Págico: Evaluating Wikipedia-based information retrieval in Portuguese

Cristina Mota, Alberto Simões, Cláudia Freitas, Luís Costa, Diana Santos

Linguateca/FCCN, CEHUM/UM, PUC-Rio, Linguateca/FCCN, University of Oslo
cmota@ist.utl.pt, ambs@ilch.uminho.pt, claudiafreitas@puc-rio.br, luis.f.kosta@gmail.com, d.s.m.santos@ilos.uio.no

Abstract

How do people behave in their everyday information seeking tasks, which often involve Wikipedia? Are there systems which can help them, or do a similar job? In this paper we describe Págico, an evaluation contest with the main purpose of fostering research in these topics. We describe its motivation, the collection of documents created, the evaluation setup, the topics chosen and their choice, the participation, as well as the measures used for evaluation and the gathered resources. The task—between information retrieval and question answering—can be further described as answering questions related to Portuguese-speaking culture in the Portuguese Wikipedia, in a number of different themes and geographic and temporal angles. This initiative allowed us to create interesting datasets and perform some assessment of Wikipedia, while also improving a public-domain open-source system for further wikipedia-based evaluations. In the paper, we provide examples of questions, we report the results obtained by the participants, and provide some discussion on complex issues.

Keywords: Information Retrieval, Question Answering, Portuguese, Evaluation, Wikipedia.

1. Introduction

This paper presents Págico, an evaluation contest organized by Linguateca, whose main goal is to foster the development of information retrieval systems that find non-trivial answers to complex information needs (topics) in Portuguese Wikipedia.

Given a set of topics, the task consists in finding the Wikipedia pages that are answers to those topics. If the answer page is not enough to justify that it is indeed the correct answer, participants must provide additional justification, which consists in a set of other Wikipedia pages that, combined, completely justify the answer. The pages provided as answers and justifications by the participants were selected from a static version of the Wikipedia created by Linguateca for Págico, as described in section 5.

The following illustrates examples of topics:

- *Which Portuguese-speaking football players played professionally in more than three different countries?*
- *Which other fighters for African independence of previous Portuguese colonies worked with Amílcar Cabral?*

Whereas the Wikipedia page about *Bebeto* is sufficient to show that *Bebeto* is a correct answer to the first topic, the page about *Agostinho Neto* does not contain information about a relationship with *Amílcar Cabral*, although it shows that *Agostinho Neto* as student started to fight for the independence of Angola. So, in this case, it is also necessary to provide the page on *Amílcar Cabral* as a justification to the *Agostinho Neto* answer.

Although Págico is a follow up of GikiCLEF (Santos et al., 2010) that builds on our previous experience, it differs in the following:

- participants need only search the Portuguese Wikipedia;
- the topics focus on a specific cultural sphere (the Portuguese-speaking one) instead of promoting cross-linguality or geographical themes;
- human participation was encouraged in addition to that of automatic systems.

Págico was announced in mid June 2011, and registration was open for systems until the end of July. Human participants could register until the beginning of the evaluation, which happened on November 4th. Systems were allowed to submit a total of three runs, and had one week to submit their results, while human participants could go on providing answers until the end of November.

The evaluation results were delivered in the beginning of 2012, so participants could write about their systems and approaches before the final meeting, a satellite workshop of PROPOR[1], the main conference on natural language processing in Portuguese, in mid April 2012.

Although 21 teams registered for Págico (6 systems and 15 human participants), only one third actually participated providing answers to the topics (2 systems, and 5 human teams, of which: 3 were individuals, one had 6 members, and another comprised 23 people organized into 8 groups). Table 1 shows the participation in numbers.

The participants were evaluated using the GikiCLEF metrics (precision and final score), and also using the following new measures: pseudo-recall, pseudo-F-measure, originality and creativity, as described in section 6.

All measures were calculated per run. Additionally we calculated originality and creativity considering different runs of the same system as a single participation, so that a system would not compete with itself, and was possible to assess the system's creativity and originality. Results per run are presented in Table 2.

