Blueprint of a Cross-Lingual Web Retrieval Collection

Börkur Sigurbjörnsson Jaap Kamps* Maarten de Rijke

Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{borkur,kamps,mdr}@science.uva.nl

ABSTRACT

The world wide web is a natural setting for cross-lingual information retrieval; web content is essentially multilingual, and web searchers are often polyglots. Even though English has emerged as the *lingua franca* of the web, planning for a business trip or holiday usually involves digesting pages in a foreign language. The same holds for searching information about European culture, sports, economy, or politics. This paper discusses the blue-print of the WebCLEF track, a new evaluation activity addressing cross-lingual web retrieval within the *Cross-Language Evaluation Forum* in 2005.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Cross-Language Information Retrieval

1. INTRODUCTION

The world wide web is a natural setting for cross-lingual information retrieval. This is particularly true in Europe: many European searches are essentially cross-lingual. For instance, when organizing to travel abroad for a business trip or a holiday, planning and booking usually involves digesting pages in foreign languages. Similarly, looking for information about European culture, sports, economy, or politics, usually requires making sense of web pages in several languages. A case in point is the current European Union, which has no less than 20 official languages.

The linguistic diversity of European content is "matched" by the fact that European searchers tend to be multilingual. Some Europeans are native speakers of multiple languages. Many Europeans have a broad knowledge of several foreign languages, and especially English functions as the *lingua franca* of the world wide web. Moreover, many Europeans have a passive understanding of even more languages.

The challenges of cross-lingual web retrieval will be addressed in WebCLEF [17], a new track at the *Cross-Language Evaluation Forum* [3, CLEF] in 2005. In this paper we provide a preliminary overview, discussing our view of the cross-lingual web retrieval task, the document collection used, EuroGOV, and the overall set-up of the WebCLEF track.

The remainder of the paper is organized as follows. In Section 2, we describe cross-lingual aspects of web retrieval in the European context. Section 3 discusses the problems involved, and outlines a test-suite for cross-lingual web retrieval. Then, in Section 4, we provide details of EuroGOV, a new web collection for cross-lingual web retrieval. Section 5 details how this collection will be used within the setting of the WebCLEF track at CLEF. Finally, in Section 6, we discuss our findings and draw some conclusions.

2. CROSS-LINGUAL WEB RETRIEVAL

In this section we discuss why the web is a natural habitat for cross-lingual information retrieval.

2.1 Multilingual Web Content and Users

The web is essentially multilingual. Although reliable statistics on web content and web usage are hard to come by, it is evident that the web is increasingly reflecting the linguistic diversity of the world's population. Let us first look at the web's content. Some indicative figures on the distribution of web content over languages are shown in Table 1.[1] On the one hand, it is clear that English still functions as the *lingua franca* of the web. English is by far the most frequently used language. On the other hand, it is also clear that there is a substantial amount of non-English content on the web. The total amount of non-English pages is approaching that of pages in English. European languages

*Currently at Archives and Information Studies, Faculty of Humanities, University of Amsterdam.

[1]Estimates are based on pages in the index of search engine `http://alltheweb.com` in 2002 [for details, see 9]. See also `http://www.cia.gov/cia/publications/factbook/rankorder/2184rank.html` for recent data on number of Internet hosts per country.

DIR 2005, 5th Dutch-Belgian Information Retrieval Workshop.

Table 1: Global Internet Content Statistics. Based on estimated figures from `http://www.netz-tipp.de/languages.html`, 2002.

Web content by language.		
Language	Internet Pages	% Web Content
English	1142.5	56.4
Non-English	2024.7	43.6
European (non-English)	536.9	26.5
Dutch	38.8	1.9
French	113.1	5.6
German	156.2	7.7
Italian	41.1	2.0
Polish	14.8	0.7
Portuguese	29.4	1.5
Russian	33.7	1.7
Scandinavian (total)	17.4	1.3
Spanish	59.9	3.0

Table 2: Global Internet User Statistics. Based on estimated figures from `http://global-reach.biz/globstats`, September 2004.

