

Portuguese at CLEF 2005

Diana Santos[†] and Nuno Cardoso*

[†]Linguatca, Oslo node, SINTEF ICT, Norway

*Linguatca, Lisbon node, DI-FCUL, Portugal

Diana.Santos at sintef.no, ncardoso at 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

To add one more language (and/or culture) to a system or evaluation framework is not just hire a translator and have the job done, although the quest for language-independent systems is still mainstream in natural language processing [1]. This is one of the reasons why Linguatca has taken the role of organizing evaluation contests for systems dealing with Portuguese [2]. In order to evaluate cross-language retrieval, however, the obvious venue is CLEF. To have Portuguese as one of the languages which the systems must process, query and/or retrieve is undoubtedly beneficial to the processing of Portuguese language in general [3].

Our experience this year at CLEF 2005 reinforced what will be a recurrent idea through the paper: you have to know well a language and culture in order to organize meaningfully evaluation contests dealing with it. Just performing translation afterwards, no matter how good, is never enough.

2 Reflections on adding Portuguese to the CLEF tasks

2.1 WebCLEF

WebCLEF is a striking example where knowing well the material is an advantage, as we believe it could have been significantly improved if people with a working knowledge of each language (and its respective Web [4]) had been involved. The Portuguese collection included in the EuroGOV collection [6] is very weak indeed. Comparing to 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, which is just a hub of links to .gov.pt pages.

Such an unbalanced collection made it furthermore quite difficult to come up with interesting topics that could reflect realistic scenarios of (crosslingual or other) search in official pages. In fact, previous studies [7] lead us to believe that, in Web navigation, user goals like research, on-line purchases and general info about travel, holidays and leisure, in addition to the usual sex and latest news, are much more common than information about public services.

2.2 GeoCLEF

Our participation at GeoCLEF was limited to the translation of the topics (and geographical relations). Still, we feel that our attempt to add Portuguese to this track succeeded in pointing

out a few serious weaknesses in it.

Our main remarks concern the geographical relations, or making sense of them. If the “relations” were supposed to convey meaning, this would have different implications for translation than if they were simply indicating 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 distinction hinged upon “in or around” versus “in and around”? In a nutshell, a clear semantics for geotopics was lacking and then, obviously, translation was hampered. We decided to do a literal translation in most of the cases, but were far from happy with the resulting “Portuguese” topics.

The lack of a precise semantics for geotopics, furthermore, brought doubts about what was scope vs. content. An example: While the original topic required documents about “Amnesty International reports on human rights in Latin America”, it got converted into the following trio: **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 right violations that took place in Latin America) seems to make considerably more sense than the quest for arbitrary AI reports that happen to be (published? referred? criticized?) in Latin America.

2.3 ImageCLEF

Our task at ImageCLEF was to translate the English captions into Portuguese, or provide a satisfactory description of the images in Portuguese. Note that these are two altogether different tasks, since what people see – and consequently take pictures of, and thereafter describes in their own language – is extremely conditioned by culture.

First of all, 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 lexical meanings of the words employed in different languages – different languages cut differently the semantic pie [8] – “people gathered at bandstand” could cover both political meetings or people just gathered for the picture taking occasion, a vagueness which could not be preserved in Portuguese; it is also remarkable that, in as much as seven cases, namely those mentioning boat, aircraft, ship, “cart or carriage”, “church or cathedral”, marketplace, and gateway, we had to either add a disjunction or use a more general or more specific term.

It was also hard to understand the user model of ImageCLEF: specialized librarians of St. Andrews, or (which) man in the street? What makes more sense, “dog in sitting position”, or “Timmy, summer holidays, 1990”? And were we justified in (inadvertently) discard, or convey, possible uniqueness presuppositions about royal visits to Scotland and monuments to Robert Burns? It obviously depends on the users we are modelling.

The most interesting reflection posed by our participation in the ImageCLEF and GeoCLEF exercises is what we call **organiser’s paradox**: if one considers state of the art CLIR systems, which use machine translation and bag-of-words approaches, the more elaborate and idiomatic translation we provide, the more we are harming recall, since the more literal the translation, the easier for the systems to get at the right original. The more natural we render a translation into a new language, the more a human user is bound to understand the topic and phrase it that way, but the less a CLIR system (at least the ones existing nowadays) is able 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 in 2002. This implied, for example, that once-only events could not be selected.

This year a new Portuguese collection was added, containing all editions of the Brazilian

newspaper Folha de São Paulo in 1994-1995.¹ So, we attempted to provide some coverage of Brazilian news in addition to Portuguese ones, as was already the case with both American and British varieties for English.