In what follows, we motivate this task on the next section, and we describe the topics involved in section 3. The paper also describes SIGA, the GikiCLEF topic management and assessment system, that was adapted to allow human participants to answer the topics and justify the answers, in section 4. The collection's construction is described in section 5. Other improvements regarding the scoring are also

[1] International Conference on Computational Processing of the Portuguese Language, see `http://www.propor2012.org/`

Participation type	Team (Run)	# Answers	# With justification
Human	Ângela Mota	157	8 (5%)
	GLNISTT	1016	255 (25%)
	ludIT	1387	489 (35%)
	João Miranda	101	60 (50%)
	Bruno Nascimento	34	1 (3%)
	Total	**2695**	
	Distinct	**2383**	
Automatic	RENOIR (1)	15000	
	RENOIR (2)	15000	
	RENOIR (3)	15000	
	Total	**45000**	
	Distinct	**28626**	
	RAPPORTAGICO (1)	1718	
	RAPPORTAGICO (2)	1736	
	RAPPORTAGICO (3)	1730	
	Total	**5184**	
	Distinct	**2343**	
	Total	**50184**	
	Distinct	**30543**	
Total		**52879**	
Distinct		**32485**	

Table 1: Participation in Págico.

Team (Run)	$\|T\|$	$\|R\|$	$\|R\|/\|T\|$	$\|C\|$	$\|\tilde{C}\|$	S	P	ρ	ϕ	$\tilde{P}$	O	K
ludIT	150	1387	9.25	1065	34	817.75	0.768	0.474	0.586	0.792	3442	3995.21
GLNISTT	148	1016	6.86	661	52	430.04	0.651	0.294	0.405	0.702	1767	2211.83
João Miranda	40	101	2.52	80	3	63.37	0.792	0.036	0.068	0.822	202	287.14
Ângela Mota	50	157	3.14	88	3	49.32	0.56	0.039	0.073	0.58	146	251.4
RAPPORTAGICO (3)	114	1730	15.18	208	13	25.01	0.12	0.092	0.104	0.128	29	297.00
RAPPORTAGICO (2)	115	1736	15.1	203	13	23.74	0.117	0.09	0.102	0.124	5	265.22
RAPPORTAGICO (1)	116	1718	14.81	181	11	19.07	0.105	0.08	0.091	0.112	22	224.72
Bruno Nascimento	18	34	1.89	23	1	15.55	0.676	0.01	0.02	0.706	37	65.67
RENOIR (1)	150	15000	100	436	38	12.67	0.029	0.194	0.051	0.032	126	745.09
RENOIR (3)	150	15000	100	398	29	10.56	0.026	0.177	0.046	0.028	54	618.50
RENOIR (2)	150	15000	100	329	25	7.22	0.022	0.146	0.038	0.024	220	609.23

Table 2: Págico results.

reported in section 6.

2. Motivation

There were four different motivation aspects for organizing Págico:

1. The first and most obvious is to continue the already long tradition of Linguateca's evaluation contests for Portuguese, with Morfolimpíadas (Costa et al., 2007), HAREM (Santos and Cardoso, 2007; Mota and Santos, 2008), IR, QA and GIR in the larger scope of CLEF, and GikiP (Santos et al., 2009) and GikiCLEF (Santos and Cabral, 2010) as precursors. The main goal is to help the development of NLP systems that deal with Portuguese and that can be evaluated in a near-realistic context;

2. The goal of looking at Portuguese Wikipedia is because this material is getting more and more one of the standard resources for NLP (as the current LREC even illustrates), in addition to having become one of the most visited and consulted knowledge resources for the person in the street. We believe that a rational and objective evaluation of Wikipedia is therefore an important goal and we hope that Págico can partially contribute to it, or at least to the start of a critical assessment of this invaluable resource.

3. The practical goal of developing frontends to Wikipedia is also important to explain: there are still a lot of questions, especially aggregate questions, that are hard to answer by Wikipedia, and for which automated systems should be of great help. In other words: we are not proposing evaluation of toy systems just to check if computers can be as good as humans.[2] Rather, we are interested in systems that really help humans in non trivial information finding goals.

[2] Although not a toy system, this was the purpose of Watson in Jeopardy (Ringel, 2011).

4. Finally, this is the first time that we implemented a side-track where also humans can participate, moving therefore to the challenging topic of non-topical factors in information access (see (Karlgren, 2000)). This is certainly a contribution to studying the different ways that people and machines solve a particular problem, but we also expect that the activity itself will provide:

   — for all the participants, learning on how Wikipedia is structured and how to navigate it;

   — for the human competitors, in particular, gaining knowledge on some of the 150 topics;

   — for the organizers, a pool of answers that will allow us to do a better evaluation of the systems (and the participants themselves) because some recall-oriented measures (and not only precision) will be possible.

