An initial proposal for cooperative evaluation on information retrieval in Portuguese

Rachel Aires[1,2], Sandra Aluísio[2], Paulo Quaresma[3], Diana Santos[1] and Mário J. Silva[4]

[1]SINTEF Telecom and Informatics, Box 124 Blindern,
N-0314 Oslo, Norway
{Rachel.V.Aires, Diana.Santos}@sintef.no
[2]ICMC-Universidade de São Paulo, NILC, C.P. 668
13560-970, São Carlos, SP, Brazil
{raires, sandra}@icmc.usp.br
[3]Departamento de Informática, Universidade de Évora
R. Romão Ramalho, 59, 7000 Évora, Portugal
pq@di.uevora.pt
[4]Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa,
Campo Grande 1749-016 Lisboa, Portugal
mjs@di.fc.ul.pt

Abstract. In this paper we discuss evaluation of information retrieval, Web search and question answering systems, paving the way for the organization of an evaluation contest on IR for Portuguese. Inspired by current international setups, we motivate the need to study the specific problems posed by Portuguese, suggesting a collection suitable for multiple tasks.

1 Introduction

Information retrieval (IR) may be broadly defined as the process of finding information satisfying a user's need [1], with the main focus of the IR community being in text retrieval, i.e. in finding specific documents in large text bases. IR research has also been extended in recent years to information available in other media, such as video and sound, and to information contained in (semi) structured forms, such as XTML and XML. An even broader definition of IR includes further processing of the information, such that more complex processes like summarization, explicit answer to a question (eventually requiring reasoning), information extraction or mining, and identification of the gist, are seen as different facets of the same discipline.

Motivation for assessing the status of IR (in a broad sense, as defined above) in Portuguese stems from several reasons. These range from the need of obtaining an indicator of "information society progress" for Portuguese-speaking countries and market oriented information providers, to a genuine research interest in issues connected with both IR and evaluation in general, and natural language processing in particular. We will not dwell in this paper on the first kind of (macro-economic) motivation. We shall rather discuss why the study and evaluation of IR in Portuguese can be of particular interest for both the development and improvement of already existing systems dealing with Portuguese, such as Web search engines or specific database query systems, and for the NLP community that deals with Portuguese and wishes to address real world problems.

For researchers of the first profile, it is advantageous to identify the challenges posed by IR in Portuguese, especially where an English-based architecture will be likely to miss the point. In addition, it is relevant to learn whether Portuguese-speaking users have different requirements and practice. For IR in the Web, it sounds misguided to assume that the distribution of subject matter and users' profiles mirrors that on the Web in general (see [2] for some comments). For a summarization task, it is also far from obvious that good abstracts in Portuguese should obey the same rules of their English counterparts, given the different rhetorical conventions (and practice) for the two languages. Also, the dialogue flow and implicit expectations are well known to differ, so one could, at least, hypothesize that the same would happen when interacting with a computer.

For the traditional NLP community, it is high time that specialized tasks and toy-like problems were replaced by real applications that deal with text and are employed by real users. This is the case of applications such as search engines or question answering systems. The aim would be not only to improve its self-confidence as an engineering discipline, but most of all to gain an opportunity to test systems and resources against real data. Information retrieval, unlike e.g. spelling correction, gives the opportunity to test semantic and pragmatic modules and theories of the language, and thus cannot be dismissed as theoretically uninteresting. Subjects like sieving terrorism information, describing commercial venues or automatically creating "who's who" directories have been approached with a simple pattern matching approach (as shown by the MUC conferences, see [3]), but they can also resort to complex models of events, fine-grained representations of time, and even non-monotonic inference; together with complex NLP techniques such as anaphor resolution, sense disambiguation, and thesaurus induction [4].