On-line population by language.		
Language	Internet Access	% On-line Population
English	295.4	35.2
Non-English	544.5	64.8
European (non-English)	285.5	35.7
Dutch	14.0	1.7
French	33.9	4.2
German	55.3	6.9
Italian	30.4	3.3
Polish	9.6	1.2
Portuguese	24.4	3.1
Russian	6.5	0.8
Scandinavian (total)	12.8	1.6
Spanish	72.0	9.0

other than English account for over a quarter of the global web content.

Let us now turn to the web's users. Table 2 gives, again, some indicative figures on the distribution of web users over languages.[2] Here, the situation is even more striking. Nearly two-thirds of the user population has a primary language other than English. Also, the European users excluding native English speakers account for one-third of the whole online user population.

The multilingual nature of the web has prompted many organizations to engage themselves in web globalization efforts [18]. This typically involves localization of web sites tailored to particular markets and users, and is proving inevitable for organizations that get their revenues from web activities, such as e-commerce. Apart from straightforward machine translation, specific cross-lingual retrieval tools and techniques have not yet been adopted by industry [5].

[2]Estimates are based on a variety of sources, e.g., on home access of Internet users [for details, see 6]. See also `http://www.cia.gov/cia/publications/factbook/rankorder/2153rank.html` for recent data on Internet users per country.

2.2 Cross-Lingual Information Retrieval

In 2002, road-map for cross-lingual information retrieval research was suggested by [5]. Gey et al. [5, p.73] list three challenges for cross-lingual information retrieval:

1. Where to get resources for resource-poor languages?
2. Who do we not have a sizable web corpus in multiple languages?
3. Why aren't search engines using our research?

The second challenge is addressed head-on in this paper. Gey et al. [5] also point out a potential problem for the evaluation methodology if English is the dominating language of web pages in a collection. Consider a set of ad hoc retrieval topics for which there are many relevant pages in English. A system focusing exclusively on English will yield very good performance, which is in contrast with the intentions of cross-lingual retrieval. There is a need for a multilingual web collection that is not dominated by one particular language. The obvious candidate is a collection based on European web content.

Cross-lingual information retrieval has been high on the agenda ever since the early years of the web. There has been an interesting shift in focus over the years. Early studies of cross-lingual retrieval, such as [10], focused on monolingual users wanting to search a collection of documents that they cannot read. Recent studies, such as [14], focus on polylingual users wanting to search documents in the languages that they can understand. One of the earliest studies taking into account users understanding multiple languages is [2]. Capstick et al. [2] investigate a system in which users can express their query in their native tongue, while retrieving documents in several languages of which the user has, at least, a passive understanding. In a series of publications, Petrelli et al. [12, 13, 14] have argued convincingly that bilingualism or polylingualism is the rule for many potential users of cross-lingual retrieval systems. As the authors put it, "it is not unusual to find people who are fluent in 4 or 5 languages." Petrelli et al. [11] highlight that many people use different languages in their everyday work, think of journalists, business analysts, professional translators, information professionals, and, of course, scientists. Their varying degrees of knowledge of the languages to search, their generic search expertise, and the final task to perform (e.g., search-only versus search-and-use) create different user classes with different information needs.

3. TOWARD A TEST COLLECTION

In this section we discuss some of the main challenges in building a cross-lingual web retrieval collection, and outline how a such a test collection could be set-up.

3.1 Requirements

As pointed out by Gey et al. [5, p.73], it is non-trivial "to define suitable criteria for the construction of a valid multilingual Web corpus for R&D." Based on our discussion of cross-lingual web retrieval in Section 2, we draft a tentative list of requirements we would like a cross-lingual web retrieval test suite to satisfy.

Ideally, a cross-lingual web retrieval test collection

- should cater for users that are polyglots;
- should address user tasks that are essentially multilingual;
- should have documents in a wide variety of languages;
- should not be dominated by a particular language, i.e., English;
- should be of sufficient size; and
- should be a natural domain for multilingual search.