As in last year’s campaign, we attempted to have some topics phrased in the Brazilian variety as well as in the one from Portugal, in order to create a competition as much variety-neutral as possible and attract broader participation [3]. We selected the topics to be conveyed in each variety randomly, without no *a priori* correlation among the variety of the topic and the variety of the document(s) that answer it. Table 1 shows that both varieties contributed fairly for the Portuguese document pool and for the final results.

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

Table 1: Distribution of relevant documents according to Portuguese collection, for the 50 topics

2.5 QA@CLEF

Compared with last year’s track, the changes in QA@CLEF were few [9], 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 changes.

There were, in any case, two modifications of this track on which we would like to cast a critical look: (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 [3] and it was consensual that there had been no objective guidelines to evaluate answers to this sort of questions. Still, the process of trying to judge them consistently arose some interesting questions: In what concerns “definition” questions about people,² we assigned a number of information pieces, and evaluated answers as incomplete (“X”) if they included any of these pieces but not all. For example, if the expected correct answer was “minister of Education of Nigeria”, any of the three pieces (minister, Education, Nigeria(n)) alone would grant the system an “X”. The justification for this procedure is that there could be contexts where just one of the pieces would satisfy the user. However, this made it no longer possible to guarantee perfect overlap (or perfect correctness, given the collections) with the golden resource, since the right answers (pieces) could be scattered among 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.

As to temporally restricted questions (T questions), they lacked a distinction between meta temporal restriction (like “temporal location” analogous to 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, although they critically are.³

From our experience as organizers and evaluators of QA systems, we believe that a real assessment of the difficulty of the questions (given the collections) should be attempted. Although the decision of not to provide nil questions that were trivial to uncover was a real improvement of this year’s track, we were still forced to re-assess our golden answer set for three different questions, which had been assumed not to have answers in the collection, and which different systems, with

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

²We will not discuss here the expansion of organizations’ acronyms, also considered a “definition”, since this is a case where multiple answers should be provided in a real world setting, and are therefore totally out of place in the present CLEF QA setting.

³In Portuguese, the use of the Imperfeito tense presupposes that there is an understood temporal period in mind; if Perfeito had been used, it would convey either a strong presupposition that there had been only one largest Italian party in the past, or that a list of all parties having had that role once was being asked for.

different strategies, were able to counterproof and 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 or wrong would be to classify answers to questions as rubbish, uninformative (empty), and dangerous questions, as we did in [9], providing a more pragmatically oriented view of evaluation. Even more to the point, we 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?⁴ Is the answer informative enough to lead to follow-up or reformulation questions from an interested user?

Finally, we believe that for the QA track to develop into something that really evaluates useful systems for real users, justification passages need be required of a QA system, 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 organization and evaluation in CLEF 2005.

References

- [1] Santos, Diana: Toward Language-specific Applications. *Machine Translation Vol.14* (1999), pp. 83–112.
- [2] Santos, Diana (ed.): *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, in print.
- [3] Santos, Diana; Rocha, Paulo: The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In: Carol Peters et al. (eds.), *Multilingual Information Access for Text Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, Springer, 2005, IST-CNR, pp. 821–832.
- [4] Gomes, Daniel; Silva, Mário J.: Characterizing a National Community Web. *ACM Transactions on Internet Technology*. Vol 5, No. 3, August 2005, ACM Press, pp. 508–531.
- [5] Silva, Mário J.: The Case for a Portuguese Web Search Engine. *Proceedings of the IADIS International Conference WWW/Internet 2003, ICWI 2003*. IADIS, Algarve, Portugal, 5-8 Novembro 2003, pp. 411–418.
- [6] Sigurbjornsson et al.: *EuroGOV: Engineering a Multilingual Web Corpus*. This volume.
- [7] Aires, Rachel; Aluísio, Sandra; Santos, Diana: User-aware page classification in a search engine. In: *Proceedings of Stylistic Analysis Of Text For Information Access, SIGIR 2005 Workshop* (Salvador, Bahia, Brazil, August 19, 2005).
- [8] Santos, Diana: *Translation-based corpus studies: Contrasting Portuguese and English tense and aspect systems*. 2004. Amsterdam/New York, NY: Rodopi.
- [9] Vallin, Alessandro et al.: *Overview of the CLEF 2005 Multilingual Question Answering Track*. This volume.

⁴An interesting example, due to Luís Costa (p.c.), is last year's Esfinge answer to the question "What country is the world football champion?", where the indoor soccer ("futebol de salão") winner was named.