But if we want to do information retrieval in Portuguese, and evaluate its success, we have to begin by agreeing on some subset of well-defined task on whose solution we can agree on — and test the whole set of applications against some common base agreed upon. This is the evaluation contest model, which is thoroughly described elsewhere (see e.g [3], [5] and [6]). Basically, participants should agree on the conditions of participation (e.g. the restrict use of some data collection); the results of participant's systems are compared using the same tasks, the same data collection (the training data, the testing data and the answer key) and the same evaluation measures. However, this definition hides all the management work and costs involved to make the contest work from the first call for participation to the evaluation workshop. The latter serves both to present and discuss results from the experiments, including failure analyses and system comparisons (see for example, the comparative analysis of the results of CLEF 2002 campaign [7], of those of the Third NTCIR Web Retrieval Task [8], and the analysis of CLEF topics [9]), and to provide a fuller presentation of the systems, e.g. describing retrieval techniques used.

One obvious approach would be to join in the existing evaluation contests such as TREC or CLEF[1], thus avoiding the huge organization efforts for the contests. Even though we have not ruled out this possibility, we are under the impression that there is very little room in these contests for experimentation with language-specific problems, let alone devise collections that reflect specific problems. Furthermore, there are already efforts for evaluation of the whole field of Portuguese processing (such as the satellite Avalon'2003, the Workshop for Evaluation Campaigns for Portuguese). Therefore, an initiative for IR would work synergistically with those for other NLP topics in Portuguese. Another advantage of this approach is to count on a collection of texts and tasks specifically devised for Portuguese, and not a mere translation or adaptation of other language users' needs. If successful the evaluation initiatives would place researchers on Portuguese on an equal standing with organizers of international contests. This would allow contests involving Portuguese to be seamlessly incorporated into international contests, thus warranting both relevance to Portuguese and more professional organization.

2 Question answering evaluation

Question answering systems aim to retrieve answers rather than documents that may contain the answers. This is a more difficult goal and several constraints have been assumed in the evaluation contest models mentioned in the previous section. For instance, in TREC a special track on Question Answering (QA) has been created only in 1999. It has suffered some changes over the years and subdivided into different subtasks. As an example, the 2001 track on (QA) had the following configuration:

1. The main task. This task aims to retrieve answers to specific questions, such as, "Who discovered Azores?" or "When did Vasco da Gama arrive in India?". Answers are composed by a short text (<50 chars) and by a document identifier. The document should contain text supporting the answer string. If the collection of documents has no answer to the question, systems should answer 'NIL'. Systems were allowed to retrieve a ranked list of up to 5 answer-pairs to each query. In 2002, this task was changed to allow only exact answers, rather than a string containing the answer.

2. The list task. This task aims to obtain a set of instances as the answer for a question. These instances may need to be retrieved from multiple documents and should have no duplicates. Examples are questions like: "Name 9 EU countries" or "Name 5 Benfica football players". Answer instances may be duplicated in the collection and it is also guaranteed that the document collection has the answer to the question (it refers the adequate number of answer instances).

---

[1] It is worth mentioning that, in the interactive CLIR track of CLEF 2002, there is a paper reporting a Portuguese-English experiment. It used an MT system to translate part of the initial sample into Portuguese, due to the scarce number of parallel English-Portuguese public corpora available [10].

3. The context task. In this task the goal was to track discourse objects through series of questions. The systems should be able to manage interaction contexts and to track discourse objects, solving linguistic phenomena, such as anaphora and ellipsis. As example might be the following series of questions: 1. Who was the first king of Portugal? 2. When did he take the throne? 3. Where was he born?

In order to evaluate the performance of systems for these tasks answer pairs (answer-string, document-id) are required, which were provided by human referees. From experience in TREC conferences, it has been concluded that one referee is sufficient for each answer pair, as relative scores remain stable despite the differences in the refereeing policy. However, in order to have absolute (rather than relative) scores it is important to have more than one referee for each question and to combine them into a single judgment.

In previous QA evaluation contests, the typical approach to deal with this difficult problem has been to classify the questions according to a taxonomy, as follows: answers to questions starting with "who" should be a person or organization; for questions with "when" the answer should be a time entity; and for "where" a place should be given. After having classified the question, the systems commonly use information retrieval techniques and some shallow parsing techniques to obtain adequate entities and to answer the initial question.

