SIGA, a System to Manage Information Retrieval Evaluations

Luis Costa, Cristina Mota, and Diana Santos

Linguateca/FCCN and University of Oslo
luis.f.kosta@gmail.com, cmota@ist.utl.pt, d.s.m.santos@ilos.uio.no

Abstract. This paper provides an overview of the current version of SIGA, a system that supports the organization of information retrieval (IR) evaluations. SIGA was recently used in Págico, an evaluation contest where both automatic and human participants competed to find answers to 150 topics in the Portuguese Wikipedia, and we describe its new capabilities in this context as well as provide preliminary results from Págico.

Keywords: Information extraction, information retrieval, evaluation, question answering, usability, wikipedia.

1 Introduction

SIGA is a web-based management and evaluation system supporting the organization of Information Retrieval (IR) evaluations, distributed by Linguateca, and included in the GIRA package[1]. Its source code is open, so anyone can improve and extend it to the particular requirements of a specific IR evaluation.

The need for this computational environment arose during the organization of GikiCLEF [1,2], because there was a considerable number of people creating and assessing topics in geographically distinct sites, dealing with large amounts of data (the Wikipedia collections for the several languages involved and many systems' submissions). SIGA has a similar structure to other systems such as DIRECT [3] or the system used in INEX to back the evaluation [4], and supports multiple user roles for different tasks. Different choices and privileges are thus available, namely topic creation, run submission and validation, document pool generation, (cooperative) assessment, system scoring and display of results. Compared to these two systems, SIGA offers an important additional capability introduced in the context of GikiCLEF: the support for topic assessment overlap (several judgements for the same answer) and the semi-automation of the subsequent conflict resolution process.

More recently, in 2011, SIGA was adapted and extended to support the organization and participation in Págico[2] [5], an evaluation contest in information retrieval in Portuguese organized by Linguateca [6] whose goal was to evaluate systems aiming to find non-trivial answers to complex information needs in Portuguese, and is a follow-up of GikiCLEF that builds on Linguateca's previous experience but focuses on a specific

[1] http://www.linguateca.pt/GikiCLEF/GIRA/

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

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 284–290, 2012.

cultural sphere (the Portuguese-speaking one) instead of cross-linguality or geographical subjects. Three main capabilities were added: automatic assessment of answers and justifications, an interface for human participants, and navigation inside a static version of Wikipedia. The current paper, although also describing SIGA, focuses on the new features required by this evaluation. For details about earlier versions and uses of SIGA, please check [1,2] or the GikiCLEF site.[3]

2 Technologies Used and Functionalities

SIGA is developed mainly in PHP and JavaScript, with data stored in a MySQL database. The choice of these technologies was driven by the following requirements:

Easy Installation. As users of the system would range from hard-core software developers to computer users with just basic knowledge, a web application was chosen, with no need for local installation.

Intuitive Interface. Catering for the broad spectrum of users, the interface should be satisfactory for the different types of users, especialy when human participants are concerned.

Ability to Deal with Large Amounts of Data, due to the large size of the document collections, of the results created by the participants and of the evaluation data created by the evaluators.

Topic Creation. SIGA supports the creation of topic sets for IR evaluations, helping topic managers look for answers in titles of Wikipedia documents included in the collections.

From the point of view of system evaluation, SIGA had a major shortcoming: it did not support adding justifications during topic creation, which entailed that the pre-assessment was only based on the comparison of answers, putting the burden of assessing the justifications on the human assessors. We improved SIGA so that it now allows the addition of justifications while creating the topics, so that the automatic pre-assessment can also be based on the justifications.

SIGA was also modified to allow a two level categorization of topics. The topic creators can associate one or more classes to each topic, and group those classes into major thematic subjects, which are then presented to the participants during the evaluation context proper.

Support for the Participation of Automatic Systems. The interface for system participation allows: (i) the download of different topic sets (evaluation, examples, testing, training), (ii) the validation and submission of runs, (iii) the inspection of the system scores and (iv) the comparison of results with the other systems. These tasks are somewhat similar to those performed by SAHARA [7], which provides a comparison between the scores of the new submitted runs and the runs officially submitted to HAREM [8].

[3] `http://www.linguateca.pt/GikiCLEF/`

Interface for Human Participation. Due to the considerable number of topics (three times more topics than in GikiCLEF), it was not expected that all participants would answer all questions. Therefore, aiming for a higher coverage of answered topics, the interface presented to each participant the topics in a different order.

Still, the participants had the option of altering the order in which they navigated in two ways, namely:

- Direct navigation to the particular topic they were interested in answering, when-viewing the list of all topics and the list of topics previously answered;
- Choice of the subject of the next topic.