   — for the whole Portuguese-speaking community, knowledge of Wikipedia's contents in Portuguese and also some insight on justification paths for different topics

More concrete pieces of motivation can also be read from the Págico website[3].

3. Topic Creation

Four people were involved in the creation of 150 topics related to the "Lusophone culture." Two criteria guided us in the creation process:

1. topics should be appealing and interesting, at least from our point of view, to the Portuguese-speaking community;

2. topic answers should not be obvious. That is, we chose topics whose justification, was, preferably, spread among different pages in Wikipedia.

As to the second point, a topic like "Musicians associated with the development of Bossa Nova", for instance, would be easily answered by browsing the Bossa Nova wiki page, and so it was discarded (though, from the systems point of view, to find the correct answer is far from being a simple task). The strategy was, therefore, to focus on questions deemed difficult from the standpoint of humans – a defensible argument when aiming at a "non-artificial" task.
In tables 3 and 4 we present the results of human and automatic participation by super-themes and by countries or locations the topics addressed.
Considering the overall spectrum of Lusophone culture underlying Págico, and also that a topic can be about more than one theme ("Which other fighters for African independence of previous Portuguese colonies worked with Amílcar Cabral," for example, can be as much about Politics as History), the 150 topics are distributed as displayed in Table 5.
Table 5 shows in bold the distribution of major themes - the broader categories, which we call super-themes. Most of the 150 topics belong to the super-theme Humanities (Letras), mainly due to the presence of the History theme, which comprehends topics associated with encyclopedic knowledge. The second most frequent category is Arts (Artes), which includes music and cinema, among others. The overall amount of topic classifications exceeds 150, since some topics were attributed to more than one theme. Table 5 also shows topic distribution by theme. History was the most frequent theme (50 topics), followed by Geography (26 topics), Music, Politics (19 topics each) and Sports (18 topics).

Super-theme	Final score		Precision	
	Hum.	Auto.	Hum.	Auto.
Letras	590.72	5.24	71.52	1.90
Artes	324.80	4.48	71.07	2.46
Geografia	268.88	8.86	71.70	3.62
Cultura	205.34	2.19	67.11	2.05
Política	107.58	0.77	65.60	1.39
Desporto	104.31	1.14	63.22	1.75
Ciência	59.08	1.88	61.54	2.57
Economia	45.10	0.32	71.59	1.61

Table 3: Comparison between human participation and systems per super-theme.

Location	Final score		Precision	
	Hum.	Auto.	Hum.	Auto.
Brasil	462.28	9.73	72.69	3.08
Lusofonia	275.89	1.47	61.86	1.22
Portugal	202.75	2.50	73.73	2.75
Geral	64.46	0.10	65.77	0.87
Moçambique	36.91	0.29	68.35	1.22
Angola	36.05	3.87	69.33	5.23
Macau	23.44	0.42	75.61	2.44
Cabo Verde	19.88	0.19	76.47	1.38
Timor	13.83	0.83	62.86	4.17
Guiné Bissau	5.44	0.00	77.78	0.39
São Tomé e Príncipe	4.45	0.03	63.64	1.14

Table 4: Comparison between human participation and systems per country.

4. SIGA and the new functionalities required by Págico

SIGA[4], the system used to support the organization and participation in Págico, was developed and designed in the context of GikiCLEF. The need for this computational environment arose from the considerable number of people creating and assessing topics, dealing with large amounts of data (the collections and systems' submissions).
SIGA supports different actions for distinct roles: manager, topic developer (creator or other), participant, assessor (basic or conflict resolver), and simple observer. SIGA takes also care of several procedures, such as validation of runs, pool creation, assessment distribution, conflict detection, score computation, and display of comparative results. For more details about SIGA's architecture and motivation

