To separate or not to separate: reflections about current GIR practice

Nuno Cardoso
Faculty of Sciences
University of Lisbon
LASIGE group
ncardoso@xldb.di.fc.ul.pt

Diana Santos
Linguateca
SINTEF ICT
Oslo, Norway
diana.santos@sintef.no

ABSTRACT

Most geographical information retrieval (GIR) systems separate the treatment of the geographical and the non-geographical part, often called "thematic". In this paper, we provide an overview of this practice, and we advance arguments for and against. We also show some experimental results that apparently substantiate the non-separation argument. We conclude with the recommendation that this practice should receive more attention by the GIR community.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Evaluation

Keywords

Geographical IR, Geographical Query, Geographical Indexing, Evaluation

1. INTRODUCTION

The interest in geographical information systems and focused geographical search as a subarea of information retrieval is no longer new, with a regular workshop since 2004, GIR [18], and an annual evaluation contest in a cross-lingual setting, GeoCLEF, since 2005 [7, 8, 14]. However, we believe that there has not yet emerged a best practice approach, and we want to discuss a possible reason for this, namely the separation of the geographical terms from the rest of the terms.

Ever since its beginning as a new discipline, geographical information retrieval (GIR) has been thought as adding geographical dimension and processing to an already existing state-of-the-art IR. Cai's paper on geo-libraries [3], although primarily concerned with merging map and text approaches, has been influential in distinguishing among two subspaces in GIR: the thematic and the geographical. The thematic space concerns the subjects or themes that are relevant to the user, while the geographical subspace deals with the scope of the documents found. The thematic space is the usual domain of information retrieval, so, in order to advance the field, geographical information retrieval should concentrate on the geographical part, properly separated from the classic thematic part.

While this may appear a sensible inference, it soon faces the difficulties of dealing with text and textual queries, and the several properties of location in text (surveyed, for example, in Santos and Chaves[26]). In fact, GIR systems to date – possibly due to GeoCLEF – have been mainly trying to solve the problem of finding place names and information in text, which is a natural language processing task. And, to come right to the point, it is hard to separate geographical from non-geographical information in text. (For example, words do not come with a flag meaning "I convey geographic meaning, and only that meaning"...)

This paper addresses this issue in more detail: we start with a survey on the dividing strategies in GIR, to clarify the different approaches taken and eventually compare them, in Section 2. Then, we discuss possible reasons or arguments why these strategies may not work, from a natural language perspective, in section 3. Section 4 adduces some empirical data in favour of the non-dividing camp, while Section 5 concludes with the suggestion that the matter be further looked into by the GIR community.

2. SEPARATING THE LOCATION PART

The most straight-forward way to develop a GIR system is to adapt an off-the-shelf, standard IR engine, and augment it with geographical information and processing modules such as named entity recognizers and gazetteers, and then evaluate how this improves the overall results of the system, for geographical queries. This is the typical GIR approach used by participants along the three editions of GeoCLEF. Yet, no significant improvements over a pure IR approach were shown, which should perhaps ring a bell for the community.

2.1 Query parsing

A very common approach is to consider that a geographical query is a concatenation of two parts: i) the thematic part, and ii) the location part. The thematic part is handled by the classical text retrieval, while the geographical part is funneled to the newly developed geographical approaches [4, 16]. This approach assumes that most geographical queries are represented on a simple "what in where" format, that can therefore be easily divided into the two parts.

The query parsing pilot task in GeoCLEF 2007 [12] illustrates this assumption: it required that participants analysed 800,000 search engine queries, splitting the geographical queries into <*what, spatial relationship, where*> triplets.

NMEIR '08 Glasgow, UK

Also, the first GeoCLEF pilot in 2005 provided an additional topic description in a similar form [8]. This was criticized in [22, 23] for lack of an adequate semantics for the relations, as well as for cross-lingual inappropriateness of the relations themselves, and was not used in later editions, although this might be reflect a lack of consensus among different organizers and not a shared position of GeoCLEF.

