GikiCLEF: Expectations and lessons learned

Diana Santos, Luís Miguel Cabral

Linguateca, Oslo node, SINTEF ICT, Norway
{Diana.Santos, Luis.M.Cabral}@sintef.no

Abstract. This overview paper is devoted to a critical assessment of GikiCLEF 2009, an evaluation contest specifically designed to expose and investigate cultural and linguistic issues in Wikipedia search, with eight participant systems and 17 runs. After providing a maximally short but self contained overview of the GikiCLEF task and participation, we present the open source SIGA system, and discuss, for each of the main guiding ideas, the resulting successes or shortcomings, concluding with further work and still unanswered questions.

1 Motivation

One of the reasons to propose and organize GikiCLEF (and the previous GikiP pilot [1]) was our concern that CLEF did not in general propose realistic enough tasks, especially in matters dealing with crosslingual and multilingual issues, both in topic/question creation and in the setups provided. In other words, while sophisticated from many points of view, CLEF setup was deficient in the attention paid to language differences (see e.g. [2, 3]) or to the task definition [4, 1].

While we all know in IR evaluation that laboratory testing has to be different from real life, and that a few topics or choices are not possible to validate a priori, but have to be studied after enough runs have been submitted and with respect to the pools and systems that were gathered[1], we wanted nevertheless to go some steps further, attempting to satisfy the following desiderata. GikiCLEF thus should:

1. provide a marriage of information needs and information source with real-life anchoring: and it is true that the man in the street does go to Wikipedia in many languages to satisfy his information needs;
2. tackle questions difficult both for a human being and for a machine: basically, we wanted a task with real usefulness, and not a task which would challenge systems to do what people don't want them to do. On the other hand, we wanted of course tasks that were possible to assess by (and satisfy) people, and not tasks that only computers could evaluate;
3. implement a context where different languages should contribute different answers, so that it would pay to look in many languages in parallel;
4. present a task that fostered the deployment of multilingual (and monolingual) systems that made use of comparable corpora.

[1] In fact, although this has been done for TREC – see [5, 6] – it still remains to be done for CLIR or MLIA, although GridCLEF [7] is a significant step in this direction.

We also note that GikiCLEF was organized after a successful GikiP pilot which had already been meant as first step in these directions: GikiP, run in 2008, offered fifteen list questions to be solved in three language Wikipedias (Portuguese, German, and English), but had only three participants. As expounded in [8], we hoped that a larger contest could be organized that would foster research in useful tasks that required cultural awareness and were not based or centered around English alone.

Given that GikiCLEF 2009's setup and results have already been described in detail in the pre-workshop working notes [9], as well as being documented in its website,[2] we devote the current text to two main subjects: a presentation of SIGA as a reusable tool for new and related campaigns; and a discussion of whether GikiCLEF really managed to address and evaluate the task of "asking culturally challenging list questions to a set of ten different Wikipedias", presenting the achievements and shortcomings of what was in our opinion accomplished. We start in any case by offering a short description of the GikiCLEF task in order that this article be self-contained.

2 Very brief description of the task

Systems participating in GikiCLEF were supposed to find, in several languages,[3] answers to questions that required or expected reasoning of some sort (often geographical, but also temporal and other).

In order to be considered as a correct answer, systems had to present it and a set of (Wikipedia) pages that justified it, in the eyes of a human being. Systems were thus invited to provide justification chains, in all the cases where the process of getting an answer involved visiting and understanding more than one Wikipedia page (see GikiCLEF's website for the exact submission format).

From the point of view of the assessment, this meant that, in order for the GikiCLEF setup to mirror a useful task, human assessors had to decide whether a given answer was (i) correct (by reading the pages or because they knew it) and (ii) justified (and, in that case, prior knowledge would not suffice).

Additionally, even if they knew better, assessors were required to "believe" Wikipedia, in the sense that even a wrong answer should be accepted as correct – according to the source, of course.

