GikiCLEF: Crosscultural issues in multilingual information access

Diana Santos*, Luís Miguel Cabral*, Corina Forascu, Pamela Forner***, Fredric Gey****
Katrin Lamm*****, Thomas Mandl*****, Petya Osenova£, Anselmo Peñas$,
Álvaro Rodrigo$, Julia Schulz*****, Yvonne Skalban&, Erik Tjong Kim Sang%**

*SINTEF ICT, **UAIC, ***CELCT, ****Univ. Berkeley, *****Univ. Hildesheim
£BTB group, IPOI, BAS, $UNED, &Univ. Wolverhampton, %Univ. Groningen
Diana.Santos@sintef.no, Luis.M.Cabral@sintef.no, corinfor@info.uaic.ro, forner@celct.it, gey@berkeley.edu
katrin.lamm@uni-hildesheim.de, mandl@uni-hildesheim.de, petya@bultreebank.org, anselmo@lsi.uned.es
alvarory@lsi.uned.es, julia-maria.schulz@uni-hildesheim.de, Yvonne.Skalban@wlv.ac.uk, erikt@xs4all.nl

Abstract

In this paper we describe GikiCLEF, the first evaluation contest that, to our knowledge, was specifically designed to expose and investigate cultural and linguistic issues involved in multimedia collections and searching, and which was organized under the scope of CLEF 2009. We present the task, its motivation, the results of the participants and the GIRA resource that is offered to the community for training and further evaluating systems with the topics gathered. We end the paper with some discussion of what we learned and possible ways to reuse the data.

1. Introduction

In this paper we describe the first evaluation contest that, to our knowledge, was specifically designed to expose and investigate cultural and linguistic issues involved in structured multimedia collections and searching, and which was organized under the scope of CLEF 2009[1].

In a nutshell, GikiCLEF[2] provided 50 topics developed with non-English users in mind, to evaluate systems that should answer open[3] natural language questions to Wikipedia, using the multilingual and crosslingual properties of this resource. Languages dealt with were: Bulgarian, Dutch, English, German, Italian, Norwegian (both Bokmål and Nynorsk writing standards, since they constitute distinct Wikipedia collections), Portuguese, Romanian, and Spanish.

GikiCLEF was a follow-up from the GikiP pilot (Santos et al., 2009), organized the previous year as a pilot under GeoCLEF (Mandl et al., 2009), and which addressed 15 questions to the German, English and Portuguese collections (snapshots of Wikipedia created in 2006).

While some features of GikiP were retained, namely providing (manually created or translated parallel) questions in all languages, and rewarding answers in more than one language, there were a number of important and interesting innovations in GikiCLEF (which actually warrant the name change), in addition to a larger number of topics, more languages and larger collections:

- a multilingual, multicultural committee (the authors of the present paper) was set up to come up with distinctly hard, and culturally-relevant, topics;
- a complex support system, SIGA, was deployed to allow for cooperatively managing many subtasks – the system is open source and available for the community, and its development largely benefited from user input and extensive use, specially during assessment;
- a different evaluation measure was introduced to deal with the ten collections and not requiring, although preferring, answers in all languages;
- provision for inter-assessor validation, which resulted in a much higher quality of the final evaluation resources created;
- the possibility to address justification issues that required more than one page/document to be returned.

All these matters will be described in turn. But first we present the task itself with complex examples, and delve into its motivation.

2. The task

2.1. Basics

The task in which we wanted to evaluate systems' performance was that of using Wikipedia to answer open list questions, that is, questions that have a variable number of answers not known in advance to the questioner.

The answers, in addition, to obey a realistic requirement from real life, had to be justified – in the sense that a human user should be able to confirm the answer correctness by simply visiting a set of Wikipedia pages.[4] In order to simplify the issue of result presentation, answers would have to be themselves entries in Wikipedia.[5]

It should be emphasized that the proposed task was hard for both men and machines because it requires browsing

[1] http://www.clef-campaign.org/

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

[3] Open in the sense of not knowing the number of right answers in advance.

[4] In addition we should stress that – as is the rule in evaluation contests – we consider correct an answer grounded in the collection, we are not after absolute truth.

[5] So, technically, the task was a merge of question answering and information retrieval because both the short answer and a document (or more) would be retrieved.