2.2 Document geo-indexing

Another frequently employed technique in GIR is the detection of location names in documents, and the creation of a separated geographical index, to store the extracted information.

For example, Leveling et al. [11] use an index of location indicators, that gather into a single index entry all location names and other derivative mentions such as adjective forms, acronyms or postal codes.

The SPIRIT project associates geometric footprints for each location in a separate index, and then used the calculation of polygon overlapping for inferring geographic similarity [9]. Kornai's approach is similar, assigning bounding boxes for each location present in the query, and using MetaCarta's local search engine [10].

With two distinct indexes serving the geographical retrieval module (a term index and a geographical index), the complexity of the GIR approach increases: with two indexes, and hence two independent ranking measures, what is the best way to combine these two relevance measures?

Although Overell et al. avoid this two-index merging problem, by converting the captured locations into unique identifiers that are also indexed along with the text, as terms [16], they are aware that they may simply be adding redundancy.

2.3 Geographic resources

Most GIR researchers rely in some way on geographical ontologies or gazetteers, that provide minimally, geographic names, classification, and coordinates. These can be accurately described as modelling separately the location relations such as inclusion, overlap, proximity and bordering.

This is, from our point of view, a natural and important addition. One has to have geographical knowledge encoded in a way that allows reasoning, and using such repositories will not be argued against, in the scope of this paper. But it is interesting to point out that, in fact, there have been also researchers who used WordNet, and Wikipedia, for getting geographical information from general resources [2, 16]. So this means that, for the sake of completeness, one could also discuss whether general ontologies (or specific ones) deal better with understanding the meaning of places in natural language (and for GIR).

One of the most common uses of such resources is for reasoning about the level of detail of a query (for example, in topic #54, "northern Europe", in an ontology, is likely to have countries such as Norway and Sweden with a "part-of" relationship). Another is to perform disambiguation, since most place names are not unique to a geographic place.

3. NOT SEPARATING THE LOCATION PART

There are nevertheless a set of arguments for not separating the location part, that we will now detail in the next subsections.

3.1 Geographical themes: a contradiction in terms?

Geographical terms are sometimes the theme of a query. To want to know something about Honolulu is as honorable and acceptable as to want to know something about judo. The difference is that the first information need has a strong geographic connotation, while the second has not. It is hard to defend that they should be treated separately a priori. (Nonetheless, it is also true that one might want to know where Honolulu is located, whereas "where is judo" does not make sense. We are not saying that geographic locations do not have different or specific properties, but this subject is not within this paper's discussion.)

3.2 Often the geographic part is contextual

Most geographically-implicit queries should not (and possibly don't) describe where the user is or comes from. This is a contextual datum which is or should be recovered by the query context and not by the query text.

In fact, this is done by major search engines that personalize or localize based on similar users, and one of the similarities may be the geographical origin.

This is the opposite of the case discussed in the previous section; here, the location is possibly extremely relevant but not necessarily expressed (if one is not already addict of search engine tricks).

3.3 Is separation at all possible?

Geographical queries (in the sense of having need for some geographical reasoning or awareness) come in several flavors. According to the typology initially suggested in [25] and then in [8], there are at least eight different kinds of queries that involve geography in some way. Just by considering those kinds of queries it becomes apparent that a separation between the geographical and non-geographical part becomes problematic.

Geographical queries like topic #40, "Cities near volcanos" or topic #56, "Lakes with monsters", just to mention two topic titles of last year's GeoCLEF, are hard to divide that way: the first because there apparently would be no non-geographical part left, the second because it is not exactly the same as the query "monsters in lakes" and therefore this query reformulation (allowing subsequent partitioning of the thematic part "monsters" and the geographical part "restricted to lakes") would miss the point. See [17, 20] for the importance of small words.

