10 September 2008.

[25] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Nicoletta Calzolari et al., editor, *Proceedings of LREC'2006*, pages 1986–1991, 22–28 May 2006.

[26] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM Press, New York, NY, USA, 2007.

[27] Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, pages 64–71, Washington D.C., USA, 1997. ACL.

[28] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, Massachusetts, USA, 2005.

[29] Jakob Voss. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics, ISSI'2005*, Stockholm, Sweden, 24–28 July 2005.

[30] Wikipedia. `http://www.wikipedia.org/`.

[31] WiQA: Question Answering using Wikipedia, 2006. `http://ilps.science.uva.nl/WiQA/`.