EX01 pt/s/a/r/Saramago.xml {pt/j/o/s/Categoria José_Saramago_8b43}
EX01 ro/j/o/s/José_Saramago_f8ad.html {}
EX01 ro/j/o/s/José_Saramago_f8ad.html {}
EX01 de/j/o/s/José_Maria_Eça_de_Queiroz_3766.html {}
LI13 en/o/t/o/Otocinclus_cocama.xml {en/c/o/c/Cocama_language.xml }
EX09 pt/g/u/a/Guaranis.xml {pt/l/í/n/Línguas_indígenas.xml, pt/l/í/n/Língua_guarani.xml}
EX09 pt/c/o/c/Cocamas.xml {}
EX09 pt/c/o/c/Cocamas.html {pt/l/í/n/Língua_cocama.html}
EX09 en/o/t/o/Otocinclus_cocama.xml {en/c/o/c/Cocama_language.xml }
EX09 it/c/o/c/Cocama-Cocamilla_24dc {}

Figure 1: Example format of a GikiCLEF submission: topic id, collection-id, justification inside brackets

and reading of a large number of documents and subsequent filtering for finding the only ones applicable.
The task, in addition, was weakly "multimedia" (joining textual and visual clues) because it was designed for satisfying people, which means that questions could be evaluated and accepted as useful by looking at maps, photos[6] or even making some use of common sense. That is, the information had not necessarily to be presented (only) by textual means, since the context of the task was user access / interaction with Wikipedia.[7]
In order to make systems invest on multilinguality, the evaluation score favoured (in fact, overemphasized) the existence of answers in more than one language. Provision was made not to harm systems if there were no answers in other languages. Here is how the score was defined:

- C: number of correct (that is, justified in at least one language) answers for the set of the 50 topics
- N: total number of answers provided by the system for the set of the 50 topics
- GikiCLEF score per language: C*C/N (so one has a score for de, pt, etc, as $C_{de} * C_{de}/N_{de}$, $C_{pt} * C_{pt}/N_{pt}$, etc.)

The final score was computed by adding every language score.

2.2. Motivation

As explained in some detail in previous papers (Santos and Rocha, 2005; Santos and Cardoso, 2005; Santos and Costa, 2007), we organizers were often unhappy with the tasks used for system evaluation, for various reasons: because these tasks tend to be artificial, have no concrete user model, and their rationale seems too often associated with a concrete system or research project, lacking a clear connection with real tasks.
So, for example, question answering evaluation contests often ask questions which are too easy or too difficult. In addition, this kind of task is often hard to understand without having a user context. While an evaluation setup has obviously to be always an approximation of (random or average) user behaviour, the lack of realistic evaluation resources is also a problem that we wished to address.
Asking Wikipedia, which is one the most visited sites on the Web according to (Alexa, 2010), appealed to us organizers, since it seems to be a natural everyday task.
As to multilinguality, an additional problem arises for evaluation. Namely, how to devise task(s) or goals that make sense to be done multilingually or crosslingually instead of just doing things in parallel for different languages.
There have been several ingenuous proposals – see for example WebCLEF (Balog et al., 2007), WiQA (Jijkoun and de Rijke, 2007) and iCLEF (Artiles et al., 2007) – but we believe one has always to make a choice between either (i) the same content in several languages, or (ii) different content in different languages.
While the first choice is obviously best for comparing performance across systems that work for different natural languages, the second offers a far more realistic motivation to go multilingual in the first place: For, if one had all content in one's own language, why would one need to process the other languages? So in GikiCLEF we chose a setup where we expected that different languages would be able to provide added value (and information) to a user question.
Now, this is seeing the world from the point of view of a shared task organizer. Commercial companies may be happy in doing everything in parallel to satisfy their customers in different languages if there is a market for it, or translate the entire content to one particular language. But our wish with GikiCLEF was to devise a truly multilingual/crosslingual task with clear advantages in processing different languages. And, in fact, another argument for this can be adduced: at least in an European context, the possibility of the users being themselves profficient to a greater or lesser extent in many languages is a real one, and therefore it makes sense to have a GikiCLEF-like system providing an answer list in several languages.[8]
The full topic list and the example topics are available from the website and were also published in (Santos and Cabral, 2009). (Cardoso, 2010; Santos et al., 2010) presented some preliminary analysis of the topics, focussing on number of answers, language bias, type of answer required, and potential relationship with geographic information retrieval.