In fact, all concrete things occur in space, and the same is true for events. So, most words in natural language refer to more than one feature of an object or concept: its location and many other properties. Often, one needs to understand the text (and the user need) to understand which facet of a particular object or location is at stake. Although this is apparently similar to the ambiguity between *Washington* as a person or as district capital, it is more complex, because we are here pointing to the very **same** concept/object which can be seen from many angles [19, 21]. So, *Brussels* can denote the city, but most often than not by metonymy it describes the EU administration; the *Vienna circle* can denote a group of philosophers or a place in Vienna; while *Lisbon youth* can denote the young people living in Lisbon or the youth of a person spent in Lisbon. In all these cases, Brussels or Vienna or Lisbon are the **same** place with all their connotations, and the co-text selects what is being put in focus/referred to.

Another way of showing the problem with the a priori separation is applying the topic/focus distinction in linguistics, and see that sometimes geo and non-geo information swaps roles: For the type of topics only with scope, such as topic #73, "Events at St. Paul's Cathedral", the focus in on the geographical part: one is interested

in whatever is happening at some place, or at whatever objects or buildings exist at a certain location. For the type of topics that are restricted to a scope, such as "Dogs in Pittsburgh" [29], the focus is on the theme: it is the inverse of the previous case. One is interested in some topic, provided it occurs (or exists) in a certain part of the world. While this may be a useful distinction to understand that it is not trivial to assign geo and no-geo roles to topics, in practice the above topic/focus distinction does not take us far. Even if it is possible, in artificial venues, to produce clear-cut topics of the two above kinds, in most real cases it is not even clear what the user focus is: if one asks for "economy in the Bosphorus region" (topic #66), is one primarily interested in economy, or in the Bosphorus area? Does it really make sense to decide?

3.4 The search argument

Keeping the example of the Bosphorus area open, a typical informed person would also search for names of companies that they knew were operating on that area, or names of economical treaties, or related products. Eventually, names of factories (or factory locations) or ship names that had been in the news. (This is a remark that is relevant as far as log analysis is concerned. Expert searchers might be looking for "economy at the Bosphorus" with other keywords which would fail to be recognized as geographically related search in the first place. See Aires and Aluisio for a pertinent discussion of user intentions versus user activities [1]).

This tells against the current practice of defining geographical queries by those mentioning a geographical term of some sort. A more informed analysis of query logs might yield that a particular set of queries had a strong geographical glue even though no places had been mentioned.

3.5 Is separation useful?

Going back to the assumption that it is possible, in most cases, to separate geographical from non-geographical terms, separate processing misses the following relevant observation: Thematic keywords are often indirectly related to geographic knowledge. For instance, shipwrecks are often found near islands, or coast of oceans, and not on top of mountains or in the Sahara desert. To dismiss all this geographic knowledge (and its implicit co-occurrences for relevance) does not seem to be wise.

3.6 Is separation technically feasible?

Another argument, of a quite different nature, can also be advanced: there is not enough maturity in NLP to be able to really separate and identify all and only geographic terms and interpretations in text. There are still a lot of mistakes (failure to identify locations) and spurious hits (names or words that are considered locations when they shouldn't).

In view of this, a careful study of the importance of such deficiencies into the processing chain might be advisable. For example, Martins et al developed CaGE, a text mining module to capture locations from Web pages, based on a geographic ontology and basic context rules, in order to compute geographic signatures for Web pages to be used in GIR[15]. However, CaGE did not manage to capture most of the geographic evidence in the text collection used in HAREM, a NER evaluation contest for Portuguese [24, 27].

In HAREM we addressed seriously the issue of finding named entities which represented locations in context (and not simply names of places out of context). We therefore produced an evaluation resource which is unique and allows one to assess the difference between a gazetteer-based (or lexically based) and the real use of names for describing locations.

3.7 Summing up

In a nutshell, the problem of identifying something as purely geographical is not an easy task, if possible at all, as will also appear conspicuously when discussing geo-topics in the next section.

All the arguments just listed seem to show that the separation of geographical information from "the rest" may not have been well enough thought of in the first place. We proceed to show that actual practice in GIR systems and their evaluation also backs us in our warnings.