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

[4] http://dinis.linguateca.pt/avalconjunta/Pagico/SIGA/

Super-theme -theme	Topics		Answers received		Correct					
					Participants		Organizers		All	
	#	%	#	%	#	%	#	%	#	%
Letras	**69**	**46.00**	**15506**	**47.73**	**961**	**48.05**	**328**	**49.03**	**1085**	**48.24**
- história	50	33.33	11124	34.24	664	33.20	243	36.32	761	33.84
- literatura	15	10.00	3610	11.11	202	10.10	74	11.06	236	10.49
- linguística	6	4.00	1312	4.04	64	3.20	32	4.78	72	3.20
- jornalismo	3	2.00	667	2.05	27	1.35	9	1.35	28	1.24
- filosofia	2	1.33	425	1.31	53	2.65	4	0.60	54	2.40
Artes	**36**	**24.00**	**7910**	**24.35**	**542**	**27.10**	**204**	**30.49**	**609**	**27.08**
- música	19	12.67	4075	12.54	194	9.70	79	11.81	222	9.87
- cinema	10	6.67	2243	6.90	216	10.80	78	11.66	237	10.54
- televisão	4	2.67	967	2.98	57	2.85	31	4.63	70	3.11
- artes plásticas	2	1.33	431	1.33	21	1.05	13	1.94	24	1.07
- artes	2	1.33	450	1.39	66	3.30	5	0.75	68	3.02
Geografia	**34**	**22.67**	**7152**	**22.02**	**509**	**25.45**	**151**	**22.57**	**563**	**25.03**
- geografia	26	17.33	5476	16.86	396	19.80	112	16.74	431	19.16
- arquitetura/urbanismo	7	4.67	1481	4.56	57	2.85	22	3.29	66	2.93
- demografia	4	2.67	894	2.75	207	10.35	36	5.38	221	9.83
- geologia	2	1.33	492	1.51	18	0.90	1	0.15	18	0.80
Cultura	**27**	**18.00**	**5612**	**17.28**	**370**	**18.50**	**79**	**11.81**	**395**	**17.56**
- antropologia/folclore	12	8.00	2544	7.83	160	8.00	40	5.98	172	7.65
- religião	7	4.67	1252	3.85	119	5.95	20	2.99	128	5.69
- culinária	5	3.33	1008	3.10	57	2.85	15	2.24	63	2.80
- cultura	3	2.00	748	2.30	75	3.75	9	1.35	75	3.33
- ensino	2	1.33	402	1.24	11	0.55	5	0.75	11	0.49
Política	**19**	**12.67**	**4164**	**12.82**	**197**	**9.85**	**59**	**8.82**	**220**	**9.78**
Desporto/Esportes	**18**	**12.00**	**3914**	**12.05**	**188**	**9.40**	**79**	**11.81**	**225**	**10.00**
Ciência	**14**	**9.33**	**2973**	**9.15**	**155**	**7.75**	**45**	**6.73**	**172**	**7.65**
- saúde	4	2.67	885	2.72	21	1.05	18	2.69	30	1.33
- zoologia	3	2.00	527	1.62	83	4.15	9	1.35	83	3.69
- ciência	2	1.33	446	1.37	20	1.00	10	1.49	27	1.20
- botânica	2	1.33	384	1.18	9	0.45	6	0.90	10	0.44
- geologia	2	1.33	492	1.51	18	0.90	1	0.15	18	0.80
- matemática	1	0.67	239	0.74	4	0.20	1	0.15	4	0.18
Economia	**6**	**4.00**	**1321**	**4.07**	**77**	**3.85**	**30**	**4.48**	**94**	**4.18**

Table 5: Detailed classification of Págico super-themes.

please check (Santos et al., 2010; Santos and Cabral, 2010; Santos and Cabral, 2009).

However, Págico also involved human participation, and we had to extend SIGA with new features and a new dedicated interface for these participants.

Since there was a considerable number of topics (three times more topics than in GikiCLEF), we were not expecting every participant to answer every question. Therefore, aiming at a higher coverage of answered topics, we opted to present the topics in a different order to each participant. The order in which the participants navigated the topics was thus determined so that altogether a most uniform coverage was achieved. However, the participants could alter the order in which they navigated in two different ways:

1. navigate directly to the particular topic they want to answer, through the list of all topics and the list of topics previously answered;

2. choose the overall (super-)theme of the next topic.

The main concepts in the interface are the topics, for which the participants are supposed to select Wikipedia documents as answers and/or justifications.

The interface provides a keyword based search on a static version of the Wikipedia (described in section 5. below), so that participants can find documents that can be assigned either as answers to the current topic, or as justifications for a particular answer on the current topic. Regarding the justifications, besides providing a list of justification documents, participants could (in the cases where just listing the documents is not an obvious justification) provide a textual description about how the list of documents constitutes a justification to the given answer.

The participants could also navigate from documents found to other documents until they come to the correct answer or appropriate justification, and select those directly as answers or justifications.