[6]Examples are: to assess left afluents of a river, to check whether mountains had snow, or to find out the colour of a flag.

[7]In fact, during human assessment, justifications were also found, for example, in one page's reference list...

[8]Note that corresponding articles in different languages are far from verbatim equivalents of each other: For example, while the German article on the (German) river Pader is very elaborate, the English one consists of only two sentences. So an English user who could read German would have a definite advantage if interested in that river.

In order to comply with the requirement of a task which would benefit from harvesting answers in different languages, and also because we expected different language Wikipedias to correspond to different cultural on-line communities in different languages[9], we were looking for a set of topics which should reflect different tastes and subject matters in different languages.

In addition, and since some of us at least do not adhere to the assumption that everything is equally well translatable, or conveyable, in every language, we have tried to elicit really culturally-laden topics, hence hard to translate, explain or even understand in other cultures or languages. For concreteness's sake, let us provide some examples of the difficulties involved: For example, *Spanish guitar* is a technical term in music that is probably not the best way to translate *violão*, the Brazilian Portuguese (original) term. Translation from the English translation into other languages would probably add a spurious *Spanish* adjective. Another case: to render the Norwegian *oppvekstroman* requires the clarification that this is close, but not exactly the same as what, in English, literature experts use the German (!) term *Bildungsroman* to express. Similarly, Romanian *balade* is probably a false friend with Spanish *balada*, and had to be translated by *romance*. Interestingly, this is again a false friend with Portuguese *romance*, which denotes what in English is called a *novel*... which, to completely close the circle, is **not** what is called *novela* in Portuguese!

Language is just one facet of culture. We are of course aware that there are cultural differences also between people interacting in the same language, see e.g. (Gumperz, 1996), and that there are other elements of culture which are not primarily visible in language, such as those studied in (Mandl, 2010). We nevertheless believe that the use of the adjective "crosscultural" for GikiCLEF is warranted because the topics chosen often made more sense to some cultures than others – or at least this was one of the criteria for their choice.

2.3. Examples of the reasoning behind topic choice

As an information consumer, we often find interesting facts about which we would like to learn more. Three of the Dutch topics were proposed with this scenario in mind.

First, for example, a Dutch music fan might discover that in 1979 young Dutchman Jaap van Zweden (19) became concertmaster of the Royal Concertgebouw Orchestra in Amsterdam. He might wonder if any other Dutch people held this position in the previous century.[10]

Second, a historian might be surprised to discover that while province capital The Hague obtained city rights in 1806, other province capitals like Haarlem (1245) and Leeuwarden (1285) obtained these rights much earlier. He could like to know if there were other province capitals that obtained these rights before 1300.

Third, a cycling fan learns that the record number of wins in the Tour of Flanders race is three times. He might wonder if there were cyclists that won the race twice, and who they were.

Yet another Dutch topic was created with a user in mind that is planning a trip to Flanders for a small group of people. She wants to include in the trip a dinner at an exclusive restaurant. Hence she wants to know which Flemish cities host exclusive restaurants (with two or three Michelin stars) since the restaurant location will have an influence on the city they will visit during the trip.

On the other hand, the choice of Bulgarian topics was made on the basis of cultural issues that had a big impact outside Bulgaria, and so were in fact often concerned with this impact (Beinsa Duno's ideas in a "outside Bulgaria" context, a fighter with the Diamond belt, a football player so famous that there are bands named after him, etc.). So, for these topics, a criterion was to have them well covered in other Wikipedias. It is relevant to note that this was not necessarily the case for other topic language(s), and corresponds to the individual choice of the researchers, and how they saw the GikiCLEF task.

Still in other cases, question choice was in fact due to practical experience with a particular user group, as happened with two widely different German topics:

Canoeists often go on weekend trips and, because time is limited, they prefer not to travel very far to get to the river; and they do not want the trip to take too long (hence a plausible river length restriction).

Students of literature, on the other hand, and given the hypothesis that Goethe used his own experiences for characters in his books, may plausibly want to visit, or at least read more about, the places where Goethe fell in love, in order to understand better the works and their settings.