The extremely simple evaluation measures should only obey two constraints: One, the more languages the participant systems were able to provide answers in, the better. Two, systems should not be penalized if there were no answers in a particular language (Wikipedia). GikiCLEF scores were thus computed as the sum, for each language, of

[2] http://www.linguateca.pt/GikiCLEF/

[3] The GikiCLEF 2009 languages were: Bulgarian, Dutch, English, German, Italian, Norwegian – both Bokmål and Nynorsk –, Portuguese, Romanian and Spanish. A remark is in order concerning Norwegian: since it has two written standards, and Norwegians keep Wikipedia in two "parallel" versions, GikiCLEF covers nine languages but ten collections. Since both written standards of Norwegian were dealt equally in GikiCLEF, we will talk loosely of ten languages in what follows.

precision times the number of correct answers. For each language, the score was C*C/N (so that one had a score for de, pt, etc, as $C_{de} * C_{de}/N_{de}$, $C_{pt} * C_{pt}/N_{pt}$, etc.).[4]

In order to avoid machine translation problems – or even the lack of MT systems for any of the language (pairs) – the 50 questions were provided in all languages, for them to be on an equal footing. This was possibly the only unrealistic bit of the GikiCLEF setup, but let us stress that even for human beings the translation was not an easy task (again, see the website and the working notes paper for details). If we had relied on the participating systems having to invoke on their own MT for the topics (which had to be provided in different languages), we believe this would introduce a lot of uninteresting noise in the system.

Due to this choice, anyway, GikiCLEF can also be conceived as ten different evaluation contests (each asking questions in ONE language to a set of ten collections). So, the GikiCLEF evaluation has also provided results per language.

3 The SIGA system

SIGA[5] follows a similar structure as other systems such as DIRECT [10] or the one used in INEX [11], encompassing multiple user roles for different tasks. Different choices and privileges are thus in action for e.g. topic creation, run submission and validation, document pool generation, (cooperative) assessment, and computation and display of results. As new capabilities of SIGA we should mention the support for assessment overlap and subsequent conflict resolution process, both within the same language/collection, and across languages/collections.

To give a flavour of SIGA, we picked the assessment and the result computation facets. SIGA's assessment interface has three methods of navigation : (i) move to next/previous; (ii) move to next/previous in my list of assessments; (iii) move to next/previous item waiting to be assessed in my list of assessments.

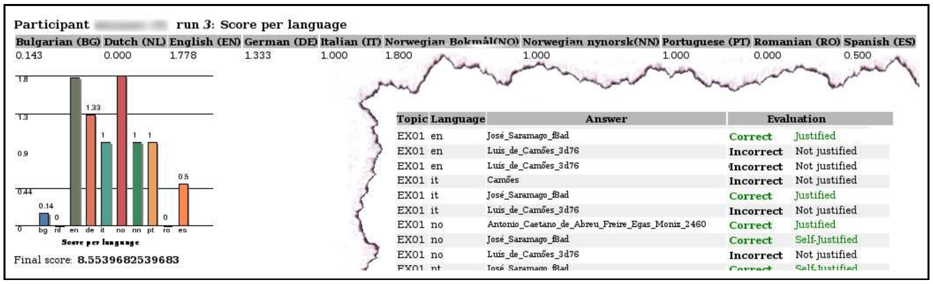
As many important tasks were dependent on JavaScript (AJAX), the interface was made compatible with the most common browsers (IE and Mozilla). An example: when assessing an answer, and to minimize waiting time for the assessors, AJAX requests were used to preview documents answers and justifications, while assessing correctness and/or the justified property (which are two different actions in the interface).

Another feature of SIGA is that it allows inspection of the (individual and aggregated) results in several tables and graphics, based on the evaluation measures adopted by GikiCLEF, as can be seen in Figure 1. (We plan to allow for the customization of these measures in future versions.)

SIGA was released with the GNU GPL open source license and we aimed at easy installation. However, given that the system was primarily built to support GikiCLEF requirements, considerable work remains to be done in the following domains: support

[4] C stands for number of correct and justified answers provided by the system in that language, N for the total number of answers that the system came up with.

[5] SIGA stands for *SIstema de Gestão e Avaliação do GIKICLEF*, Portuguese for "Management and Evaluation System of GikiCLEF". The word *siga* means "Go on!" (imperative of verb *seguir*, "continue").



