Multilingual Search in Libraries. The case-study of the Free University of Bozen-Bolzano

R. Bernardi*, D. Calvanese*, L. Dini†, V. Di Tomaso†, E. Frasnelli*, U. Kugler*, B. Plank*

* Free University of Bozen-Bolzano,
Bolzano, Italy
{bernardi, calvanese, plank}@inf.unibz.it, {Elisabeth.Frasnelli, Ulrike.Kugler}@unibz.it
† CELI s.r.l
Torino, Italy
{dini, ditomaso}@celi.it

Abstract

This paper presents an on-going project aiming at enhancing the OPAC (Online Public Access Catalog) search system of the Library of the Free University of Bozen-Bolzano with multilingual access. The Multilingual search system (MUSIL), we have developed, integrates advanced linguistic technologies in a user friendly interface and bridges the gap between the world of free text search and the world of conceptual librarian search. In this paper we present the architecture of the system, its interface and preliminary evaluations of the precision of the search results.

1. The problem

In this paper, we present the MUSIL (MUltilingual Search In Libraries) system developed within an on-going project on the enhancement of an OPAC (Online Public Access Catalog) search system with multilingual access. The project aims at integrating advanced linguistic technologies in a user friendly interface and bridging the gap between the world of free text search and the world of conceptual librarian search.

The need of multilingual access to textual information is perceived worldwide and is particularly relevant for libraries that operate in a multi-cultural context, like the Library of the Free University of Bozen-Bolzano (FUB), which is a multilingual (Italian, German, and English) university offering several international study programs. Moreover, due to the collaboration with both German and Italian libraries, FUB librarians use both the Italian and German subject headings systems ("Soggettario italiano" and "Schlagwortnormdatei" (SWD)) for cataloging bibliographic items. Additionally the "Library of Congress Subject Headings" (LCSH) are used to catalog bibliographic items in English.

The evaluation of the FUB Library OPAC search logs shows that a considerable portion of the queries are "duplicated" in two languages and that many are even repeated in three languages, namely Italian, German, and English. This situation is clearly a major barrier in accessing the Library Catalog: besides wasting the users' time, this search method does not guarantee users to find all possibly relevant books either written or cataloged in a language different from the one of the query terms.

To address the problem of multilingual search, a possible approach is to exploit possibly complex mappings between terms in different languages (Landry, 2004). A complementary approach is based on Information Retrieval, which offers well established solutions to this problem (cf., (Peters et al., 2003)). However, the proposed techniques are often statistically based and as such do not perform well enough when information available is scarce. This is precisely the case in library databases, where documents are described by means of only few words, typically authors, title, and subject headings, sometimes also summary and abstract. Moreover, standard query translation approaches usually range over a limited domain, but the domain of a general purpose library such as a university library is by definition the whole of human knowledge. Therefore, the present project aims at tailoring standard IR methods, and in particular CELI's search engine –DOCDIGGER– to the library catalogs structure and the library users' needs.

Currently, MUSIL combines linguistic knowledge, like stemming, grammars, dictionaries and thesauri, with statistical methods for data retrieval to improve precision and recall of the search. It provides automatic translation of query terms into Italian, German, and English and suggestions of related terms on the basis of the semantics of the query (thesaurus look-up). Moreover, it performs a free text search looking for the words entered by the user in the titles, table of contents, notes, Subject Headings (SHs)[1]. By giving a search term in a language of his/her choice the user is able to search the library catalogs and find relevant documents written or cataloged in any one of the three languages. To enrich the search results, the user can also choose to expand the query by searching also for related terms (broader, narrower, synonyms, etc.) obtained by means of thesauri (Dini et al., 2005). This option is of particular relevance when the input and target languages do not match and thus translation is required.

We are currently working on optimizing search results, specifically to improve the ranking of the retrieved documents. We further plan to work on the improvement of the search interface, by presenting users relevant portions of a thesaurus which s/he can exploit to refine query terms (Dongilli et al., 2004). Another priority will be the development of adaptive systems able to handle the rapid evolution of library catalogs and addition of new words.

The next Section describes the system architecture and its

[1] We are currently digitizing the abstracts of the archived documents and extend MUSIL search to them too.

interface. Section 3. presents the preliminary results of a laboratory test on a sample of the Library Catalogs, and Section 4. summarizes the identified needs for further research.

2. System description

This section briefly describes the MUSIL architecture and illustrates the multilingual search features and its interface.

2.1. Architecture

MUSIL is based on two main components, DOCDIGGER and OPAC, whose functionalities are briefly described below.

DOCDIGGER[2] is an information retrieval and search engine. It extracts an index from sets of given documents that are first converted into a suitable format by means of stemming and part of speech tagging[3]. It searches for the query terms and their morphological variations, as well as their translations and expansions.