Finally, two of the Italian topics have been created with a user in mind being a tourist visiting Italy and having some interest in knowing and tasting Italian food and specialties. The cassata, for example, is a typical, traditional cake from Sicily, and one might be interested in knowing how it is prepared and which are its ingredients. Likewise, a wine connoisseur coming to Italy will undoubtedly have heard about Chianti, a famous red wine produced in Tuscany, and may want to visit the places where it is actually produced.

2.4. From a participant point of view

The Wikipedia snapshots – henceforth referred to as the GikiCLEF collection – were made available December 2008, both in HTML and in XML, to cater for different participants preferences.

Participants had to fetch the topic set in XML format (the 50 topics were made available in all ten languages), from 15 May 2009 12:00 GMT until 31 May 2009, and had exactly five days to upload the result runs (maximum of three runs). The run format is illustrated in Figure 1.

Participants knew that only justified answers would be counted as correct, but that it was enough that justification were found in one language only. So, once a correct as well as justified answer was found in one language, to return all other aligned answers in different languages would be an obvious way of improving the system's score, which seems

[9] In addition to Veale's remark that the global Wikipedia has an obvious bias on science fiction and imaginary worlds due to the cultural preferences of its contributers mass (Veale, 2007).

[10] Interestingly, the topic owner's original hypothesis was that there would be few Dutchmen in this position, which turned out not to be the case.

Name	Institution	System name	Langs.	NL
Ray Larson	University of California, Berkeley	cheshire	all	en
Sven Hartrumpf &	FernUniversität in Hagen &	GIRSA-WP	all	de
& Johannes Leveling	& Dublin City University			
Iustin Dornescu	University of Wolverhampton	EQUAL	all	en
TALP Research Center	Universitat Politécnica de Catalunya	GikiTALP	en,es	en,es
Gosse Bouma &	Information Science,	JoostER	du,es	du,es
& Sergio Duarte	University of Groningen			
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) &	bbk-ufrgs	pt	pt
	UF Rio Grande do Sul (Brazil)			

Table 1: Participants in GikiCLEF 2009: *Langs.* stands for languages of participation, *NL* stands for native language of the system, if not all equally treated.

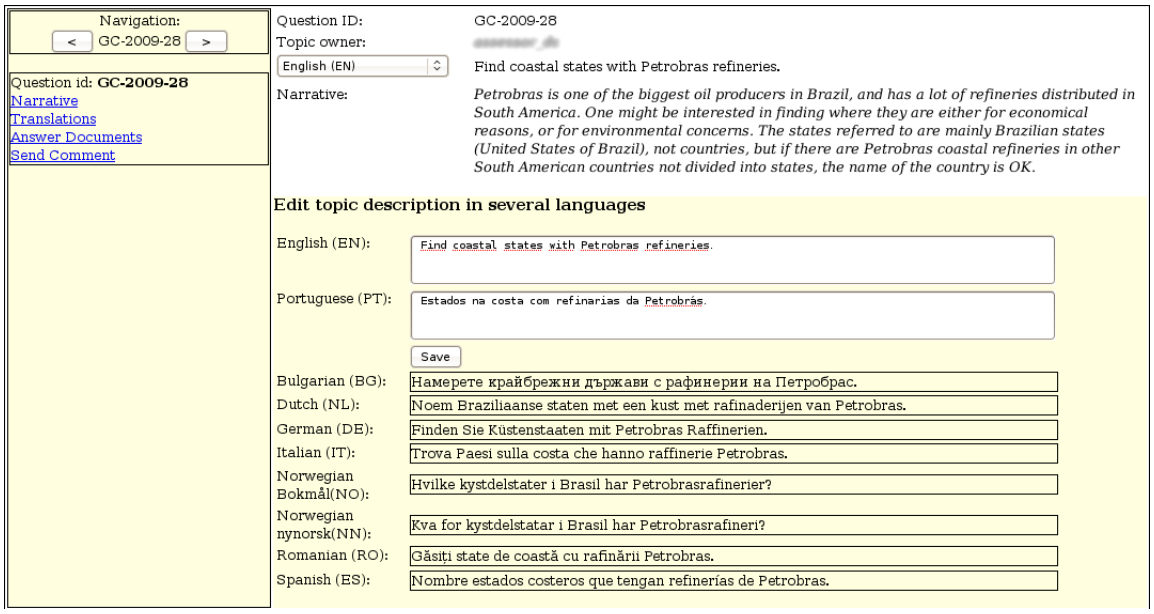

Figure 2: SIGA interface for creating topics: The screenshot was taken after the topic had been translated, which is of course artificial in that during creation the other language slots are void. (Reprinted from (Santos and Cabral, 2009).)

to have been what most people did.