Participant [redacted] run *3*: Score per language

Bulgarian (BG)	Dutch (NL)	English (EN)	German (DE)	Italian (IT)	Norwegian Bokmål(NO)	Norwegian nynorsk(NN)	Portuguese (PT)	Romanian (RO)	Spanish (ES)
0.143	0.000	1.778	1.333	1.000	1.800	1.000	1.000	0.000	0.500

Final score: **8.5539682539683**

Topic	Language	Answer	Evaluation	
EX01	en	José_Saramago_f8ad	Correct	Justified
EX01	en	Luis_de_Camões_3d76	Incorrect	Not justified
EX01	en	Luís_de_Camões_3d76	Incorrect	Not justified
EX01	it	Camões	Incorrect	Not justified
EX01	it	José_Saramago_f8ad	Correct	Justified
EX01	it	Luís_de_Camões_3d76	Incorrect	Not justified
EX01	no	Antonio_Caetano_de_Abreu_Freire_Egas_Moniz_2460	Correct	Justified
EX01	no	José_Saramago_f8ad	Correct	Self-Justified
EX01	no	Luís_de_Camões_3d76	Incorrect	Not justified
EX01	pt	José_Saramago_f8ad	Correct	Self-Justified

Fig. 1. SIGA in result mode: on the left, a graphic with language score; on the right, the assessment of each answer.

for internationalization, easy addition of more metrics and plot solutions, and dealing with collections other than Wikipedia.

We have recently added a new functionality to SIGA, namely the possibility to try out new runs and provide corresponding additions to the pool for post-campaign experiments. This is extremely relevant for participants to do fine-grained error analysis and also to allow cooperative improvement of GikiCLEF resources.

In any case, it should be stressed that it is hard to do a system that remains useful for a long time when it deals with a dynamic resource such as Wikipedia. It is well known that Wikipedia has a steady growth, which may be accompanied by changes in format or structure. For example, differences in that respect were even noticeable among GikiCLEF languages. So, while SIGA currently allows to inspect answers (that is, Wikipedia pages, stripped of images and other links) in HTML, in XML[6] as well as in the current online version (last tested November 2009), changes in Wikipedia format and directives may occur so that future adaptation of SIGA may be required, as was incidentally the case when adapting WikiXML to GikiCLEF purposes.

Table 1. Sizes of different Wikipedia collections: For GikiCLEF, in addition to count the number of pages whose name starts by *Category* we provide also the number provided by the WikiXML conversion

	INEX Collection		GikiP collection		GikiCLEF collection		
Language	No. docs	No. cats	No. docs	No. cats	No. docs	No. cats	No. cats in Wikimedia
en	659,388	113,483	2,975,316	189,281	5,255,077	365,210	390,113
de	305,099	27,981	820,672	34,557	1,324,321	53,840	53,610
pt	–	–	286,296	22,493	830,759	51,001	48,761
nl	125,004	13,847	344,418	22,110	644,178	38,703	37,544
es	79,236	12,462	–	–	641,852	65,139	60,556

To provide some quantitative data on collection size, we show in Table 1 a comparison with two previous Wikipedia-based collections for evaluation, namely the ones

[6] Converted with the WikiXML tool created by the University of Amsterdam, available from `http://ilps.science.uva.nl/WikiXML/`.

used in GikiP (from November 2006) and in the INEX collection [12], from January/February 2006. From INEX to GikiCLEF, it can be seen that Wikipedia grew up to ca. 800% for Spanish and English, as far as the number of documents is concerned. The number of categories has also grown considerably, up to almost 500% for Spanish.

4 Addressing the crosslingual and crosscultural issue

Amassing information needs In GikiP, most answers had been found in the three languages, therefore reducing the value of having a multilingual collection. So we decided to take the bull by the horns and heavily turn to culturally-laden questions, that is, questions about which one would expect a particular language or culture to display far more information than others.

In order to do that, we gathered a large organization committee with people from eight different countries/languages: there were Bulgarian, Dutch, German, Italian, Norwegian, Portuguese, Romanian and Spanish native speakers in the topic group, and we expressly requested that they came up with GikiCLEF topics that were not too global.

However, we had not foreseen that, by requiring people to choose topics of interest for their own language and culture, they would often choose those that their compatriots had carefully stored in the English Wikipedia as well, so that in fact the topic set became a sort of star with English as pivot. Table 2, borrowed and slightly modified from [13], displays the (current known) extent of the topics in the several GikiCLEF languages, by language/culture bias. One can see that most topics, no matter their cultural origin, had most hits in the English Wikipedia.