4. EXPERIMENTAL RESULTS

We start by reminding readers that, after three years of GeoCLEF, there is not a single GIR approach that clearly outperforms pure IR systems for the same GeoCLEF tasks [13]. This is indeed negative evidence of some strength for the need for a separate GIR strategy.

In this section, we will present a particular system developed by the first author and colleagues, and the results of the analysis of its performance in GeoCLEF. Although we are perfectly clear that there might be other design flaws in this system, the fact that explicitly investigating the issue of separation showed that it did not work for the particular architectures seems to be yet another valid counter-argument for it.

4.1 A case-study of a GIR system

XLDB's GIR system, co-developed by the first author, participated in all three editions of GeoCLEF, as part of a research project to give geographic capabilities for a Portuguese web search engine [28]. The architecture of the GIR system is shown in Figure 1, and described in detail in [4].

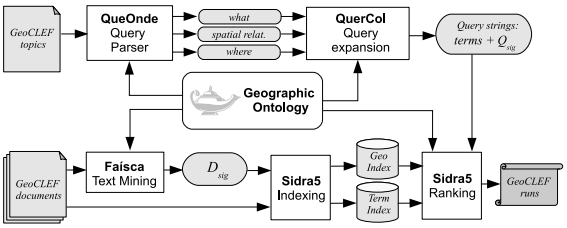

Figure 1: The architecture of the XLDB's GIR system.

In the 2007 edition, the GIR system embraced a purely segregational approach: the QueOnde query parser module divided the GeoCLEF topic titles into <*what, spatial relationship, where*> triplets; the QuerCol query expansion module had different strategies – blind relevance feedback for the thematic part, and an ontology-driven expansion for the geographic part – in order to generate a final query string; finally, the Sidra5 indexing module generates separated term and geographic indexes.

4.2 General analysis of its results

From a preliminary analysis of XLDB's GIR system, we came across the following practical results or doubts:

- The term query expansion (QE) approach adopted is based on blind relevance feedback set, using the top-5 documents and adding the top-8 expanded terms that were weighted higher by the $w_t(p_t\text{-}q_t)$ algorithm [6]. For the 2007 GeoCLEF topics, the QE step re-introduced geographic terms that were later injected in the thematic part.

	Portuguese topic title	English topic title
51	Extracção de petróleo e gás entre o Reino Unido e o continente europeu	Oil and gas extraction found between the UK and the European Continent
52	Crime perto de Santo André	Crime near St Andrews
53	Investigação científica em universidades da costa leste da Escócia	Scientific research at east coast Scottish Universities
54	Prejuízos causados por chuvas ácidas no Norte da Europa	Damage from acid rain in northern Europe
55	Mortes causadas por avalanches na Europa excluindo os Alpes	Deaths caused by avalanches occurring in Europe, but not in the Alps
56	Lagos com monstros	Lakes with monsters
57	Uísque de ilhas escocesas	Whisky making in the Scottish Islands
58	Problemas em aeroportos londrinos	Travel problems at major airports near to London
59	Cidades em que houve reuniões da comunidade dos países andinos	Meetings of the Andean Community of Nations (CAN)
60	Baixas em Nagorno-Karabakh	Casualties in fights in Nagorno-Karabakh
61	Acidentes de avião perto de cidades russas	Airplane crashes close to Russian cities
62	Reuniões da OSCE na Europa de Leste	OSCE meetings in Eastern Europe
63	Qualidade da água na costa mediterrânica	Water quality along coastlines of the Mediterranean Sea
64	Acontecimentos desportivos na Suíça francesa	Sport events in the french speaking part of Switzerland
65	Eleições livres em áfrica	Free elections in Africa
66	Economia no Bósforo	Economy at the Bosphorus
67	Pistas em que Ayrton Senna correu em 1994	F1 circuits where Ayrton Senna competed in 1994
68	Rios com cheias	Rivers with floods
69	Morte nos Himalaias	Death on the Himalaya
70	Turismo no Norte da Itália	Tourist attractions in Northern Italy
71	Problemas sociais na Grande Lisboa	Social problems in greater Lisbon
72	Costas com tubarões	Beaches with sharks
73	Ocorrências na catedral de São Paulo	Events at St. Paul's Cathedral
74	Tráfego marítimo nas ilhas portuguesas	Ship traffic around the Portuguese islands
75	Violações dos direitos humanos na antiga Birmânia	Violation of human rights in Burma

Table 1: Portuguese and English topic titles of GeoCLEF 2007.