However, there were only two participants who provided justification pages, which means that the correct answers, when found, were in the vast majority of cases self-justified. This fact made GikiCLEF more akin to pure information retrieval than we had presumed.[11]

Although almost thirty interested parties enrolled in the beginning, we had only eight participants that actually submitted seventeen runs for the task (see Table 1).

3. Description of SIGA

Although (Santos and Cabral, 2009) already offers a thorough description of the work behind the scenes, we provide here a sketch of the many tasks that had to be organized.

Since there was a considerable number of people creating topics in different languages, and an even larger set of assessors (30) after submissions had been sent in, there was a need for a computational environment to manage the large amounts of data, and also to provide an inspection facility against the collections, for both topic owners and assessors. So, SIGA, standing for *SIstema de Gestão e Avaliação do GIKICLEF*[12] in Portuguese, was developed, offering different actions for five separate roles: manager, topic developer (owner or other), participant, assessor (basic or conflict resolver), and simple observer. SIGA was in charge of several procedures, such as validation of runs, pool creation, assessment distribution, conflict detection, scores computation, and display of comparative results.

Details on the topic creation mode and the possibility of viewing the collection for checking the existence of answer candidates can be found in (Santos and Cabral, 2009); see also Figure 2. Note that, after topics had been created and translated into English, together with the "user model"/narrative in English, they had to be translated into each of the other eight languages, and possible answers in

[11] In any case, it is important to note that an answer could be self-justified in one language and not in another, since the information of "parallel" pages often differed widely.

[12] GikiCLEF Management and Evaluation System

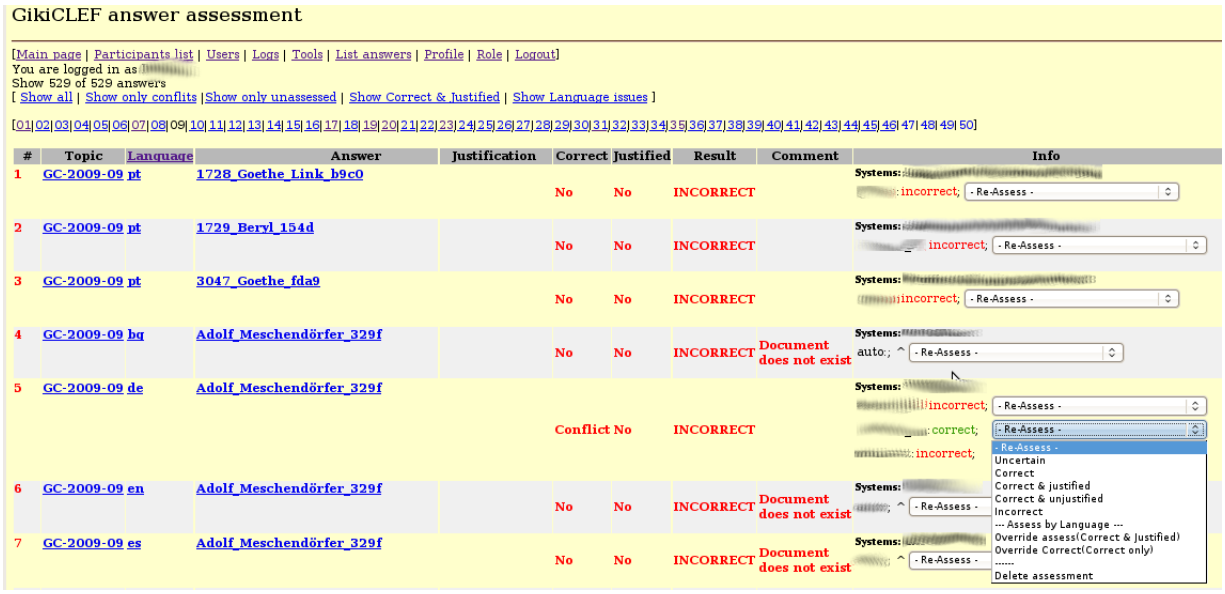

Figure 3: Solving conflicts about assessement of the same topic in the same language

other languages had to be added to the "expected answer pool", tagged as self-justified or not.