In the background the system logs all keyword searches and documents viewed by the participants, to enable the later study of the strategies used to find answers to the topics, and justifications to answers when necessary; see (Costa et

al., 2012) for a preliminary analysis of human navigation.
We also improved the interface for topic management, extending it with the addition of the required justifications, allowing a more effective and complete automatic assessment of the participants answers and justifications.

5. Collection

The collection used in Págico was created from a Wikipedia snapshot of April 25th, 2011, and was made available both as a single zip file, and as a collection of one hundred smaller zip files of XHTML documents.

5.1. Collection Construction

Wikipedia mark-up syntax is very rich and powerful. It can be used as a standard wiki mark-up language but it also supports complex macros. These features are of great relevance, making Wikipedia easier to maintain, but make the language harder to parse.
Although Wikimedia source code is available, the formatting behavior is not available as a stand alone method. Therefore, when processing Wikipedia, researchers have to write their own parsers, use some of the (incomplete) parsers available in the Web, or just hit the Wikipedia web page with crawlers.
In order to both making this parsing task easier and the evaluation task simpler, we decided to pre-process the Portuguese version of Wikipedia, converting it to XHTML, normalizing its names and, equally importantly, defining an official collection to be used by all participants.
As mentioned, robust tools to convert Wikimedia syntax to XHTML are lacking. After some analysis, we chose the `mwlib`[5] Python library to perform the format conversion. To make the Wikipedia Portuguese snapshot easier to process, we used the `MediaWiki::DumpFile`[6] Perl module.
When using the `mwlib` software we found yet another problem, concerning localization: English Wikipedia *Template* namespace was renamed to *Pré-Definição*, some of the redirection pages use *redirecção* instead of *redirection*, and so on. Although `mwlib` code is mostly parameterized by a language code, we found that this feature intends to make future versions language-aware, but is not yet fully functional. This led to some in-house software development to process Wikipedia XML snapshot and create a cache of macro templates, replacing some of them in the XML file. This replacement was however not possible for all macros because dealing with some of them would damage the XML well-formedness of the document.
Some of these macros were relevant but were lost for this Págico edition. An example of such macro is the well known *Infobox* table.
The collection was further processed in order to convert all links from external to internal so that it could be easily navigated, also performing some corrections to the mark-up generated by `mwlib`.
Redirection pages were detected, and replaced by an empty HTML page just with the link to the redirection target.
Finally, the obtained pages were organized in folders, accordingly with their title first letters, and filenames were normalized.

[5] `http://pediapress.com/code/`
[6] `http://search.cpan.org/dist/MediaWiki-DumpFile/`

5.2. Collection Characterization

Wikipedia contains different kind of pages, not all of them relevant for providing answers in Págico, such as the template pages mentioned above, the disambiguation pages, the redirection pages and the media pages. Only article pages are relevant for Págico, but the collection delivered to the participants included all these kinds of documents, as summarized in Table 6.

Page type	# Documents
Templates	32 900
Disambiguation	5 006
Redirection	574 077
Media	9 678
Articles	856 005

Table 6: Document distribution per page type.

In addition to the type of the Wikipedia pages, we also looked at the categories assigned to the pages, which are available as part of the pages markup. They can be considered as tags of a folksonomy (Sinclair and Cardew-Hall, 2008). We found a rather anarchic distribution in the Portuguese Wikipedia, which contains 95 446 categories to classify 681 058 documents: there are more than 8 500 documents not yet classified), and, as can be seen in Table 7, a large number of categories (32 652) classify only one document, and the vast majority of categories (59 775) classify less than 66 documents.

# Documents	# Categories	Percent
$]0, 1]$	32 652	34.21%
$]1, 66]$	59 775	62.63%
$]66, 130]$	1 789	1.87%
$]130, 194]$	507	0.53%
$]194, 260]$	231	0.24%
$]260, 345]$	166	0.17%
$]345, 442]$	108	0.11%
$]442, 592]$	84	0.09%
$]592, 862]$	68	0.07%
$]862, \infty[$	65	0.07%

Table 7: Number of documents per category number.

Conversely, Table 8 synthesizes the distribution of documents according to the number of categories that classify them, showing 8 771 documents with no categories, and that the majority of documents (676 705) has between 1 and 8 associated categories.