Even though evaluation of IR systems for Portuguese is likely to have many features in common with that for English, different parsing strategies and named entity recognition requirements may turn specific systems into very different ones. In addition, we are aware of specificities of Portuguese interrogatives that require treatment, such as *O primeiro rei de Portugal tinha que nome? Em que igreja casou?*. In these cases, there are no clue words to specify the kind of information required, as demonstrated by comparing with *Em que partido votou?*

3 Web search evaluation

According to Saracevic [11], one can conceive three ways of evaluating IR, namely system evaluation; user evaluation; and evaluating the system from a user point of view. We give a short overview of each paradigm in the following subsections, in what concerns Web IR.

Web search engines are now well established as one of the most fundamental and widely used components of the Internet infrastructure. Search engines use many of techniques developed over the last decades for full-text document retrieval, but are also quite different in many aspects [12]. Users interact with these systems in a very different way: queries tend to be much shorter and only the first or second results pages are examined in most cases. On the other hand, the hyperlinks between pages represent a source of data that can be used to rank results effectively [13].

Web search engine evaluation has been the subject of substantial research work [14]. TREC has had a Web track in past editions, which will continue this year[2]. In

[2] http://www.ted.cmis.csiro.au/TRECWeb/2002/index.html

this track, evaluations are performed against a document set that is a snapshot of the Web (in 2002, this was a collection of web pages published under the .GOV domain). In the last edition, this track focused on two problems: topic distillation, which involves finding key resources in a particular topic area, and named page finding, or locating a page which has been named by the user. There are multiple classes of algorithms and approaches being used or researched for computing lists of resources to present in response to web search tasks, ranging from keyword matching algorithms to algorithms for finding related pages [15] (using, for instance, bibliometric techniques [16, 17]) or for clustering matched resources by topic [18].

Among the many possibilities to initiate evaluation of web search engines for Portuguese webpages, we highlight the one that consists in defining a set of tasks, execute them against several search engines and evaluate them against several criteria, such as:

- Coverage, the amount of web pages indexed in the Portuguese language.
- Retrieval performance, or the precision of results and response times

To conduct this task, hundreds of search engines could be considered. A large number of search engines could be employed, including global search engines such as Google[3] and Alltheweb[4], and Portuguese-specific national or cultural search engines, such as tumba![5] [19] and todobr[6].

The selection of tasks to benchmark could follow as guidelines the tasks of the TREC Web search track or the Portuguese terms most frequently searched on web search engines (using, for instance, the access logs of one or more search engines as a sample). However, an evaluation conducted at this level would only provide results on the performance of these systems as seen by the end users, without taking into account differences among the collections of web pages used by each search engine or additional database that could be used to improve the quality of ranking algorithms. Researchers on Web IR algorithms demand comparisons on equal terms. To satisfy this requirement, a common, publicly available, collection of web pages extracted from a list of well-known sites could be considered. This collection would be annotated for evaluation purposes and given to the participants. A substantially large document collection of Portuguese web pages, such as governmental web pages, could be gathered from public sources for this purpose and given in advance to participants in a web search evaluation track.

No matter the size and indexing capabilities, IR systems may still fall short of basic user needs, and a user-based study typically focuses on questions such as whether the particular needs are solved, whether the information retrieved was useful in the context of use, whether the interface was friendly, what are the typical shortcomings, and so on. In addition to simple precision and recall measures (incidentally, based on a ill-defined and often problematic view of relevance), several other measures have been proposed to encode subjective factors: novelty (how many retrieved items were new to

---

[3] http://google.com
[4] http://alltheweb.com
[5] http://tumba.pt
[6] http://www.todobr.com.br

the user), coverage (how many were already known to the user), effort (how many irrelevant items before a relevant one), etc. (See [20]).

The problem with this kind of studies is that they are extremely expensive and time consuming. Furthermore, they require users to comprise a (relatively) homogeneous set, with whom researchers can communicate directly. This is at odds with the situation on the Web, where users form a large and mostly heterogeneous group, with whom typically information providers do not have direct contact.