The process of (monolingual) assessment was also described and discussed in (Santos and Cabral, 2009). Prior to it, answers not in the collection or corresponding to a filetype not accepted (such as disambiguation list or image) were automatically discarded. Also, answers already occurring in the expected answer pool and which had been considered by the topic owners as self-justified were automatically judged correct, while those which were not self-justified were marked as correct and not justified. This entailed a significant reduction in assessment work, as can be appreciated from the numbers of Table 2.

Answers received	21,251
Different answers	18,152
Different answers with complex justification	215
Different manually assessed answers	6,974
Manual assessments	10,332
Automatically assessed answers as incorrect	10,588
Automatically assessed answers as correct	283
Answers resulting in conflicts	383
Correct and justified answers	1,327
Correct but not justified answers	1,415

Table 2: Numbers on the assessment process.

After the simple assessment was concluded, and since a large percentage of answers had been assigned to more than one assessor, automatic discovery of conflicts took place, and conflict resolution was performed. The conflict resolution mode of SIGA can be seen in Figure 3.

We believe that, at least within CLEF, this was the first case where multiple assessments were used, and in fact the number of initial conflicts was so large that we found out that the assessment guidelines (and the task itself) were not clear enough: for example, the type checking that we took for granted while devising topics was not accepted or understood by many participants and assessors, and this led to a massive reassessment. We discuss in the final section the consequences this has for the task definition and to the possibility of actually devising a reliable and realistic evaluation set.

After monolingual conflict resolution had finished, alignment between answers in different languages was performed, and a second kind of conflict resolution had to be carried out, namely between cases where different assessors had concluded different things based on material in different languages.

This had to be specially tailored to have two kinds of cases as exceptions:

- those where different languages Wikipedias actually contradicted each other: in that case, no propagation, alignment, or other-language justification was possible;
- one case where by mistake there was a different requirement (for river length) in one language: then obviously answers in different languages were not comparable.

The results were then finally computed. Only correct and justified answers were considered for the systems' scores, although there were several cases of correct but not justified answers.

We repeat that, once justified in one language, an answer would be considered correct and justified in all the others. From a score computation point of view, this means that the very same unjustified answer in one language (Wikipedia) could be considered correct and justified for one run (and thus getting full score) and not correct (because no justification was provided) for another (and thus not bewing rewarded by our score), depending on the set of multilingual answers returned by that run. We did not consider, however, cases of cross-language justification, in the sense of

having the justification on one language to be provided in another language: a justification (set) was a set of pages in one Wikipedia that together provided an answer, as can be seen in Figure 1.

4. Results

The results obtained by the participants are shown in Table 3.

The resulting resource (a sizeable number of correct answers to each topic, in many languages) was made publicly available in November 2009 from `http://www.linguateca.pt/GikiCLEF/GIRA/`.

In short it contains the collections, the topics, the assessments, the results, the programs, and a number of documentation issues related to the particular topics and their assessment.

Table 2 shows a quantitative description of the most important data regarding the assessment effort which have an obvious bearing on the resource size. By "different answers" we mean answers together with justification lists that are unique.

As to language variety, Figure 4 displays the different number of answers per language in the pool.

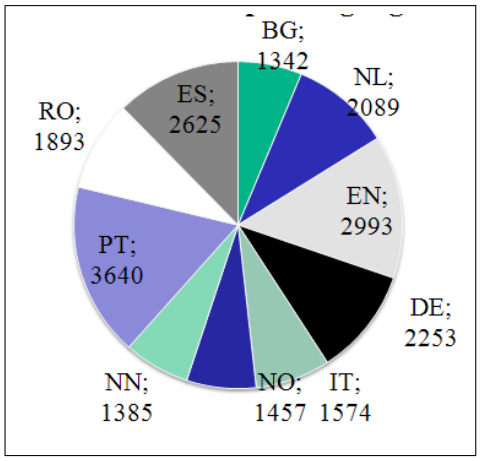

Figure 4: Answers per language returned by the set of all systems

Figure 5 represents the correct answers in GikiCLEF per language.