Table 2. Best languages per topic bias: in gray are the languages with the largest number of hits per topic. The rows describe the cultural topic bias as analysed by Nuno Cardoso, *none* meaning that no particular GikiCLEF language should a priori be best in it.

	total	bg	de	en	es	it	nl	nn, no	pt	ro
none	3	3	3	3	3	3	3	3	2	3
europe	6	1	4	5	1	2	2	1	2	1
bg	3	**3**	0	3	0	0	0	0	0	0
de	14	0	**10**	6	0	1	0	0	0	0
en	2	0	0	**1**	0	1	0	0	0	0
es	4	0	2	2	**1**	1	1	0	1	0
it	11	2	6	6	4	**7**	3	3	4	2
nl	6	1	1	2	0	0	**2**	0	0	0
nn, no	3	0	0	1	1	0	0	**1**	0	0
pt	2	0	2	1	0	0	0	0	**1**	0
ro	5	0	0	5	1	0	0	0	0	**2**

Guaranteeing difficult, non-trivial questions On the issue of finding user needs that required complex navigation and browsing in Wikipedia, therefore in need of automated help – that as far as we know is still not available for querying Wikipedia –, there was no doubt we succeeded.

The trouble might have been that the questions or topics were too difficult for systems as well, and thus GikiCLEF has been described as well ahead in the future. For more information on the topics and how they matched the collections, see again [13, 9]. There were nine topics for which no correct answer was returned.

The added value of crosslinguality and multilinguality Another issue was whether multilingual systems could get some value by using or reusing a comparable and parallel resource such as Wikipedia.

In one aspect, it is undeniably true that a bunch of participant systems were able, with this setup, to provide answers in languages they did not cover in any detail. This is advantageous because it shows that with a minimum work one can significantly widen the range of users one can satisfy, so we believe this should count as a GikiCLEF success.

Let us note this is not only a matter of following blindly language links from different language versions of Wikipedia (as it was almost always the case in GikiP): in fact, we were careful to provide a mechanism of crosslingual justification, in the sense that an answer was considered correct if it had a sibling which was justified. This was one could answer questions in Portuguese whose justification was only in Romanian or Bulgarian. This is obviously an added value of using other languages, even only in a monolingual setup.[7]

However, the value of processing different languages instead of English was not at all ascertained, as already described in subsection 4 and as we will show further in the next section.

Table 3. Participants in GikiCLEF 2009: *Langs.* stands for languages of participation, *NL* stands for native language of the system, if not all equally treated.

Name	Institution	System name	Langs.	NL
Ray Larson	University of California, Berkeley	cheshire	all	en
Sven Hartrumpf & & Johannes Leveling	FernUniversität in Hagen & & Dublin City University	GIRSA-WP	all	de
Iustin Dornescu	University of Wolverhampton	EQUAL	all	en
TALP Research Center	Universitat Politécnica de Catalunya	GikiTALP	en,es	en,es
Gosse Bouma & Sergio Duarte &Sergio Duarte	Information Science, University of Groningen	JoostER	du,es	du,es
Nuno Cardoso et al.	GREASE/XLDB, Univ. Lisbon	GreP	all	pt
Adrian Iftene et al.	Alexandru Ioan Cuza University	UAICGIKI09	all	all
Richard Flemmings et al.	Birkbeck College (UK) & UFRGS (Brazil)	bbk-ufrgs	pt	pt

[7] A pedantic user could wish to know in each language was it actually justified, but most users asking for list questions would be satisfied knowing that the system had justified the answer some way.

Actual participation and subsequent answer pool In fact, GikiCLEF 2009 was not able to provide a setup where seriously processing languages other than English provided a considerable advantage. The particular group of participants in GikiCLEF (see Table 3) should also in a way be held responsible for this conclusion, as we proceed to explain.

In fact, an unexpected detail in GikiCLEF that also conspired against our initial goals was that there were very few participating groups from non-English languages, which meant that the pool (the results we actually got) are much better in English. This is hardly surprising if the bulk of the processing was made in English. Figure 4 shows this clearly.

Let us stress this here: Our pool does not necessarily mean that the answers to the questions were better answered by the English Wikipedia, no matter its larger size. It is also equally a consequence of the particular group of participant systems.

More concretely, we emphasize that there were no pure Bulgarian, Italian, Norwegian or Romanian participants, which means that most answers got in those languages came from following links from other languages.[8] Likewise, there was only one Dutch and one German participant, while Spanish and Portuguese, although having more devoted participants, were not able to significantly gather more answers because of that, given that some of these dedicated systems had hardly any correct answer to contribute to the pool.