While search engine developers may be interested in comparing algorithms, as mentioned above, other researchers may try to assess search engine users' satisfaction and some of their problems without requiring large user-based evaluations. We may use simple usability techniques such as cognitive walkthroughs and questionnaires, or we can perform large-scale non-intrusive studies by looking at real users in some semi-controlled setups. Web user logs massaging [21] and click-through data [22] are two technologies that we propose to apply for Portuguese.

For our language there are some variables with obvious usability importance that can be studied almost independently of the evaluation setup, such as character encoding, clitics handling, hyphenated words, and a particular kind of spelling mistakes produced by missing or wrong accents. Also, even the simple "language" variable has turned out to be defective in major search engines. For example, in December 2002 only one third of the pages in Portuguese indexed by Google have been correctly assigned the "Portuguese language" label!.

4 An initial proposal for cooperative evaluation

What are we going to evaluate? Is it possible to find a subset or intersection of the above goals and ways to evaluate, discussed in the sections above, which allows a common endeavor? No matter the final answer, one should attempt to devise a setup that could be eventually reused for the different sub-areas. It is in any case necessary to start a common reflection – of which this paper represents the beginning – around IR in Portuguese. Basically, one has to agree on what to evaluate, and how to evaluate it. Almost all questions dealing with evaluation have not yet been answered for Portuguese. Let us illustrate some relevant ones in what follows.

Concerning search engines: Which one behaves best for the most frequent questions in Portuguese? What is the size of the index of each engine? What subjects do they cover? What is their overlap? Concerning users: Which are the most frequent questions? What kind of goals do they have? What kind of mistakes are they most likely to make? What kind of content do they search in Portuguese as opposed e.g. to doing the search in English? Concerning specific QA systems: How are question types categorized? What is defined as a valid answer? What is the proportion of questions that have an answer in the database? Concerning QA: Which kind of questions are users interested in? What is the prior influence (if any) of their exposure to Web search? Concerning information extraction systems, again depending on the information to be extracted, one would like to answer a number of questions, and so on.

Abiding by the evaluation contest paradigm, we have in any case to define a common "document collection". Probably we should have several, or at least two: a closed collection of documents (probably also taken from the Web, given its higher availability), and a large Web collection of documents (open collection). In addition, we need to agree on an initial set of queries or topics, together with some measures of relevance or quality.

We suggest that these collections should be prepared having in mind from the start the requirement that they should be iteratively refined with judgments and annotation relative to more and more complex tasks. Thus, from basic topic detection (classical IR) to QA, (multi-document) summarization, and MUC-like tracks such as ENR, coreference, etc. as well as more linguistically oriented tasks such as terminology extraction or thesaurus induction, or more application oriented tasks such as geographical information mining. In fact, we even propose to maximize the set of common data with other kinds of evaluation contests which are usually kept apart, since evaluation activities for Portuguese are all starting almost simultaneously. One could use a subset of the texts for (machine) translation and (some subsets of) parsing.

Another question to bear in mind is that, though the language is common, Brazilian and Portuguese culture and users may differ considerably. This also applies to the coverage of the indexes of national search engines. Some applications may require one universe only, others may benefit from accessing both Portuguese and Brazilian information. Special care must thus be taken in order to find realistically interesting tasks that pertain to both variants of Portuguese if we want to cooperatively evaluate our systems. One possibility would be to have separate panels of judges for the two variants, so that they might rank differently relevance according to their national (and/or geographical) bias. Examples (for ease of exposition they are framed as questions, but of course the same would apply if they were cast as topics): *Que políticos foram acusados de corrupção ultimamente? Quanto ganha um jogador de futebol? Quais os restaurantes de comida japonesa mais perto? Quem foi Guimarães Rosa?* One would expect different information to be relevant to users from different countries…

Though we acknowledge that many hurdles will have to be overcome, we wish to start with the work to verify or reject our hypotheses, by actually setting up the first collection of queries and documents for Portuguese. This process has begun, and we expect to present its first instantiation at Avalon'2003.