- several geographical clues came in the form of landmarks (whose location is known), but which were missed because they were not in the geographic ontology.
- most geographical terms in our geographic signatures did not concern the geographic scope of the document: they could be case of metonymies or simply different facets of that term.

More specifically, a detailed analysis topic by topic, showed the following major sources of problems:

- local conveying property or association: Russian planes are not necessarily in Russia, Scottish research is not necessarily presented only in Scotland, France Press is not only read in France... in other words, the location association is hardly ever a restriction on geographical scope.
- as already referred, many query expansion terms are geographic, but not necessarily relevant for that either... it might be that the most significant expansion for football were Rio de Janeiro, but the topic one was interested in was "Italian football". Then, adding geographical terms outside Italy would probably only diminish performance.
- mention of theme and location in a document may not mean they were related in it: in fact, there was talk about acid rain in one context, and a location in Sweden in another context, and the document was returned as relevant. This is of course a general problem in IR – and thus not specific of GIR – but it tells against providing **one** geographical scope to a document based on the locations discussed in it.

4.3 Query expansion

As mentioned above, by analysing the behavior of the XLDB's GIR system on the GeoCLEF evaluation task revealed that the QE step re-introduced geographical terms in the thematic part, even considering that the initial query was stripped from all geographic names.

We have done an in-depth analysis of the results of this step for the 25 GeoCLEF topics of 2007. Table 1 list them both in English and in Portuguese, for convenience of the reader, but the results and the analysis was done for the Portuguese subtask.

Table 2 presents the top-8 terms re-introduced by the QE module, during the blind relevance feedback step. In bold stand the terms that are considered geographical by the GIR system: it is significant that, out of 192 terms, 71 (37%) are of clear geographic nature.

5. CONCLUDING REMARKS

We believe to have amassed enough data to raise doubts about whether an *a priori* separation between geographic and non-geographic information is appropriate for GIR, a separation we already theoretically attacked in 2006 [26].

Although we are aware that there are several different applications and contexts of use for GIR, and that we are speaking mainly from a GeoCLEF perspective, that is, one of querying geo-topics in newspaper text (and not Web pages or GIS papers), we believe that this reflection can be useful to the whole community, and we make a plea for people to test the particular separation flavour(s) they use in their systems with an open mind.

In particular, we believe that many further empirical studies – especially from the other architectures based on this separation – are required, as well as empirical studies of more general nature, both on

- linguistic issues: how geographical information is encoded in natural language(s) and which other clues may be relevant. For this, the recent trend of relation identification in information extraction may be an important one, see [5, 30].
- user studies: how do location matters really matter for users (of different IR systems). Most probably, different issues will be required for different kinds of task and different kinds of text. Maybe the new pilots at GeoCLEF this year will shed some light on this latter issue (one on Wikipedia and one on image search).