The interface provides a keyword based search on a static version of the Wikipedia. Participants can use this functionality to find documents, which can then be selected either as answers to the current topic, or as justifications for a particular answer of the current topic. As for justifications, in addition to providing a list of justification documents, participants have the option of providing a textual description of how the list of documents constitutes a justification to the given answer, if they feel that just listing the documents still does not make the justification obvious.

The result of a search is an ordered list of documents whose titles match the keywords provided by the participants. Each document can be visualized, allowing the participant to decide whether it is an appropriate answer or justification. If this is the case, it should be simple to add it as answer or justification using the buttons on top of the page being visualized.

The current system logs the participants' actions in the background: visited topics, keyword searches, documents viewed and the answers and justifications given. This allows the study of the time used for each topic and answer, as well as investigate the strategies used by the participants to find answers and justifications.

Support for Assessment. SIGA provides extensive support in the assessment phase of IR evaluations. The first task consists in pooling all answers returned by the participants. Answers and justifications provided by automatic and human participants are pooled together and treated almost the same way in the assessment process. The only difference is that the human participants have the possibility of providing a textual explanation on how the justification documents support the answers, which is displayed to the assessors in the assessment interface.

The assessment interface of SIGA allows assessors are able to judge the candidate answers, and check the correctness of their justifications. Prior to this, there is an automatic process where answer and justification documents are marked as correct if they had already been listed as answers and justifications by the topic manager during the topic preparation period.

Assessment using SIGA is thus performed in three steps:

1. All answers and justifications provided by the participants which were listed as answers and justifications during the topic preparation period are automatically classified as correct.
2. The remaining answers are then assessed by the assessors, and classified either as Incorrect, Correct, or Dubious. If the answer is classified as Correct, assessors

must also indicate whether it is Justified or Not Justified (by looking at the answer document and possibly at the chain of justification documents).

3. Finally, for answers for which different assessments exist, conflict resolution is performed. This process allows the assessors to discuss and become aware of complications and/or mistakes or mistaken assumptions. The administrator can choose to question the diverging assessors, or decide in straightforward cases.

Scoring, and Display of Comparative Results. The final scores are automatically computed after the completion of the assessment task and made available to participants, who have access to several different measures (precision, pseudo-recall, originality, tolerant-precision, etc.[4]) and to the detailed assessment of their answers, which they can contest or comment upon. (See [9] for an overview of the assessment problems.)

3 Towards Better Portuguese IR Systems with SIGA

SIGA was used in Págico where both automatic and human participants competed to find (justified) answers to 150 topics in the Portuguese Wikipedia.

Altogether the (6 human and 2 automatic) participants in Págico submitted a total of 32,488 unique answers, see Tables 1 and 2, which still reflect preliminary numbers.

Table 1. Statistics about answers and justifications

	Topic owners	Participants
Auto-justified answers	635	31,714
Answers that include justification	72	774
(with one document)	(67)	(678)
(with more than one document)	(5)	(96)
Total	707	32,488

This version of SIGA allowed us to significantly enlarge the set of topic answers and justifications, given that human participants provide more reliable answers than automatic systems[5], hence resulting in a better evaluation contest.

The human participation brought in the challenging topic of non-topical factors in information access (see [10]). With this amount of data we have produced a large base of information for subsequent statistical processing of data in Portuguese, as opposed to the case of GikiCLEF, where most of the answers were in English, and very few were only found in Portuguese.

[4] Pseudo-recall is computed by considering the set of all correct answers jointly returned or prestored, as if they were all correct answers in Wikipedia. Originality measures the capacity of finding answers that others have not found. Tolerant-precision is more lenient in that it also accepts correct but not justified answers.

[5] Table 3 shows that human precision is always above .5, while automatic participation was not higher than .1173.

Table 2. Answers per participant: AA=assessed automatically; HA=assessed by humans; C+J=correct and justified; C+nJ=correct but not justified; nC=incorrect

Participant	Sent	AA C+J	AA C+nJ	AA C+nJ	AA nC	AA nC
				but HA C+J	but HA C+J	but HA C+nJ
ludIT	1387	263	135	18	683	22
GLNISTT	1016	180	64	1	419	52
Ângela Mota	157	46	0	0	42	3
João Miranda	101	25	26	4	29	3
Bruno Nascimento	34	19	1	0	4	2
RAPPORTAGICO	5184	226	7	0	355	38
RENOIR	45000	305	12	0	856	81

Table 3. Participant scores: score=C*C/N, where C=correct answers, N=number of answers given by participant