3.2 Challenges and Solutions

Of course, not all languages are equally acceptable as a vehicle for conveying information to a particular user. It would be natural and attractive to conduct retrieval experiments in a setting where users store profiles in which they list the languages they can read. This brings us directly to one of the main challenges involved in building a truly cross-lingual web retrieval collection: just as users are not able to read all languages, so will individual assessors be unable to provide relevancy judgments for pages in each and every language in a cross-lingual web collection (let alone be a native speakers of each language). At the same time, making relevance judgments based on topical relevance would require the assessors to judge the content of a page regardless of the language in which it is expressed [15, 16].

This is a fundamental challenge that we cannot, and will not be able to resolve, but we can try to minimize the extent to which it affects the cross-lingual web retrieval test collection. Our strategy is based on two key ingredients.

Known-item Search We will focus on *known-item search* exclusively, that is, on tracking down pages known to exist in the collection. This will imply that the (original) target page is fairly unique, although identical mirrors of the page, or translations of the page's content into other languages may occur. Known-item search is a natural task in a web environment, and has some obvious further advantages in the limited assessment effort needed to create the test collection.

Monolingual Topics Our test collection will be built from sets of monolingual topics targeting a particular language or domain of the collection. This implies that topics should have a national focus, making them unlikely to occur in other languages/domains. The original topics will be translated into English and, potentially, other languages, allowing for bilingual and multilingual retrieval against the original monolingual judgments.

To sum up, we face the challenge that users and assessors are polyglots, but not "omniglots." By focusing on monolingual known-item search, assessors should primarily judge pages in their native tongue. If the site provide translations of the target page, these are generally easy to identify. The occurrence of an English version of an originally non-English page is frequent, and sometimes there are translations to a whole range of languages. Can we exclude that relevant pages occur in an unexpected language or domain? No, we cannot. Think of a foreign embassy hosting a content-wise identical page in a different language and a different domain. In this sense the recall base may be incomplete, but we expect that this will not affect the quality of the test-suite.

Note that our focus on known-item search also avoid us falling victim to the concerns of [5]: English pages will not dominate the set of relevant pages for such topics. This does not imply that we are not interested in general ad hoc topics, just that we want to start with known-item search topics. For assessing general ad hoc information needs, the problem of having to assess pages in all the collection languages is unavoidable. The pragmatic solution would be to include a limited set of target languages, i.e., those languages that the topic creator can read, in the topic statement. Note that this may lead, again, to the dominating language problem if one particular language, i.e., English, can be read by all topic creators.

3.3 Outline of a test-suite

Based on our discussion above, we envision the following sets of known-item search tasks.

Monolingual Tasks a set of 50 known-item search topics in a single language, targeting pages in the same language.

Mixed Monolingual Task a set of 200–500 known-item search topics in multiple languages in which the language of the topic statement is typically the language of the target page.

Bilingual Tasks a set of 50 known-item search topics in English, targeting pages in a single other (i.e., non-English) language or domain.

Multilingual Task a set of 200–500 known-item search topics in English, targeting pages in any other (i.e., non-English) language or domain. This may require the extraction of language cues, for example, a topic like "Danish minister of . . . " is likely to target pages in Danish, or from the `.dk` domain.

The mono- and bilingual tasks can be organized out as subtasks of the mixed monolingual and multilingual tasks, respectively.

More complex mixtures are possible, also revealing more information:

Language Identification Model the use of language identification tools: *What language do pages in the collection have? What language does the topic have?*

Language Cue Extraction Model the use of language cue extraction: *What language does the targeted page have?*

Search Intentions We could also model user interaction by revealing part of the searcher's intentions: *What language or domain does the targeted page have?*

4. EUROGOV COLLECTION

Cross-lingual web retrieval requires a new document collection to be constructed, containing web content in many languages. Of course there are many options for creating such a collection. Multi-lingual documents are abundant on the web. We have chosen to focus on pages of European government-related sites, where collection building is less restricted by intellectual property rights. We baptize

Table 3: Main domains in the EuroGOV collection, and the dominant languages (based on preliminary page counts).