51	[**mar**, empresa, **unido**, **norte**, **reino**, gas, natural, mil]
52	[**santo**, luiz, oswaldo, silva, criminal, delegado, cruz, clodovil]
53	[**edimburgo**, efeito, gases, aquecimento, temperatura, **irlanda**, lugar, cientistas]
54	[cento, **dinamarca**, novo, reduzir, 2005, **oslo**, gases, **florestas**]
55	[**alemanha**, neve, **rios**, chuva, **holanda**, mau, assolar, continuam]
56	[**loch**, **ness**, **lago**, famoso, **ilha**, **mar**, volumoso, passada]
57	[bebida, **ilha_islay**, turfa, scotch, bourbon, single_malt, maltes, casa]
58	[**aeroporto**, **londres**, sido, **heathrow**, voo, passageiros, nomeadamente, contra]
59	[**tomarense**, igat, pedro, marques, autarquia, **tomar**, assistirem, praticava]
60	[**nagorno_karabakh**, **crimeia**, contra, itar_tass, presidente, kremlin, guerra, boris_ieltsin]
61	[**siberiana**, tupolev, 154, **irkutsk**, passageiros, **russo**, companhia, tripulantes]
62	[**hungria**, pacto, estabilidade, **europeia**, **checa**, nato, **leste**, apresentar]
63	[**mar**, objectivo, marinhas, **efluentes**, nascem, ecologistas, reivindicam, cento]
64	[**saas**, **valais**, final, esquiadores, esquiar, slalom, mil, **lausanne**]
65	[**senegal**, **marfim**, **costa**, **ruanda**, saraui, milhares, **ruandesa**, **ruandeses**]
66	[**capital**, sob, acordo, petroleiro_cargueiro, medidas, **turca**, **turcos**, **estreito**]
67	[**silverstone**, gp, pilotos, **pistas**, piloto, **pista**, lehto, grande]
68	[chuvas, problemas, **rio**, urbanos, abastecimento, parque, lercas_adjudicadas, suficiente]
69	[**himalaias**, evereste, alpinistas, gokyo, encontrados, monte, corpos, tinha]
70	[**veneza**, turistas, san, **veneziano**, piazza, comparados, guias, turista]
71	[**oeiras**, **xira**, **loures**, **amadora**, **cascais**, **sintra**, **franca**, **vila**]
72	[steven, brancos, entrar, atacando, comem, alimentam, spielberg, recife]
73	[]
74	[**ilhas**, **ilha**, **miguel**, **faial**, **graciosa**, **jorge**, milhas, **horta**]
75	[suu, nobel, **myanma**, aung, kyi, paz, san, anistia]

Table 2: Top 8 expanded terms for the GeoCLEF 2007 Portuguese subtask.

In fact, for each particular system following the separation architecture, one can always blame the lack of coverage of the ontology or the low recall level of the NER system employed, but this may only be masking a design flaw, which we bring to the consideration of the reader: that of trying to separating what cannot be separated.

Acknowledgements

This work was jointly funded by the European Union (FEDER and FSE) and the Portuguese government, under contracts ISFL/13/408 (FIRMS-FCT), 339/1.3/C/NAC (Linguateca)and PTDC/EIA/73614/2006 (GREASE II). The first author acknowledges FCT grant SFRH/BD/29817/2006.

6. REFERENCES

[1] R. V. X. Aires and S. M. Aluísio. Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em Português. *Revista Ciência da Informação*, 32(1):5–16, 2003. in Portuguese.

[2] D. Buscaldi, P. Rosso, and E. Sanchis. A WordNet-Based Indexing Technique for Geographical Information Retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *Lecture Notes on Computer Science*, pages 954–957. Springer-Verlag, 2007.

[3] G. Cai. GeoVSM: An Integrated Retrieval Model for Geographic Information. In *Proceedings of the Second International Conference on Geographic Information Science, GIScience'02*, pages 65–79, London, UK, 2002. Springer-Verlag.

[4] N. Cardoso, D. Cruz, M. Chaves, and M. J. Silva. The University of Lisbon at GeoCLEF 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[5] J. Chu-Carroll and J. Prager. An Experimental Study of the Impact of Information Extraction Accuracy on Semantic Search Performance. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07*, Lisbon, Portugal, 6–8 November 2007.

[6] E. N. Efthimiadis. A user-centered evaluation of ranking algorithms for interactive query expansion. In *Proceedings of the 16th Conference on Research and Development in Information Retrieval, SIGIR'93*, pages 146–159, 1993.

[7] F. Gey, R. Larson, M. Sanderson, K. Bishoff, T. Mandl, C. Womser-Hacker, D. Santos, P. Rocha, G. D. Nunzio, and N. Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *Lecture Notes on Computer Science*, pages 852–876. Springer-Verlag, 2007.