References

1. Voorhees, E., Harman, D.: Overview of TREC-2001. Proceedings of TREC'2001 (2001). Available at http://trec.nist.gov
2. Aires, R., Santos, D.: Measuring the Web in Portuguese. In: Brian Matthews, Bob Hopgood & Michael Wilson (eds.): Euroweb 2002 conference (2002) 198-199
3. Hirschman, L.: The evolution of Evaluation: Lessons from the Message Understanding Conferences. Computer Speech and Language 12 (4) (1998) 281-305

4. Quaresma, P., Rodrigues, I.: PGR: Portuguese Attorney General's Office Decisions on the Web. In: Bartenstein, Geske, Hannebauer, and Yoshie (eds): Web-Knowledge Management and Decision Support. LNCS/LNAI 2543 Springer-Verlag (2003) 51-61
5. Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. Proceedings of COLING-96 (1996) 466-471
6. Santos, D., Rocha, P.: AvalON: uma iniciativa de avaliação conjunta para o português. In: Actas do XVIII Encontro da Associação Portuguesa de Linguística 2002 (forthcoming).
7. Peters, C.: Results of the CLEF 2002 Cross-Language System Evaluation Campaign. In: Peters, C. (Ed.): Working Notes for the CLEF 2002 (2002). Available at: http://clef.iei.pi.cnr.it:2002/2002WN.html
8. Eguchi, K., Oyama, K., Ishida, E., Kando, N., Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop. NII Technical Report, No.NII-2003-002E (2003)
9. Mandl, T., Womser-Hacker, C.: Linguistic and Statistical Analysis of the CLEF Topics. In: Peters, C. (Ed.): Working Notes for the CLEF 2002 Workshop (2002)
10. Orengo, V., Huyck, C.: Portuguese-English Experiments using Latent Semantic Indexing. In: Peters, C. (Ed.): Working Notes for the CLEF 2002 Workshop (2002)
11. Saracevic, T.: Evaluation of evaluation in information retrieval. Proceedings of SIGIR 95 (1995) 138-46
12. Arasu, A. Cho, J. Molina, H. Paepcke, A., Raghavan, S.: Searching the Web. ACM Transactions on Internet Technology 1(1) (2001) 2-43
13. Brin, S., Page,L.: The Anatomy of a Large-Scale Hypertextual (Web) Search Engine. Computer Networks and ISDN Systems, 30(1-7) (1998) 107-117
14. Hawking, D., Craswell, N., Bailey, P., Griffiths,K.: Measuring Search Engine Quality. Information Retrieval l4 (1) (2001) 33-59
15. Dean, J. Henzinger, M.: Finding related pages in the World Wide Web. Proceedings of the Eighth International World Wide Web Conference (1999) 389-401
16. Garfield, E.: Citation Analysis as a tool in journal evaluation. Science 178 (1972) 471-479
17. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents Journal of American Society for Information Science, 24, 4, (1973) 265-269
18. Zamir, O., Etzioni, O.: Web document clustering: A feasibility demonstration. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998) 46-54
19. Silva, M. J.: The Case for a Portuguese Web Search Engine. DI/FCUL Technical Report 03-3 (2003). Available at http://www.di.fc.ul.pt/tech-reports/03-3.pdf
20. Gwizdka, J., Chignell M.: Towards Information Retrieval Measures for Evaluation of Web Search Engines. Unpublished manuscript (1999). Available at http://www.imedia.mie.utoronto.ca/~jacekg/pubs.html
21. Burton, M., Walther, J.: The value of Web log data in use-based design and testing. Journal of Computer-Mediated Communication, 6 (3) (2001). Available at http://www.ascusc.org/jcmc/vol6/issue3/burton.html
22. Joachims, T.: Evaluating Retrieval Performance Using Clickthrough Data. Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval (2002). Available at http://www.cs.cornell.edu/People/tj/publications/joachims_02b.pdf