GikiP: Evaluating Geographical Answers from Wikipedia

Diana Santos
Linguateca, SINTEF ICT, Norway
Diana.Santos@sintef.no

Nuno Cardoso
Univ. of Lisbon, Faculty of Sciences, LaSIGE
ncardoso@xldb.di.fc.ul.pt

ABSTRACT

This paper describes GikiP, a pilot task that took place in 2008 in CLEF. We present the motivation behind GikiP and the use of Wikipedia as the evaluation collection, detail the task and we list new ideas for its continuation.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms: Design, Measurement

Keywords: Geographical IR, Evaluation, Question Answering, Wikipedia, Crosslingual IR

1. MOTIVATION

In this paper we present GikiP, an evaluation contest on extracting information from Wikipedia in the form of a list of answers (corresponding to articles) that have some geographical component. Our main interest is investigate how humans access the information or knowledge they need or want. To that goal, it is fairly immaterial which way they query or the form of the result. For this reason, we believe hard boundaries between question answering (QA) and information retrieval (IR) are not useful [1]. One subject which is frequently associated to human queries is geography, in ways far richer than just adding a geographical restriction.

So, in GikiP we tried to come up with several different ways to involve geographical knowledge, beyond the scope restriction. From the typology described in [2], we had topics of the following kinds: 3 - geographic subject restricted to a place; 4 - non-geographic subject associated to a place, such as people with some geographical property (from nationality or born in to quite complex geographically-correlated concepts, as belonging to the Vienna Circle); and in general 5 - (geo or non-geo) subjects that are a complex function of place – temporally anchored properties of a place, places/buildings in a particular temporal phase, and events taking place in areas variable in time.

2. THE GIKIP TASK AT GEOCLEF

GeoCLEF, `www.uni-hildesheim.de/geoclef/`, is the first evaluation forum for evaluating geographical information retrieval. As part of CLEF, it is crosslingual and multilingual, and in the last years has provided (parallel) topics in English, German and Portuguese against a multilingual newspaper collection (from Brazil, Germany, Portugal, Scotland, Switzerland and the US) from 1994-1995. The sister forum for QA evaluation at CLEF, QA@CLEF, has used frozen snapshots of Wikipedia for the last two years, in addition to newspaper collections.

For GikiP we have therefore made those German, Portuguese and English Wikipedia snapshots the target of our GIR topics, attempting a marriage between QA and IR in that short answers would be the title of documents (read: Wikipedia entries) that satisfied the queries/questions. Topics were provided both as standard IR topics with title and description, and in a question form, the title (see Table 1).

2.1 Task definition

So, the particular specific task to be solved/tried by this year's participants was:

> Find Wikipedia entries / articles that answer a particular information need which requires geographical reasoning of some sort.

Fifteen topics were made available on the 2th June 2008, from the GikiP site (`www.linguateca.pt/GikiP`), where eight example topics in the three languages had already been published. Participants had 10 days to return the results, in the form of a list of Wikipedia articles.

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 names of people (painters and scientists), and not names of boats, or of countries. Conversely, if the question was about countries, the type of results should be places of the country type, and neither wars nor 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 were typically much less than that number).

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. We delivered (and assessed) topics in Portuguese, German and English, and therefore, an additional bonus was computed for multilinguality, *mult*, which

Table 1: Topic titles, pool size and correct answers in GikiP 2008

Topic nr.	English topic title	Results	Correct
GP1	Which waterfalls are used in the film "The Last of the Mohicans"?	5	1
GP2	Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany?	31	7
GP3	Portuguese rivers that flow through cities with more than 150,000 inhabitants	28	8
GP4	Which Swiss cantons border Germany?	79	21
GP5	Name all wars that occurred on Greek soil.	69	21
GP6	Which Australian mountains are higher than 2000 m?	36	7
GP7	African capitals with a population of two million inhabitants or more.	90	33
GP8	Suspension bridges in Brazil.	49	2
GP9	Composers of Renaissance music born in Germany.	49	17
GP10	Polynesian islands with more than 5,000 inhabitants.	53	2
GP11	Which plays of Shakespeare take place in an Italian setting?	35	23
GP12	Places where Goethe lived	51	25
GP13	Which navigable rivers in Afghanistan are longer than 1000 km?	9	4
GP14	Brazilian architects who designed buildings in Europe.	60	6
GP15	French bridges which were in construction between 1980 and 1990.	18	2
Total		662	179