Participant	Run	Topics	Answers	Correct	Precision	Pseudo-recall	Score
ludIT	1	150	1387	1067	0.7693	0.5105	820.8284
GLNISTT	1	148	1016	660	0.6496	0.3158	428.7402
João Miranda	1	40	101	80	0.7921	0.0383	63.3663
Ângela Mota	1	50	157	88	0.5605	0.0421	49.3248
RAPPORTAGICO	3	116	1730	207	0.1197	0.0990	24.7682
RAPPORTAGICO	2	115	1736	202	0.1164	0.0967	23.5046
RAPPORTAGICO	1	114	1718	180	0.1048	0.0861	18.8591
Bruno Nascimento	1	18	34	24	0.7059	0.0115	16.9412
RENOIR	1	150	15000	437	0.0291	0.2091	12.7313
RENOIR	3	150	15000	399	0.0266	0.1909	10.6134
RENOIR	2	150	15000	330	0.0220	0.1579	7.2600

Table 4. Analysis of correct and justified answers: SHA=answers given both by systems and humans; HA=answers given only by humans; SA=answers given only by systems

	SHA	HA	SA	Total
Assessed automatically	133	254	49	436
Assessed by judges	195	886	302	1383
Total	328	1140	351	1819

Furthermore, by logging the human participants navigation through the topics while they are trying to find the answers and justifications, we can provide important information for IR system developers and interface designers that can help them to identify some of the strategies used by people when seeking information in Wikipedia, as the following figures illustrate.

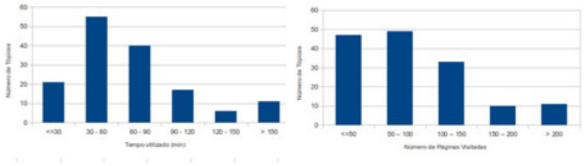

Fig. 1. A bird eye’s view of human participation in Págico

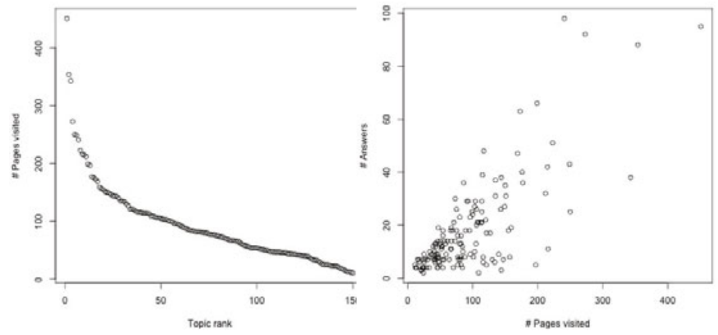

Fig. 2. Interplay between topics and number of different pages scrutinized

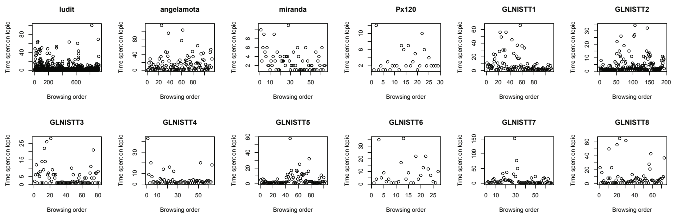

Fig. 3. Comparison between time spent on answers and browsing order

Acknowledgments. Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), UMIC, FCCN and FCT. Págico is also supported by the Universities of Oslo, PUC-Rio, and Coimbra. We are most grateful to the rest of the Págico team and to the Págico participants.

References

1. Santos, D., Cabral, L.M.: Summing GikiCLEF up: expectations and lessons learned. In: Peters, C., Nunzio, G.D., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G., Borri, F., Nardi, A., Peters, C. (eds.) Multilingual Information Access Evaluation. Text Retrieval Experiments, vol. I, pp. 212–222. Springer (2009)
2. Santos, D., Cabral, L.M., Forascu, C., Forner, P., Gey, F., Lamm, K., Mandl, T., Osenova, P., Peñas, A., Rodrigo, Á., Schulz, J., Skalban, Y., Sang, E.T.K.: GikiCLEF: Crosscultural Issues in Multilingual Information Access. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, European Language Resources Association, ELRA (May 2010)
3. 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, pp. 424–424. ACM, New York (2008)
4. Lalmas, M., Piwowarski, B.: INEX 2006 relevance assessment guide. In: INEX 2006 Workshop Pre-Proceedings, pp. 389–395 (2006)
5. Santos, D.: Porquê o Págico? Linguamática 4 (2012)
6. Santos, D.: Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. Linguamática 1(1), 25–59 (2009)
7. Gonçalo Oliveira, H., Cardoso, N.: SAHARA: an online service for HAREM Named Entity Recognition Evaluation. In: The 7th Brazilian Symposium in Information and Human Language Technology, STIL 2009 (2009)
8. Mota, C., Santos, D. (eds.): Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca (2008)
9. Mota, C., Freitas, C., Costa, L., Rocha, P.: O que é uma resposta? notas de uns avaliadores estafados. Linguamática 4 (2012)
10. Karlgren, J.: Stylistic Experiments for Information Retrieval. PhD thesis, Stockholm University (2000)