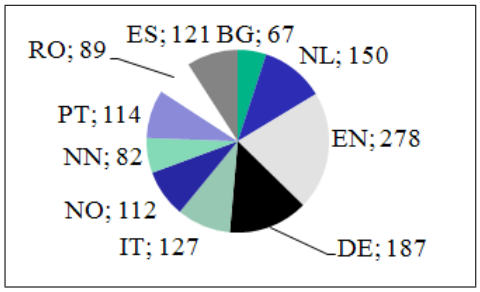

Figure 5: Correct answers per language returned by the set of all systems

The two figures show that the material gathered in GikiCLEF can be used to train, deploy or at least evaluate IR and question answering systems in the ten languages. In addition, it should be easy to augment the data / pool for each language and customize the SIGA system for particular domains, languages or tasks, and we hope to be able to do this in the future.

5. What was learned

Let us now provide a critical assessment of GikiCLEF, touching upon what could have been improved, and what can be considered design flaws.

As discussed in detail in (Santos and Cabral, 2010), the fact that English was a pivot language, both in the GikiCLEF team and in Wikipedia in general, caused a tremendous bias towards English, which became the language with by far more justified answers. This made it possible to to reach a relatively high score at GikiCLEF by just processing English, which is no doubt a clear design flaw of GikiCLEF: In fact, we produced, by juxtaposition of different (and hopefully) realistic users, a "non-existent" multi-cultural user who was equally well versed (and interested) in Bulgarian religious leaders and American museums featuring Picassos.

So, our current conclusion is that further organization of GikiCLEF-like contests has to give more weight to one or two cultures and not to ten or more.

We also observed that there were hardly any current systems – at least among the participants – which were able to do the task. So, probably not much was gained by organizing GikiCLEF with such high stakes.

Another problem with the topic choice was the quality of the related Wikipedia pages. Even though the topics were carefully selected by us organizers, and so there was good material on the particular subjects at least on the language of the topic owner (and probably in English as well), the fact that none of us was multilingual in the other nine languages – and, of course, had not enough in-depth knowledge of all the subjects – prevented real quality control of all the possible answers/Wikipedia entries. So, pages in languages other than the topic owners' were often of bad quality or had wrong data. The fact that this happened to a level of contradiction for three of the 50 topics (6%) is something that is also relevant: there is still a lot of rubbish in Wikipedia.

Another interesting issue came up during assessment, that made us reflect on the task definition itself: how important or relevant for a natural language processing task is strict type checking or type correctness? In other words, a useful answer for a user, although not exactly to the point (according to logically strict principles), seems to be preferable to a perfectly logically correct answer which is however redundant with a previous one. And so many participants (and some assessors) complained that an answer such as "flag of Argentina" should have been considered correct, or almost correct, if one issued a question such as "Which countries have flags such and such?", instead of considering it incorrect because the type required was "country".

Also, it was obvious that some answers were much easier to justify than others, and that some answers were "direct" while others required a number of complex cycles of indirection. A thorough study of the difficulty of each

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		

Table 3: Scores per language and total score. 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.

topic given the collection is thus needed to understand more closely the reasons and the requirements for system behaviour.

We intend to annotate the resources (topics and answers in each language) with this information, in order to see whether they correlate in any way with system's behaviour and whether these are relevant features to assign in future evaluations.

Finally, we have requested from prospective and actual GikiCLEF participants an answer regarding whether they wanted also to address within GikiCLEF the following issues:

- Improve presentation of the results: To devise user-friendly systems, an unordered list of answers is often not enough, especially when multiple answers can be related. So, from the point of view of the scoring procedure, one might reward ordered lists (for instance by granularity given a particular ontology, or by time if the question concerns a particular temporal journey).
- Investigate geographical diversity: Another subject that is now receiving some attention is how to take geographical diversity into account: depending on the kind of topic, one might want to boost diversity instead of mere quantity. In fact, for some users and uses, returning too (geographically) close hits may be considered annoying instead of relevant.

Although no system was prepared to work in either regard, we still believe they are interesting alleys to explore.

Another related subject on which we are aware a lot can be done to improve a contest of the GikiCLEF kind is devising more appropriate and complex evaluation measures, also taking into account recall-oriented measures, and difficulty estimates for different kinds of topics. We believe that experimentation with other measures is made easy by the availability of the GIRA resource, where different scoring procedures can be implemented and its impact evaluated on the actual runs. Although we have no space here to provide a full overview of the participants' approaches, we would like to state that they have shown a wide variety of different methods and priorities, as had already been the case in GikiP even with only three participants. So both semi-interactive approaches, using a human-in-the-loop, semantic-oriented QA systems, and IR traditional methods were used to try to get at the answers in GikiCLEF 2009.