5.3. Answers in Págico

Focusing now on the subset of the collection constituted by the answer and justification documents amassed by Págico (both those provided by topic creators those submitted by participants), Figure 1 provides the distribution of the number of answer documents per topic. For exactly half of

# Categories	# Documents	Percent
0	8 771	1.271%
$]0, 8]$	676 705	98.097%
$]8, 15]$	4 008	0.581%
$]15, 23]$	314	0.046%
$]23, 33]$	25	0.004%
$]33, \infty[$	6	0.001%

Table 8: Number of categories per document.

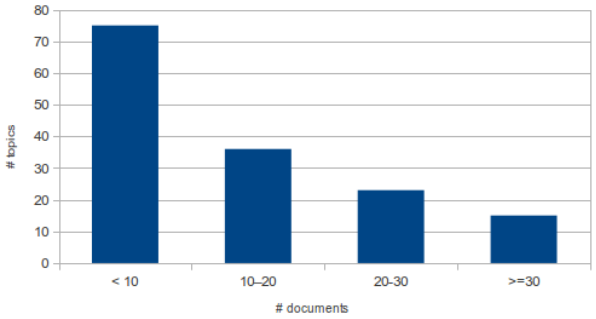

Figure 1: Topics according to the average number of answer documents found.

the topics this amounts to less than ten documents and for more than two thirds of the topics this number is lower than twenty answers.

The number of words in the answer documents varies substantially, as illustrated in Figure 2. Some topics have less than one thousand words, whereas others have more than one hundred thousand words.

Figure 3 illustrates the number of categories in which the answer documents are classified. For most of the topics this number is not higher than four categories.

Table 9 show the topics with highest and lowest number of answer documents. The latter category corresponds often to African themes, which may indicate that the Portuguese Wikipedia lacks information about such themes. On the

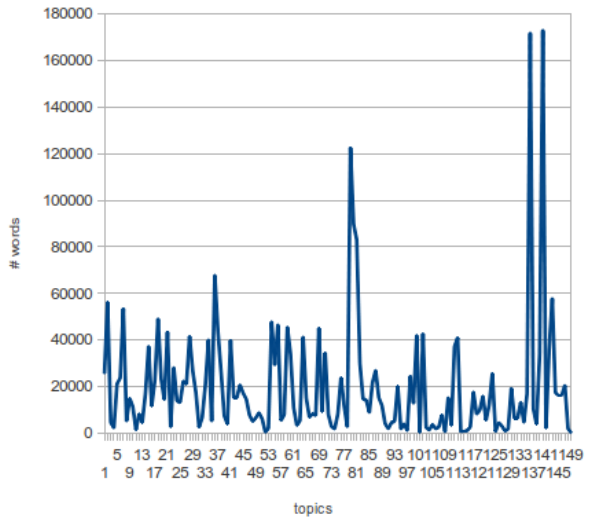

Figure 2: Number of words of the answer documents per topic.

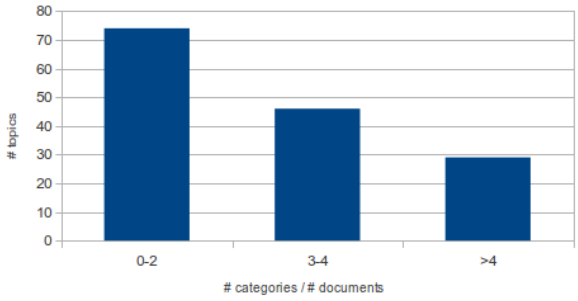

Figure 3: Topics according to the number of categories assigned to their answers.

other hand, the topics with many correct answers seem to have an intrinsic high number of answers such as 'Indigenous tribes living in the Amazon Rainforest' and 'Museums in capitals of Lusophone countries'.

Topic	#
Indigenous tribes living in the Amazon Rainforest	95
Museums in capitals of Lusophone countries	62
Locations mentioned in "Os Lusíadas"	51
Indigenous Brazilian peoples considered extinct.	50
Viceroys of the Portuguese India	48
...	
Politicians from Portuguese speaking African countries who studied in the Soviet Union	2
Churches in Rio de Janeiro constructed by Afro-Brazilian confraternities.	1
Members of Parliament from FRELIMO	1
Mozambican writers who received Prémio Camões	1
Foreign writers who visited Portugal in the 19th Century and published descriptions of their travels	1

Table 9: Topics with most and least answer documents.

6. Evaluation measures

The participants were evaluated using the GikiCLEF measures (precision and final score), and also using the following new ones: pseudo-recall, pseudo-F-measure, originality and creativity.