This means that, in fact and although expected otherwise, what GikiCLEF 2009 amounted to was to ascertain how well systems can answer multilingual/multicultural questions by simply processing English and following the Wikipedia links (although some systems tried the reverse as well). This is a relevant and interesting issue in itself, but it must be emphasized that it is very far from the research question we had in the first place.

5 Was GikiCLEF in vain?

The final balance we do is therefore mixed. Although the initial purpose was not achieved, several resources were gathered and deserve further study. We have also laid the foundations for organizing future venues which are similar in spirit, as well as offered a system that allows easy gathering of further empirical data.

The first and obvious lesson learned was that generalization or extension from a pilot is not free from danger. While we may have correctly diagnosed that GikiP was not interesting enough because one could get the very same data by processing only one language, the suggested fix had the opposite effect, by effectively electing English as the best language to win at GikiCLEF.

[8] This is a truth with modifications, since the UAICGIKI09 system actually processed all languages in parallel. However, its contribution to the pool was rather poor. Note also that we are not interested in the country of origin of the researchers but simply whether their systems treated in a special and knowledgeable way a particular language. When systems participated in a partially interactive run, it is even more difficult to decide what languages really were independently natively processed.

Table 4. Results in GikiCLEF 2009: The last row indicates how many participants per language, and the last column the number of languages tried in that run. Eight runs opted for all (10) languages, four tried solely 2 languages, and five one only.

System	bg	de	en	es	it	nl	nn	no	pt	ro	Score	L
EQUAL	9.757	25.357	34.500	16.695	17.391	21.657	9.308	17.254	15.515	14.500	181.933	10
GreP	6.722	12.007	13.657	11.115	8.533	8.258	9.557	11.560	7.877	6.720	96.007	10
Cheshire	1.091	9.000	22.561	4.923	11.200	9.132	3.368	7.043	4.891	7.714	80.925	10
GIRSA 1	1.333	3.125	1.800	3.000	2.250	2.250	2.000	3.000	3.000	3.000	24.758	10
GIRSA 3	3.030	3.661	1.390	2.000	1.988	1.798	3.064	2.526	2.250	1.684	23.392	10
GIRSA 2	2.065	1.540	0.938	1.306	1.429	1.299	1.841	1.723	1.350	1.029	14.519	10
JoostER 1	——	——	1.441	——	——	0.964	——	——	——	——	2.405	2
GTALP 3	——	——	1.635	0.267	——	——	——	——	——	——	1.902	2
GTALP 2	——	——	1.356	——	——	——	——	——	——	——	1.356	1
GTALP 1	——	——	0.668	0.028	——	——	——	——	——	——	0.696	2
bbkufrgs 1	—-	——	——	——	——	——	——	——	0.088	——	0.088	1
UAICG 2	0.000	0.002	0.002	0.006	0.002	0.002	0.000	0.002	0.002	0.000	0.016	10
bbkufrgs 2	——	——	——	——	——	——	——	——	0.012	——	0.012	1
UAICG 1	——	——	——	0.006	——	——	——	——	——	0.000	0.006	2
UAICG 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10
bbkuf 3	——	——	——	——	——	——	——	——	0.000	——	0.000	1
JoostER 2	——	——	——	0.000	——	——	——	——	——	——	0.000	1
Runs	8	8	12	12	8	9	8	8	11	9		

But note: we came to realise as well that GikiCLEF was very far from a realistic situation. Quite the opposite, it will strike anyone who gives it some thought that the topic collection is very far from representing any single individual or the usual needs or interests of one particular community: it is a mix of a set of ten or more individuals – each of whom has probably tried to come up with diverse questions, and not even his or her own real interests.

So, in hindsight, we may say that the GikiCLEF topic set was fairly unrealistic and that we had better concentrate on a specific area or kind of user to really test systems for a particular application. Of course, this is often something that is hard to do in the context of a general evaluation: if one wants to evaluate tasks and have a broad participation, one cannot concentrate in a too narrow (but realistic) set of users: those interested in Romanian literature, for example.