Our conclusion is therefore a positive one: although we might have been too ambitious for the state of the art, GikiCLEF has shown that it is possible to implement systems that answer in many languages, by using a multilingual collection. Also, our work has produced a resource that can be further used in the development of Wikipedia-based information access systems in the years to come.

Acknowledgements

We thank the remaining GikiCLEF organizers, Sören Auer, Gosse Bouma, Iustin Dornescu, Danilo Giampiccolo, Sven Hartrumpf, Ray Larson, Johannes Leveling, and Constantin Orasan; the other assessors, Anabela Barreiro, Leda Casanova, Luís Costa, Ana Engh, Laska Laskova, Cristina Mota, Rosário Silva, and Kiril Simov; Paula Carvalho and Christian-Emil Ore for help on Portuguese and Norwegian topics respectively; and of course the participants, without whom GikiCLEF would not have existed.

Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC, MCTES, 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) for GikiCLEF assessment.

Álvaro Rodrigo has been partially supported by the Education Council of the Regional Government of Madrid and the European Social Fund.

6. References

Alexa. 2010. Alexa top 500 sites. `http://www.alexa.com/site/ds/top_sites?ts_mode=global`.

Javier Artiles, Julio Gonzalo, Fernando López-Ostenero, and Víctor Peinado. 2007. Are Users Willing to Search Cross-Language? An Experiment with the Flickr Image Sharing Repository. 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 195–204, Berlin. Springer.

Krisztian Balog, Leif Azzopardi, Jaap Kamps, and Maarten de Rijke. 2007. Overview of WebCLEF 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 803–819, Berlin. Springer.

Nuno Cardoso. 2010. GikiCLEF topics and Wikipedia articles: Did they blend? In Carol Peters et al, editor, *Multilingual Information Access Evaluation, VOL I: Text Retrieval Experiments*. Springer, Setembro.

John J. Gumperz. 1996. The linguistic and cultural relativity of conversational inference. In J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 374–406, Cambridge. Cambridge University Press.

Valentin Jijkoun and Maarten de Rijke. 2007. 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.

Thomas Mandl, Paula Carvalho, Fredric Gey, Ray Larson, Diana Santos, and Christa Womser-Hacker. 2009. GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, pages 808–821. Springer.

Thomas Mandl. 2010. Cultural and International Aspects of Social Media. In Panagiotis Papadopoulou, Panagiotaand Kanellis and Drakoulis Martakos, editors, *Handbook of Research on Social Computing Theory and Practice Interdisciplinary Approaches*. Idea Group Reference.

Diana Santos and Luís Miguel Cabral. 2009. GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia. In Francesca Borri, Alessandro Nardi, and Carol Peters, editors, *Cross Language Evaluation Forum: Working notes for CLEF 2009*, 30 September - 2 October.

Diana Santos and Luís Miguel Cabral. 2010. GikiCLEF : Expectations and lessons learned. In Carol Peters et al, editor, *Multilingual Information Access Evaluation, VOL I: Text Retrieval Experiments*. Springer, September.

Diana Santos and Nuno Cardoso. 2005. Portuguese at CLEF 2005: Reflections and Challenges. In Carol Peters, editor, *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005)*, Vienna, Austria, 21-23 September. Centromedia.

Diana Santos and Luís Costa. 2007. QolA: fostering collaboration within QA. 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. Alicante, Spain, September, 2006. Revised Selected papers*, pages 569–578, Berlin / Heidelberg. Springer.

Diana Santos and Paulo Rocha. 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. 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, Bath, UK, September 15-17, 2004, Revised Selected Papers*, pages 821–832. Springer, Berlin/Heidelberg.

Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, pages 894–905. Springer.

Diana Santos, Nuno Cardoso, and Luís Miguel Cabral. 2010. How geographical was GikiCLEF? A GIR-critical review. In *6th Workshop on Geographic Information Retrieval (GIR'10)*, 18-19 February.

Tony Veale. 2007. Enriched Lexical Ontologies: Adding new knowledge and new scope to old linguistic resources. In *European Summer School on Language, Logic and Information (ESSLLI 2007)*.