

Portuguese at CLEF 2005

Diana Santos and Nuno Cardoso

Linguatca, Oslo node, SINTEF ICT, Norway
Linguatca, Lisbon node, DI-FCUL, Portugal
diana.santos@sintef.no, ncardoso@xldb.di.fc.ul.pt

Abstract. In this paper, we comment on the addition of Portuguese to three new tracks in CLEF 2005, namely WebCLEF, GeoCLEF and ImageCLEF, and discuss differences and new features in the adhoc IR and the QA tracks, presenting a new Brazilian collection.

1 Introduction

In order to evaluate cross-language retrieval, the obvious venue is CLEF. However, to add one more language (and/or culture) to a system or evaluation framework is not just to hire a translator and have the job done. This is one of the reasons why Linguatca has taken on the role of organising evaluation contests for systems dealing with Portuguese [1]. Another reason is that to have Portuguese as one of the languages which systems must process, query and/or retrieve within CLEF is undoubtedly beneficial to the processing of the Portuguese language in general. [2].

Our experience at CLEF 2005 reinforced what will be a recurrent idea through this paper: you have to know a language and culture well in order to organise meaningfully evaluation campaigns which include it. Just performing translation of query formulations created in another language, no matter how good, is never enough.

2 Reflections on Adding Portuguese to the CLEF Tasks

2.1 WebCLEF

WebCLEF is a striking example of where knowing the material well would be an advantage. The track could have been significantly improved if people with a working knowledge of each language (and its respective Web [3]) had been involved. The Portuguese collection included in the EuroGOV collection [4] is quite weak. Judging from the Portuguese Web crawls made by tumba! (www.tumba.pt), a Portuguese web search engine [5], we estimated that half of present-day government hosts are absent from the EuroGOV .pt set. In addition, over 70% of the crawl contained webpages from a single site (www.portaldocidadao.pt), just a hub of links to .gov.pt pages. Such an unbalanced collection made it quite difficult to come up with interesting topics that could reflect realistic scenarios of (crosslanguage or other) search in official pages.

2.2 GeoCLEF

Although our participation in GeoCLEF was limited to the translation of topics (and geographical relations), we feel that our attempt to add Portuguese to this track succeeded in pointing out a few serious weaknesses in it. This mainly concerned making sense of the geographical relations. If the “relations” convey meaning there are different implications for translation than if they simply indicate prepositions. However, we could not see a way to express the distinction between “in the south of” and “south of”, in the sense of a subpart of a larger region versus adjacency or simply relative location. Conversely, which fine distinctions hinged upon “in or around” versus “in and around”? In an nutshell, a clear semantics for geotopics was lacking and, thus, translation was obviously hampered. We decided to do a literal translation in most cases, but were far from happy with the resulting “Portuguese” topics.

The lack of a precise semantics for geotopics also caused doubts about scope vs content. For example, a source topic requiring documents about “Amnesty International reports on human rights in Latin America”, was converted to: concept: Amnesty International Human Rights Reports, spatial relation: “in”, location: “Latin America”, which is altogether a different question. Of course, one may claim that the original topics were only a source of inspiration to create new geotopics, but the original user need (reports about human rights violations in Latin America) seems to make considerably more sense than the quest for arbitrary AI reports that happen to be (published?, refereed? criticized?) in Latin America.

2.3 ImageCLEF

Our task at ImageCLEF was to translate English captions into Portuguese, or provide a satisfactory description of the images in Portuguese. These are two different tasks, since what people see – and consequently take pictures of, and then describe in their own language – is extremely conditioned by culture. Most images are not self-explanatory and translation will not help if you do not know the subject, as was obvious for pictures like “golfer putting on green” or “colour pictures of woodland scenes around St Andrews”. Likewise, due to the different meanings of words employed in different languages – different languages cut differently the semantic pie [8] – “people gathered at bandstand” could cover both musical events or people just gathered to take a photo, a vagueness which could not be preserved in Portuguese.

It was also hard to understand the user model of ImageCLEF: specialised librarians of St Andrews or (which) man in the street? Which makes more sense, “dog in sitting position”, or “Timmy, summer holidays, 1990”? And were we justified in (inadvertently) discarding, or conveying, possibly unique presuppositions about royal visits to Scotland and monuments to Robert Burns? It obviously depends on our users.

The most interesting reflection posed by our participation in the ImageCLEF and GeoCLEF exercises is what we call the **organiser’s paradox**. Considering state of the art CLIR systems, which use machine translation and bag-of-words approaches, the more idiomatic translation we provide, the more we harm recall, since the more literal the translation, the easier the system finds the relevant target information. The more natural a translation into a new language, the more understandable it is for a human but the less easy for a CLIR system (at least existing ones) to get sensible answers.

2.4 AdHoc CLEF

Given the addition of new languages with newer collections, topics for this year's adhoc track had by necessity to be more restrictive, since they would have to feature hits both in 1994-1995 and 2002 news documents. This implied, for example, that once-only events could not be selected.