EuroGOV Collection.		
Domain	Predominant language	# of Pages
.cz	Czech	690,673
.de	German	887,260
.es	Spanish	735,310
.eu.int	Mixed	3,710,000
.fi	Finnish	868,100
.fr	French	1,399,653
.hu	Hungarian	230,830
.it	Italian	570,506
.nl	Dutch	388,470
.pt	Portuguese	186,783
.ru	Russian	185,000
.se	Swedish	312,000
.uk	English	829,740

this collection EuroGOV. We think of this collection as an European counterpart of the .GOV collection.

Our initial plan was to obtain a focused crawl from the European Union seed `.eu.int`. However, restricting a crawler to government-related sites proved highly non-trivial. The collection we want to crawl is fairly heterogeneous, for example in the number of document languages. For some governments the crawling is smooth and we can easily filter out governmental pages (notable examples include `.gov.uk` and `.regeringen.se`). Most governmental sites, however, have more complex structures, and we could only focus the crawl by providing an explicit list of domains. As an example, we crawled 13 different domains to gather pages from the Finnish government. As the following domain list shows there is no easy way of identifying Finnish governmental domains:

> `defmin.fi`, `formin.finland.fi`, `intermin.fi`, `ktm.fi`, `minedu.fi`, `mintc.fi`, `mmm.fi`, `mol.fi`, `om.fi`, `stm.fi`, `vm.fi`, `vnk.fi`, and `ymparisto.fi`

These differences in domain naming traditions will make it difficult to guarantee completeness of some governments. As a result, EuroGOV will contain, as a minimal requirement, a fairly complete content of

- main government portals
- main ministries

4.1 EuroGOV Collection Characteristics

The EuroGOV collection will contain over 10 million pages; this is an indicative figure, based on our current set of seeds and the coverage of these domains by Internet search engine `http://google.com/`. For practical reasons, we will only release a 3 million page subset of the full EuroGOV collection for the WebCLEF 2005 evaluation campaign. The countries and domains included for EuroGOV are chosen in accordance with current CLEF interests and plans. Table 3 gives the preliminary page counts for each of the main domains in the collection. The distribution of the main domains is visualized in Figure 1. Further countries from whose government portals are being considered for inclusion include

- `at`, `be`, `cy`, `dk`, `ee`, `gr`, `ie`, `lu`, `lv`, `mt`, `pl`, and `sk`.

Note that pages in the languages of these domains will 'creep in' anyway. For example, the `eu.int` domain have ample pages in all the 20 official languages of the European Union.

The EuroGOV collection will feature more languages and countries than used in WebCLEF 2005 tasks. We made a deliberate choice to go for this extended list of countries and domains. On the one hand, this will facilitate future task extensions for cross-lingual web retrieval, or re-use of the collection for other purposes. On the other hand, we feel that this reflects the natural situation when building a 'European' search engine. Of course, participating teams at WebCLEF will be free to select only parts of the collection to index for a specific task.

4.2 EuroGOV Availability

The EuroGOV Collection will be made available in January 2005 [17]. The crawled pages will have been cleaned-up and put in a uniform format. The resulting pages will be bundled and compressed in manageable sizes. The collection will be distributed over the Internet; if a participant's local band-width is not sufficient, DVDs can be shipped on request. The collection will be distributed under a license restricted to research use only.

5. WEBCLEF TRACK AT CLEF

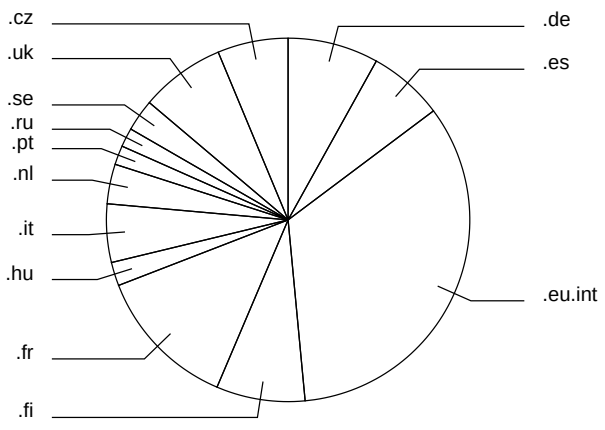
The precise WebCLEF Track guidelines will be determined in early 2005. We intend to involve track participants actively during topic creation and peer-assessments.