6.1. Precision and Tolerant Precision

$$P_{p,r} = \frac{|C_{p,r}|}{|R_{p,r}|} \quad (1)$$

$$\tilde{P}_{p,r} = \frac{|C_{p,r}| + |\tilde{C}_{p,r}|}{|R_{p,r}|} \quad (2)$$

We computed the (normal) precision, $P_{p,r}$, given by the number of correct and justified answers, $|C_{p,r}|$, over the total number of answers $|R_{p,r}|$ provided by run r of participant p (see equation 1). In addition, we defined the tolerant precision $\tilde{P}_{p,r}$, that does not take into account whether the correct answers are also correctly justified (see equation 2, where $|\tilde{C}_{p,r}|$ is the number of answers that are correct but not justified).

6.2. Pseudo-recall

$$\rho_{p,r} = \frac{|C_{p,r}|}{|C_{Pagico}| + |C_{aval}|} \quad (3)$$

Although topic creators provided many answers and justifications, the set is not complete, and hence it is not possible to calculate a (true) recall measure. Nonetheless, we defined a pseudo-recall, $\rho_{p,r}$ (see equation 3), that uses as reference key not only the answers and justifications provided by the topic creators $|C_{Pagico}|$[7], but also the participant answers that do not belong to C_{Pagico} and were considered correct and justified by the evaluators $|C_{aval}|$.

6.3. Pseudo-F-measure

$$\phi_{p,r} = 2 \times \frac{P_{p,r} \times \rho_{p,r}}{P_{p,r} + \rho_{p,r}} \quad (4)$$

Given that we defined pseudo-recall, we also calculated a pseudo-F-measure, $\phi_{p,r}$, given by equation 4, that combines precision and pseudo-recall in a single value.

6.4. Originality and Creativity

We defined two different metrics to reward the uniqueness of the answers in a run:

- Originality, $O_{p,r}$, which measures the number of original correct answers given by the run r of participant p, i.e., the number of correct answers that no other participant or topic creator came up with (see equation 5). The more participants $p(i)$ tried to answer the same topic i, the more original is the answer (as shown by equation 6, that computes the answer originality $o(r_{p,r,i,j})$).
- Creativity, $K_{p,r}$, which measures how creative are the answers (see equation 7), in the sense that a answer may not be original but be more or less creative if there are less or more participants answering the respective topic. Hence, the answer creativity $k_{p,r,i,j}$, formulated by equation 8, is inversely proportional to the number of participants that gave the same answer, $c(r_{p,r,i,j})$, and directly proportional to the number of participants that tried to reply to the topic in question $p(i)$.

It should be noted that if all correct answers are original, then $O_{p,r} = K_{p,r}$.

$$O_{p,r} = \sum_{i}^{T} \sum_{j}^{R_{p,r,i}} o(r_{p,r,i,j}) \quad (5)$$

$$o(r_{p,r,i,j}) = \begin{cases} p(i) & r_{p,r,i,j} \in C_{aval} \quad \wedge \\ & r_{p,r,i,j} \notin C_{Pagico} \quad \wedge \\ & r_{p,r,i,j} \notin \bigcup_{m \neq p, n \neq c} R_{m,n} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

[7] In some cases, the topic creators provided correct bt not fully justified answers. These were not taken into account when computing pseudo-recall.

$$K_{p,r} = \sum_{i}^{T} \sum_{j}^{R_{p,r,i}} k(r_{p,r,i,j}) \quad (7)$$

$$k(r_{p,r,i,j}) = \begin{cases} \frac{1}{c(r_{p,r,i,j})} \times p(i) & r_{p,r,i,j} \in \\ & C_{Pagico} \bigcup C_{aval} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$O_p = \sum_{i}^{T} \sum_{j}^{R_{p,i}} o(r_{p,i,j}) \quad (9)$$

$$o(r_{p,i,j}) = \begin{cases} p(i) & r_{p,i,j} \in C_{aval} \quad \wedge \\ & r_{p,i,j} \notin C_{Pagico} \quad \wedge \\ & r_{p,i,j} \notin \bigcup_{m \neq p} R_m \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

$$K_p = \sum_{i}^{T} \sum_{j}^{R_{p,i}} k(r_{p,i,j}) \quad (11)$$

$$k(r_{p,i,j}) = \begin{cases} \frac{1}{c(r_{p,i,j})} \times p(i) & r_{p,i,j} \in C_{Pagico} \bigcup C_{aval} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$p(i)$ = # participants in topic i
$c(r_{p,r,i,j})$ = # participants that gave answer $r_{p,r,i,j}$