This year a new Portuguese collection was added, containing the Brazilian newspaper *Folha de São Paulo* for 1994-1995¹. As in 2004, we phrased some topics in the Brazilian variant of Portuguese as well as that of Portugal, in order to create a competition as variant-neutral as possible and attract broader participation [2]. We selected the topics to be conveyed in each variant randomly. The table shows how both varieties contributed in the Portuguese document pool and in the final results.

Candidates in Folha	Relevant in Folha	Candidates in Público	Relevant in Público
8213	1,035	12,326	1,869

2.5 QA@CLEF

Compared with last year's track, the changes in QA@CLEF were few [7], which may either denote that a stable setup has been found, or that the large number of languages involved (nine) actually brings some inertia and prevents change. In any case, we would like to discuss two changes in this track: (a) the increase in the amount of definition questions; and (b) the introduction of temporally restricted questions.

Definitions were unchanged from last year, although we had advocated their exclusion in [2]. There are no objective guidelines to evaluate answers of this sort of question and the process of trying to judge them consistently raised some interesting questions. For definition questions about people, we assigned a number of information pieces, and evaluated answers as incomplete ("X") if they included some of these pieces but not all. For example, if the expected correct answer was "Minister of Education of Nigeria", any of the three items (Minister, Education, Nigeria) alone would gain the system an "X". The justification for this procedure is that there could be contexts where just one of the items would satisfy the user. However, this made it no longer possible to guarantee perfect overlap (or perfect corrections given the collections) with the golden resource, since the right answers (items) could be scattered over different documents. In fact a system could get an "X", while nil stood in the golden collection, since there was no document that provided a full answer.

The temporally restricted questions (T questions) lacked a distinction between meta temporal restriction (like "temporal location" as in geoCLEF) and factual temporal restriction (inside the text), which allowed systems to answer them with no special provision. On the other hand, questions involving anaphoric reference to time, like "Which was the largest Italian party? meaning "was but no longer is" not classified as "T", were not considered temporally dependent, even though they are.

From our experience as organizers and evaluators of QA systems, we believe a real assessment of the difficulty of the questions set should be attempted. Although the decision not to provide easily identifiable nil questions was a real improvement this

¹ See the Portuguese CLEF site at <http://www.linguateca.pt/CLEF/>

year, we were still forced to reassess our golden answer set for three different questions for which it had been assumed that there were no answers in the collection, and for which different systems with different strategies were able to actually find a satisfactory answer. Some criteria for ranking QA pairs according to difficulty could be: (a) literal answers, (b) answers in the same sentence (or clause) but with a wording different from the question, (c) answers in separate sentences, (d) answers requiring some reasoning from a human (although not necessarily from a system).

We also suggest that more helpful than right and wrong would be to classify answers to questions as rubbish, uninformative (empty), and dangerous, as we did in [7], providing a more pragmatic view of evaluation. We also suggest that human evaluation should assess things like the following: Is the answer nonsensical so that any user can discover this at once by consulting the alleged justifying passage? Is the answer incomplete but useful? Is the answer complete and right but not supported? Is the answer wrong but (at least apparently) supported? Is the answer informative enough to lead to follow-up or reformulation questions from an interested user?

Finally, if the QA track is to develop into something that really evaluates useful systems for real users, we believe that systems must provide justification passages, in addition to the short answer, instead of just providing the whole document id.

Acknowledgements. We thank Público and Folha de São Paulo for allowing us to use their material, and respectively José Vítor Malheiros and Carlos Henrique Kauffmann for making this practically possible. We acknowledge grant POSI/PLP/43931/2001 from the Portuguese Fundação para a Ciência e Tecnologia, co-financed by POSI, and are grateful to our colleagues at Linguateca who helped with the organization and evaluation in CLEF 2005.

References

1. Santos, D. (ed.): Avaliação conjunta: un novo paradigma no processamento computacional da língua portuguesa, in print.
2. Santos, D., Rocha, P.: The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In: C. Peters et al. (eds.) Multilingual Information Access for Text, Speech and Images. Vol. ??? LNCS, Springer (2005), 821-832.
3. Gomes, D., Silva, M.J.: Characterizing a National Community Web. ACM Transactions on Internet technology. Vol. 5(3), 2005, ACM Press, 508-531.
4. Sigurbjornsson et al.: EuroGOV: Engineering a Multilingual Web Corpus. This volume.
5. Silva, M.J.: The Case for a Portuguese Web Search Engine. Proceedings of the IADIS International Conference WWW/Internet 2003, 411-418.
6. Santos, D.: Translation-based Corpus Studies: Contrasting Portuguese and English Tense and Aspect Systems (2004). Amsterdam/NewYork, Rodopi.
7. Vallin, A. et al.: Overview of the CLEF 2005 Multilingual Question Answering Track. This volume.