5.1 Topic Creation

Topic creators will be asked to create monolingual topics, targeting pages in the topic's language. We ask participants to

- focus on a particular domain/language combination,
- create $\pm$ 50 monolingual known-item search topics, and

Figure 1: Composition of the EuroGOV collection (based on preliminary counts).



- provide English translations of the topics.

The monolingual topics form the core of the monolingual tasks, as well as for the mixed monolingual task. The translated topics will be used for the bilingual and multilingual tasks. Depending on interest and available language expertise, further translations may be provided to accommodate other language combinations.

Bilingual topics will be the result of the translation of the original monolingual topics. This will require to make make default assumption explicit, e.g., consider the Dutch monolingual topic

- "*minister van buitenlandse zaken.*"

A literal translation of this topic into English would be

- "minister of foreign affairs."

However, the implicit assumption underlying the Dutch topic, because it is formulated in Dutch, is to find information about the

- "**Dutch** minister of foreign affairs."

By making these default assumptions explicit in the translated topic, the relevance judgments on the original Dutch topic statement will carry over to the translated version.

5.2 Assessment and Evaluation

There will be a rather limited assessment stage, in which we verify that the original target page is unique. Topic creators will be asked to assess their own topics, and identify whether similar or identical content occurs on other pages, possibly in a different language. The results will be evaluated by the familiar measures for early precision, including mean reciprocal rank of the first found relevant page, and success at 1, 5, and 10.

5.3 WebCLEF 2005 Tasks

In the first incarnation of the WebCLEF track, we will focus on a small number of core tasks. Building the WebCLEF test collection will be a community effort, and the specific languages for which topic and judgments are available will depend on the available language expertise. We expect to provide, at least, topics and judgments in the following eight target languages: Dutch, English, French, German, Italian, Portuguese, Russian, and Spanish.

At the time of writing, we consider focusing on two main tasks.

Mixed monolingual Using 50 topics per target language, for at least the eight languages mentioned above. (This is a natural task when building a 'European' search engine.)

Multilingual Using the topic language English, based on the translations of the monolingual topics provided by the topic authors. (This is a natural task when catering for the information needs of a polyglot.)

As a side-product we will also evaluate on the individual monolingual and bilingual retrieval results for each of the target languages.

The topics will contain additional topic fields revealing the topic language, and the language and domain of the intended target page.

6. DISCUSSION AND CONCLUSIONS

The history of building web retrieval test collections has been very well documented in [8]. Our decision to focus on known-item search is largely based on experiences during the web tracks at the Text REtrieval Conference (TREC). An important difference with earlier web retrieval test collections is that we decided to crawl the fairly complete content of a number of sites, rather than letting the crawler navigate freely on the web. This may have interesting effects on the link-structure of the resulting collection. Earlier efforts put much stress on replicating a natural link structure in a web collection [1, 7]. In our collection, we anticipate that each individual site will exhibit a fairly dense network of navigational links, but that the network linking individual sites will be much less dense. Hence, the resulting collection could be viewed a heterogeneous collection of sites.

The EuroGOV web retrieval collection discussed in this paper is distinct from the collections used at TREC by its focus on cross-lingual retrieval. However, web retrieval collections have also been constructed outside the TREC framework. Of particular interest is the web task at NTCIR-3 [4]. Here, a crawl of the Japanese `.jp` domain is used, containing mostly Japanese (90%) and English (8.3%) pages. Our proposal for a cross-lingual web retrieval collection substantially extends the number of languages in that collection, and the planned cross-lingual web track will enable the evaluation of a variety of monolingual, bilingual, and multilingual web retrieval tasks.