Rephrasing the problem: we had too few questions of each kind of subject / language, together with the fact that a miriad of other factors also played a non-despicable role. If we had 50 topics each of interest for one given language/culture/community, then we might be able to smooth the role of individual differences. But – just to give a striking example – there was only one question that was specifically related to a Portuguese-speaking culture (about Brazilian coastal states). So, a system working only or primarily in Portuguese would have (probably) advantage for that topic only, while for 25 topics it would have had absolutely no answer in Portuguese (according to [14]). In other words, such a system in GikiCLEF had the possibilities of getting a good score halved from the start. And the same unprivileged situation applied to Bulgarian or Dutch, even if they had 3 or 4 biased topics in the total 50.

As already noted in e.g. [15], as far as we know there has never been an investigation on the role/weight of language/culture in previous CLEF contests. Adhoc tracks were often designed to have answers in most languages, but it was hardly discussed whether these answers[9] were similar or distinct, in the sense of representing the "same" information (but just discussed or presented in a different way). So, we have no idea whether multilingual search was beneficial in those tracks, neither how contrived and/or general the topics had to be in order to be chosen, and it may well be that the conclusions and failures reported for GikiCLEF apply to these other setups as well.

As reported in [9], organizing GikiCLEF in 2009 allowed us to amass a significant number of resources as far as judgements are concerned, of which the most important were possibly the 1,009 correct and justified answers for 50 topics in 10 Wikipedias (1,621 if we count only correct, not necessarily justified). But we know that a lot of work still remains to be done to have a good evaluation resource.

All resources have been joined in the GIRA package, available from `http://www.linguateca.pt/GikiCLEF/GIRA/`, which we expect to improve in the future by delivering further versions.

In fact, we think it is important and worth while to enhance this data with actual work done by users genuinely interested in the particular topics and with native or good competence in the several languages, in order to get a better overview of the knowledge that is included in Wikipedia(s) and the upper limit that systems could get at.

Another course of action relatively easy to implement would be to provide recall measures based on the improved pool, as suggested for example by Iustin Dornescu [16].

Also, one should be able to gather a better overview of the relative difficulty of the different questions, if we were able to get this job done by human volunteers, as Ray R. Larson [17] one of the participants started to do. For example, the pool is uneven even due to the fact that the cheshire system, for lack of time, only delivered answers to the first 22 topics.

Note that there are two difficulties with the two suggestions just made, though: (i) GikiCLEF topics were most often than not meant to be discovery topics, that is, the topic owner did not know all answers beforehand, so we may never be sure about absolute recall; and (ii) many questions may require huge human labour to be answered.

Incidently, although the main target of GikiCLEF was open list questions, some closed questions were inadvertently included, and also even some with only one answer. This second issue, however, in our opinio only makes GikiCLEF more realistic, in the sense that an ordinary questioner might not know that there was a unique answer.[10]

[9] Which were documents and not strings, that is, not precise answers.

[10] Only to reject would be those questions where that was presupposed in the question formulation, such as "Who is the unique...", which we declared as uninteresting fro GikiCLEF. But we are aware that this was just an evaluation contest limitation, obviously similar (and not list) questions involving some kind of ranking are often equally interesting and important to answer, such as who was the first, which is highest, and so on, and should not be harder or different to anwer by GikiCLEF systems, were it nor for the fact that often these properties are also mentioned in the text on an entry, and have this easy shortcuts.

A final issue which in our opinion deserves further study, is to consider more carefully how far the "same" answer can be said to be given/present in different languages. In addition to the already mentioned fact mismatches reported e.g. in [1, 8], other more subtle problems concern categories: we enforced category type checking – which was often a problem for assessors as reported in [9] – but in some cases categories were not alignable across languages. For example, in some languages the category "ski resorts" was not available, even if all information was duly described in the corresponding village or mountain pages.[11] Also, cases where lexicalization was different – and thus lexical gaps exist – provide obvious problems for language linking. So, a study of the misalignability of the different Wikipedias is relevant in itself, not only for GikiCLEF-like systems, but also for the large number of other NLP systems out there who rely on Wikipedia as a multilingual or translation resource.

In a nutshell, we have made the obvious discovery that, if one wants to go beyond a quite basic simplicity level, one has to deal with all philosophical and intriguing questions that natural language understanding poses.