OPAC provides functionalities for catalog search via bibliographic information (such as author, title, ISBN, etc.) or via subject headings and classifications. It offers also services for the library users (e.g., access to the library account, ordering an item, etc.).

The goal of MUSIL was to enhance the search functionalities of OPAC with the capabilities offered by DOCDIGGER, while preserving the additional services of OPAC not related to search. A further goal was the adaptation of DOCDIGGER to the specific requirements of the library domain. In MUSIL, the integration between DOCDIGGER and OPAC was achieved by developing an interface module that preserves the existing services of OPAC (e.g., access to library account, etc.), while offering the multilingual search functionalities of DOCDIGGER. More specifically, the integration of the two systems has been carried out along the following lines:

i) The user accesses the search functionalities via the OPAC interface, and the query s/he poses is forwarded to DOCDIGGER. Note that this required extending the OPAC interface with the functionalities for specifying the additional parameters related to multi-lingual and thesaurus-enhanced search.

ii) DOCDIGGER exploits the data in the library database to compute the answers to queries. More precisely, DOCDIGGER periodically accesses the library catalog database and incrementally updates an index of the data therein. Queries are then answered by accessing the index only (see Figure 1).

iii) The result of the multi-lingual search is transferred from DOCDIGGER to the OPAC interface and shown to the user. Again, this required extending the OPAC interface, e.g., to allow for grouping the results according to the language, and for visualizing the terms obtained by translating and expanding the original query terms. An example of the displayed results is given in Figure 3.

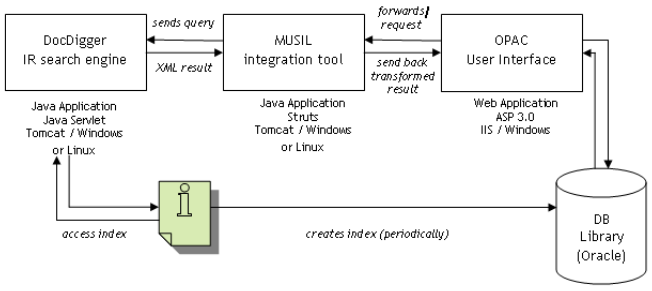

Figure 1: MUSIL architecture

2.2. Language Functionalities

In MUSIL the traditional OPAC search interface has been extended with the following language related functionalities (see Figure 2). The user must state the search term language (English, Italian or German) and can choose different search modes:

Translate. The system looks for the search term's translations. The user can specify the target language of the translation (document language), by default "all languages" is chosen;

Expand. The system looks for the search term and conceptually related terms only in documents of the same language of the search terms;

Translate and Expand. The system looks for the search term's translations and their conceptually related terms, too.

In the default mode, when both translate and expand are disabled, the system looks only for the search term and its linguistic variations (i.e., terms sharing the root of the search term, e.g "security" and "securities", "light" and "lighting".)

The retrieved documents are clustered per language and ranked on the base of their relevance, as shown in Figure 3. When the Translate and/or Expand modes are activated, the system shows all the terms used for the search. For instance, for the English query term "probability" by clicking on "Explanations" the users will be given the following translated terms: Wahrscheinlichkeit (German), probabilità, simile, verosimiglianza, accidentalità, casualità (Italian) (see Figure 3.)

Figure 2: MUSIL user interface: functionalities

[2] http://www.celi.it/motore_ricerca.html

[3] The current version integrated in MUSIL has the PoS tagging feature disabled.

Figure 3: MUSIL user interface: output

3. Evaluation

The analysis of OPAC logs shows that users looking for books by using field search (Subject headings and Author/Title) are 27%, while the ones using free-text search are 73%. For the former, the best solution for multilingual access is probably the use of mappings between different national Subject Headings (SHs) (which is the aim of the MACS (Landry, 2004) project). On the other hand, for the latter group this solution cannot be completely satisfactory unless integrated with language technology tools for free text search. (See (S. Michos and Fakotakis, 1999) for an evaluation with real-life users of multilingual tools to access Library Catalogs). Moreover, also at the level of the output, using language information to cluster documents helps the user finding the relevant ones more easily. This would be, for instance, the case of books found by searching for query terms like "computers" that are used in several languages. Similarly, presenting a ranked list of retrieved books can help users speed up his/her search. However, both free text search and ranking techniques must be tailored to the specific application of Library Catalogs, as our first evaluation experiments show.

Evaluating an OPAC search system presents several challenges and asks for decisions to be made by the investigators (Tague-Sutcliffe, 1997). In order to evaluate the current version of MUSIL, we have started with a laboratory test and have postponed the operational one with real-life users to a later stage. The laboratory test has been set up following (Wien, 2000) as described below.