Summarizing, in this paper we discussed a wide range of facets of cross-lingual web retrieval. We analyzed the diversity of languages of pages on the web, as well as the native languages of web users. We distinguished two user types for cross-lingual retrieval: on the one hand an essentially monolingual user who searches pages in languages that she cannot read, and, on the other hand, a polyglot who searches in languages that they have some level of proficiency in. We decided to focus on the second type of user, and outlined a blueprint for a cross-lingual web retrieval test collection. The proposed test suite has a document collection, baptized EuroGOV, containing documents from various European governmental sites. This will avoid the dominance of a single language, i.e., English, and provides a natural setting for multilingual web search. We highlighted the problem that users and assessors may be polyglots but no "omniglots," i.e., they will not be able to read all languages in the collection. As a result, we plan to build the test collection around monolingual, known item search topics. By providing translations of the topics, we will create a true bilingual and multilingual test sets.

In its first incarnation in 2005, WebCLEF will focus on two main tasks, *mixed monolingual retrieval* and *multilingual retrieval* using the English topic set. We view this as a stepping stone toward realizing the full potential of cross-lingual web retrieval. Building a test-collection for cross-lingual web retrieval is essentially a community effort: we depend on participants providing language expertise, natural cross-lingual information needs, and relevance judgements. Future editions will address, amongst other things, resource-poor languages, more natural cross-lingual search scenarios, and search engine efficiency.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, and 612.069.-006.

REFERENCES

[1] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, 39:853–871, 2003.

[2] J. Capstick, A. K. Diagne, G. Erbach, H. Uszkoreit, A. Leisenberg, and M. Leisenberg. A system for supporting cross-lingual information retrieval. *Information Processing and Management*, 36:275–289, 2000.

[3] CLEF. Cross language evaluation forum, 2004. `http://www.clef-campaign.org/`.

[4] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. An evaluation of the web retrieval task at the third NTCIR workshop. *SIGIR Forum*, 38:39–44, 2004.

[5] F. C. Gey, N. Kando, and C. Peters. Cross language information retrieval: a research roadmap. *SIGIR Forum*, 36(2): 72–80, 2002.

[6] Global Reach. Global internet statistics by language, 2004. `http://global-reach.biz/globstats`.

[7] C. Gurrin and A. F. Smeaton. Replicating web structure in small-scale test collections. *Information Retrieval*, 7:239–263, 2004.

[8] D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[9] Netz Tipp. Internet statistics: Distribution of languages on the internet, 2004. `http://netz-tipp.de/languages.html`.

[10] D. W. Oard. Serving users in many languages: Cross-language information retrieval for digital libraries. *D-Lib Magazine*, December, 1997.

[11] D. Petrelli, M. Beaulieu, and M. Sanderson. User participation in CLIR research. In F. C. Gey, N. Kando, and C. Peters, editors, *Cross Language Information Retrieval: A Research Roadmap*, pages 43–47, 2002.

[12] D. Petrelli, M. Beaulieu, M. Sanderson, G. Demetriou, P. Herring, and P. Hansen. Observing users, designing Clarity: A case study on the user-centered design of a cross-language information retrieval system. *Journal of the American Society for Information Science and Technology*, 55: 923–934, 2004.

[13] D. Petrelli, P. Hansen, M. Beaulieu, and M. Sanderson. User requirement elicitation for cross-language information retrieval. *The New Review of Information Behaviour Research*, 3, 2002.

[14] D. Petrelli, S. Levin, M. Beaulieu, and M. Sanderson. Which user interaction for cross-language IR? Design issues and reflection. *Journal of the American Society for Information Science and Technology*, 56, 2005.

[15] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

[16] E. M. Voorhees. The philosophy of information retrieval evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2002.

[17] WebCLEF. Cross-lingual web retrieval, 2004. `http://ilps.science.uva.nl/webclef/`.

[18] J. Yunker. *Beyond Borders: Web Globalization Strategies*. New Riders Publishing, 2002.