Acknowledgements

We are grateful to Nuno Cardoso for preparing the colections, and to the remaining organizers – Sören Auer, Gosse Bouma, Iustin Dornescu,Corina Forascu, Pamela Forner, Fredric Gey, Danilo Giampiccolo, Sven Hartrumpf, Katrin Lamm, Ray Larson, Johannes Leveling, Thomas Mandl, Constantin Orasan, Petya Osenova, Anselmo Peñas, Alvaro Rodrigo, Julia Schulz, Yvonne Skalban, and Erik Tjong Kim Sang – for hard work, supportive feedback and unfailing enthusiasm.

We also thank the larger set of further assessors – including the organizers, but also Anabela Barreiro, Leda Casanova, Luís Costa, Ana Engh, Laska Laskova, Cristina Mota, Rosário Silva and Kiril Simov – who performed the assessment, as well as Paula Carvalho and Christian-Emil Ore, who helped in an initial phase suggesting Portuguese and Norwegian-inspired topics, respectively.

Our work was done under the scope of the Linguateca project, jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC, UMIC and FCCN. We also gratefully acknowledge support of the TrebleCLEF Coordination Action, ICT-1-4-1 Digital libraries and technology-enhanced learning (Grant agreement: 215231).

References

1. Santos, D., Cardoso, N., Carvalho, P., Dornescu, I., Hartrumpf, S., Leveling, J., Skalban, Y.: GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., J.F.Jones, G., Kurimo, M., Mandl, T., Peñas, A., Petras, V., eds.: 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) 894–905

[11] In that case we had to relax the type checking constraint during assessment and conflict resolution modes.

2. Santos, D., Rocha, P.: The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers. Springer, Berlin/Heidelberg (2005) 821–832
3. Santos, D., Cardoso, N.: Portuguese at CLEF 2005: Reflections and Challenges. In Peters, C., ed.: Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005), Vienna, Austria, Centromedia (21-23 September 2005) no pp
4. Santos, D., Costa, L.: QolA: fostering collaboration within QA. In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers, Berlin / Heidelberg, Springer (2007) 569–578
5. Zobel, J.: How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In: SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM (1998) 307–314
6. Voorhees, E.M., Buckley, C.: The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. (2002) 316–323
7. Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF Pilot Track Overview. This volume.
8. Santos, D., Cardoso, N.: GikiP: Evaluating geographical answers from Wikipedia. In: Proceedings of the 5th Workshop on Geographic Information Retrieval (GIR'08), Napa Valley, CA, USA (30 October 2008) 59–60
9. Santos, D., Cabral, L.M.: GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia. In Borri, F., Nardi, A., Peters, C., eds.: Cross Language Evaluation Forum: Working notes for CLEF 2009. (30 September - 2 October 2009)
10. Dussin, M., Ferro, N.: Direct: applying the dikw hierarchy to large-scale evaluation campaigns. In Larsen, R.L., Paepcke, A., Borbinha, J.L., Naaman, M., eds.: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, ACM 424–424
11. Lalmas, M., Piwowarski, B.: INEX 2006 relevance assessment guide. INEX 2006 Workshop Pre-Proceedings (2006) 389–395
12. Denoyer, L., Gallinari, P.: The Wikipedia XML corpus. ACM SIGIR Forum **40** (14 April 2006) 367–272
13. Cardoso, N.: GikiCLEF topics and Wikipedia articles: did it blend? In: CLEF2009 Workshop, Corfu, Greece (30 September - 2 October 2009)
14. Cardoso, N., Batista, D., Lopez-Pellicer, F., Silva, M.J.: Where in the Wikipedia is that answer? The XLDB at the GikiCLEF 2009 task. In Borri, F., Nardi, A., Peters, C., eds.: Cross Language Evaluation Forum CLEF 2009 Workshop. (30 September - 2 October 2009)
15. Santos, D., Cardoso, N.: Portuguese at CLEF. In Peters, C., Gey, F., Gonzalo, J., Müller, H., Jones, G.J., Kluck, M., Magnini, B., de Rijke, M., eds.: Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised selected papers. Volume 4022 of LNCS., Berlin, Springer (2006) 1007–1010
16. Dornescu, I.: EQUAL Encyclopaedic QA for Lists. In: CLEF2009 Workshop, Corfu, Greece (30 September - 2 October 2009)
17. Larson, R.R.: Interactive probabilistic search for GikiCLEF. In Borri, F., Nardi, A., Peters, C., eds.: Cross Language Evaluation Forum: Working notes for CLEF 2009, Corfu, Greece (30 September - 2 October 2009)