is 1, 2 or 3 depending on the number of languages they searched/provided results. System's results are thus evaluated according to the following formula: $mult * N * N/total$ where N is the number of correct hits, $N/total$ is precision, and $mult$ rewards multilinguality. The system's final score is given by the average of the individual scores.

3. PARTICIPANTS AND RESULTS

Participating systems and their results are presented in Table 2, where A stands for the number of answers, C the correct ones, AP average precision and S the final score. Each system is fully described in [3]. Interestingly, the systems were able to find hits for two topics where a human run had found no candidates.

Table 2: GikiP results in 2008

Run	A	C	AP	S
GIRSA-WP (best run)	79	9	0.107	0.704
RENOIR	218	122	0.554	10.946
WikipediaListQA@wlv	123	93	0.632	**15.815**

4. IDEAS FOR FUTURE EDITIONS

We believe that GikiP is ideal to experiment with new varieties of geographical information retrieval, given that it targets a truly multilingual and multimedia collection, Wikipedia. Just 15 topics allowed us to uncover several interesting issues and foresee improvements to the task, from the need to provide an adequate ranking, to check true translations, and to decide based on contradictory evidence in different languages.

In addition, and due to the vagueness inherent in natural language, even the apparently most straightforward topics showed to benefit from either interactive refinement or at least a more structured presentation form.

For example, the granularity level as far as places are concerned would definitely ask for a structured answer to a question about places where Goethe lived, instead of a list containing all from countries to cities to buildings.

For future editions, and attempting to provide a more natural query environment, we believe that one should also provide topics with more than one restriction, such as "Portuguese cities founded before 1500 with rivers larger than 100 km and featuring a Moorish castle". This is closely related to requiring composite answers, i.e, having to return more than one Wikipedia article as answer. (For example, for each Portuguese city, the system would point also to the corresponding river and Moorish castle).

Another important issue would be a more thorough and direct justification of the answers, either obtained by selecting a justifying snippet in the Wikipedia article (as in INEX, www.inex.otago.ac.nz), or by spelling out the reasoning employed. For instance, producing an answer to GP11 that, after (i) pointing to "Much ado about nothing", stated that (ii) this play has some scenes in Messina, and that (iii) Messina is in Italy. Still another possibility is to formulate questions focusing on paths, such as: retrieve an ordered set of places where Magellan was along his voyage, or a set of chronologically ordered battles in a war. This could in the longer run provide a map or path as an answer.

In fact, the interaction among several modes is a hot topic nowadays, and GikiP could be naturally extended to tasks such as: find images of particular places, or objects in images. After all, Wikipedia is a multimodal resource, allowing us to address cross-modal and/or multimodal queries such as "Which Swiss cantons have a lion on their flag", or "Find portraits of married women in the 18th century".

Finally, and closer to WiQA's spirit, ilps.science.uva.nl/WiQA, queries could focus on specific disagreements between articles, which might provide different answers for the same question in Wikipedia, in different languages or even in different articles written from other points of view.

Acknowledgements

5. REFERENCES

[1] N. Cardoso and D. Santos. To separate or not to separate: reflections about GIR practice. In *Novel methodologies for evaluation in information retrieval, NMEIR 2008 (ECIR'2008 Workshop)*, Glasgow, UK, 30 March 2008.

[2] F. Gey et al. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In C. Peters et al., editor, *CLEF 2006*, volume 4730 of *LNCS*, pages 852–876. Springer, 2007.

[3] D. Santos, N. Cardoso, P. Carvalho, I. Dornescu, S. Hartrumpf, J. Leveling, and Y. Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. In F. Borri et al., editors, *Working notes of CLEF 2008*, Aarhus, Denmark, 17-19 September 2008.