Another aspect worth pointing out is that both originality and creativity are proportional to the number of participants that tried to answer the topic, instead of being proportional to the number of different runs that tried to answer the topic. Otherwise, systems would be penalized because they sent more than one run, and it is likely that different runs share a sizeable part of the answers (in Págico, RAPPORTAGICO answers have an average frequency of 2.2 times whereas RENOIR answers are repeated 1.6 times).
Nonetheless, correct answers that exist only in different runs of the same system do not contribute to the run originality, and the creativity of answers that exist in different runs of the same system is also lower than the creativity of answers that appear in only one of the runs of a system (something that penalizes not only the run creativity but also the run creativity of the participants who gave the same answer). Consequently, we decided to also compute the participant's originality, O_p, and creativity, K_p (in addition to the run originality $O_{p,r}$ and creativity $K_{p,r}$), considering different runs of the same system as a single run.

6.5. Final Score in Págico

$$M_{p,j} = |C_{p,r}| \times P_{r,j} \quad (13)$$

Although we defined various metrics to evaluate the participants through different perspectives, the final score in Págico is the same as the final score per language used in GikiCLEF (Santos and Cabral, 2010), herein named $S_{p,r}$ (see equation 13). This measure, based on precision, allows distinguishing participants who have the same number of

correct answers but different total number of answers, and assigns a higher value to runs that contain more correct answers.

7. Concluding Remarks

Págico was an innovative evaluation contest, but not without problems.

On the one hand, we were not able to gather enough participation to be able to generalize – something that speaks against dedicating an evaluation exercise to a single language. Also, we were not able to develop, in a short time, an environment in which people were happy with to browse Wikipedia: most human participants reported using the ordinary Wikipedia and just went to the Págico system to register the answers. So, one of our goals, namely to study human problem solving with the help of the logs, could not be achieved.

On the other hand, we can report some progress in those areas, in the sense that a public, open source, system, SIGA, now exists in a much better condition than before. And we were able to create an interesting evaluation collection (CARTOLA[8]) for further research and development in the area of information access, something which is crucial for further work in these matters.

We were also able to produce some evaluation of the Portuguese Wikipedia, which to our knowledge is one of the first, based on cultural aspects, as well as detect a set of technical problems that have to be solved for non-English versions.

While we were not able to do justice, in this paper, to everything that was learned in this exercise, we have also managed to extensively document most of the issues, problems and solutions in the special edition of the Linguamática journal of April 2012 (Santos et al., 2012).

8. Acknowledgments

We are most grateful to the Págico participants, for without them we would not have gathered any resources worth making public.

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, Coimbra and FCT grant SFRH/BPD/73011/2010.

9. References

Luís Costa, Paulo Rocha, and Diana Santos. 2007. Organização e resultados morfolímpicos. In Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, pages 15–33. IST Press, Lisboa, Portugal, 20 de Março.

Luíss Costa, Cristina Mota, and Diana Santos. 2012. SIGA, a Management System to Support the Organization of Information Retrieval Evaluations. In *Proceedings of PROPOR'2012*, pages 284–290. Springer.

Jussi Karlgren. 2000. *Stylistic Experiments for Information Retrieval*. Ph.D. thesis, Stockholm University, Department of Linguistics, Stockholm.

Cristina Mota and Diana Santos, editors. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, December, 31.

Matrin Ringel. 2011. IBM Watson and Jeopardy! Presentation at Notur conference, oslo, June 2011.

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*.

Diana Santos and Luís Miguel Cabral. 2010. GikiCLEF : Expectations and lessons learned. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda, editors, *Multilingual Information Access Evaluation, VOL I*, number 1, pages 212–222. Springer, Setembro.

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

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 Pe nas, 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*. Springer.

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, and Erik Tjong Kim Sang. 2010. Gikiclef: Crosscultural issues in multilingual information access. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Diana Santos, Cristina Mota, Cláudia Freitas, and Luíss Costa, editors. 2012. *Linguamática 4(1)*. April.

James Sinclair and Michael Cardew-Hall. 2008. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, February.

[8] `http://www.linguateca.pt/Cartola/`