[8] F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In C. Peters, F. Gey, J. Gonzalo, H. Müeller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF'2005. Revised Selected papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 908–919. Springer, 2006.

[9] C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the Third International Conference on Geographic Information Science, GIScience'2004*, pages 125–139, Adelphi, MD, USA, 20-23 October 2004.

[10] A. Kornai. Evaluating Geographic Information Retrieval. In C. Peters, F. Gey, J. Gonzalo, H. Müeller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised selected papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 928–938. Springer-Verlag, 2006.

[11] J. Leveling and S. Hartrumpf. University of Hagen at GeoCLEF 2007: Exploring Location Indicators for Geographic Information Retrieval. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[12] Z. Li, C. Wang, X. Xie, and W.-Y. Ma. Query Parsing Task for GeoCLEF 2007 Report. In A. Nardi and C. Peters,

editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[13] T. Mandl, F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. *GeoCLEF 2007: the CLEF 2007 Cross Language Geographic Information Retrieval Track Overview.* Presentation held at CLEF 2007, Budapest, Hungary, 20 September, 2007.

[14] T. Mandl, F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[15] B. Martins, M. J. Silva, and M. S. Chaves. O Sistema CaGE no HAREM - Reconhecimento de Entidades Geográficas em Textos da Língua Portuguesa. In *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, chapter 11*, pages 199–215. Linguateca, 2007. In Portuguese.

[16] S. Overell, J. Magalhães, and S. Rüger. GIR experiements with Forostar at GeoCLEF 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[17] K. Pastra, H. Saggion, and Y. Wilks. Extracting relational facts for indexing and retrieval of crime-scene photographs. *Knowledge-Based Systems*, 16(5-6):313–320, 2003.

[18] R. Purves and C. Jones. Workshop on Geographic Information Retrieval. *Computers, Environment and Urban Systems*, 30(4):375–377, 2006.

[19] J. Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, USA, 1995.

[20] E. Riloff. Little Words Can Make a Big Difference for Text Classification. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 130–136, 1995.

[21] D. Santos. *What is natural language? Differences compared to artificial languages, and consequences for natural language processing*. Invited lecture at SBLP'2006 and PROPOR'2006, Itatiaia, RJ, Brazil. 15 May, 2006.

[22] D. Santos and N. Cardoso. Portuguese at CLEF 2005: Reflections and Challenges. In C. Peters, editor, *Cross Language Evaluation Forum: Working Notes for the CLEF Workshop, CLEF'2005*, Vienna, Austria, 21–23 September 2005.

[23] D. Santos and N. Cardoso. Portuguese at CLEF. In C. Peters, F. Gey, J. Gonzalo, H. Müeller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised selected papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 1007–1010. Springer-Verlag, 2006.

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

[25] D. Santos and M. S. Chaves. *The place of place in geographical IR*. Presentation held at the Geographic Information Retrieval workshop, held at SIGIR'2006. `http://www.linguateca.pt/Diana/download/acetSantosChavesGIR2006.pdf`.

[26] D. Santos and M. S. Chaves. The place of place in geographical IR. In *Proceedings of the 3rd Workshop on Geographic Information Retrieval, GIR'2006 (held at SIGIR'2006)*, pages 5–8, Seattle, WA, USA, 10 August 2006.

[27] D. Santos, N. Seco, N. Cardoso, and R. Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, and D. Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'2006*, pages 1986–1991, Genoa, Italy, 22-28 May 2006.

[28] M. J. Silva, B. Martins, M. S. Chaves, A. P. Afonso, and N. Cardoso. Adding Geographic Scopes to Web Resources. *CEUS - Computers Enviroment and Urban Systems*, 30(4):378–399, 2006.

[29] B. Yu and G. Cai. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR'07 (held at CIKM'07)*, pages 49–54, Lisbon, Portugal, 2007. ACM.

[30] S. Zhao and R. Grishman. Extracting Relations with Integrated Information Using Kernel Methods. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05*, pages 419–426, Morristown, NJ, USA, 2005. ACL.