Sample DB A sample database has been extracted from the Library Catalogs, by choosing those subject areas of the classification system (RVK) with number of records higher than 300 for each language. Out of the 22 subject areas, we were left with 7 categories, viz. Psychology and Philosophy (C), Pedagogy (D), Law and Sociology (M), Law (P), Economics (Q), Mathematics and Computer Science (S), Agriculture and Technology (Z). We extracted 1.099 records per language chosen by means of stratified random sampling to come up with a balanced distribution among categories (viz. 157 records per category). Finally, we removed duplicates.

Queries Queries have been generated automatically, by first selecting as *candidate query terms* those terms occurring both in the titles of the sample DB and in the Controlled Vocabularies of the SHs of the corresponding language. This selection resulted into the following groups: 1.022 terms for German, 404 terms for English, 551 for Italian[4]. Then, these sets of candidate terms have been reduced to *query terms* by selecting only those matching subject headings assigned to records of the corresponding language contained in the sample DB resulting into: 473 terms for German, 224 for English and 282 for Italian.

When carrying out an evaluation of a (multilingual) Library Catalog search system, the first mayor difficulty to face is to define the set of documents to be considered relevant for a given query term, and against which to evaluate precision and recall of search results. Different criteria could be used. Relevant documents can be considered (a) the ones containing the query term in the Subject Headings (Wien, 2000), where the document is of the same language of the query term; (b) the ones containing a linguistic variation of the query term in the SH; (c) the ones considered as relevant by area experts; (d) the ones that satisfy real-life users needs, etc. Secondly, when evaluating multilingual access these criteria have to be extended to the set of documents in the translation target language. So far, we have based our evaluation on the criterion (a) in order to compare search based on controlled vocabulary vs. free text search.

Titles are not always indicative of the topic of the record, for instance, the query term "God" matches the "Democracy - the God that failed", but clearly the document cannot be considered as relevant. The SHs assigned to it (Economics / Political aspects Economics / Moral and ethical aspects Economics / Moral and ethical aspects Economic policy Monarchy Democracy Anarchy) would be much more helpful and significant. This example brings evidence in favor of the (a) criterion above. On the other hand, a perfect matching between the query term and the assigned SHs

[4]The differences among the candidates query terms per language are due to the differences in the controlled vocabularies at disposal. In particular, the Library has full access to SWD (German SHs) whereas it has only a limited access to LCSH and "Soggettario Italiano" (the latter is not available in electronic form yet hence the available SHs are only the ones used by the FUB librarians).

could be too restrictive as relevance criterion. This is already an obvious consequence of the fact that SWD uses mostly singular SHs whereas the LCSH and the "Soggetario Italiano" use plural SHs more often.

Both the classical OPAC search system and MUSIL look for the query term in the titles, subtitle, notes, and assigned SHs. Moreover, MUSIL (also in the default mode) looks also into table of contents and "related terms" (narrower, broader, ...) of the Controlled Vocabularies[5]. Hence, by assuming the (a) criterion, both system have 100% recall, while they could differ in their precision. On the one hand, OPAC retrieves also (i) documents of languages different from the query term's language, and as such not relevant; these documents are not clustered together with the relevant ones by MUSIL. On the other hand, MUSIL, in the default mode, retrieves also (ii) documents containing linguistic variations of the query terms, as well as (iii) those documents that have the query term or a linguistic variation of it in the "related" SHs. Neither of these documents are found by OPAC. We are currently evaluating the frequencies of these cases, and will then check user judgments on the relevance of the documents (ii) and (iii). The difference between (i) and (ii)-(iii) are not well represented in the precision results summarized in Table 1, which are obtained on the sample DB as the average of the precision over all selected query terms (DOCDIGGER has been used in the default mode).

	OPAC	MUSIL
German	0,61	0,65
English	0,50	0,49
Italian	0,49	0,49

Table 1: OPAC and MUSIL precision

Notice that both criterion (a) and the generation of query terms that literally match words in titles and SHs does not shed light on the added value of search by linguistic variations used by DOCDIGGER in the default mode.

To evaluate the translation functionality of MUSIL we have adapted the (a) criterion to this mode. We have started our analysis from English query terms to retrieve German documents too. FUB librarians have mapped the 224 English query terms, mentioned above, into the set of German SHs assigned to the records contained in the sample DB: 75 English terms didn't have a counterpart into this subset of SWD and 16 have been mapped to more than one term. For each English query term having a mapped German term, we considered as relevant those German documents containing the mapped term in the SHs. For all English query terms the German terms proposed by the librarians were among the translated terms used by MUSIL. Hence, the relevant books were all found. We still have to measure the precision of the search, but it's already clear that too many non relevant books are also found. Similar results are obtained for the Expand mode.

[5]In the case of the FUB Library, these terms are only available in the German controlled vocabulary (SWD).

4. Conclusions and Further work

We have presented ongoing work on the MUSIL project, in which the traditional OPAC of the Free University of Bozen-Bolzano is extended with multilingual search capabilities. We are currently working on several extensions of the system, on the one hand aimed at improving search results, and on the other hand at offering additional functionalities to the user through an improved interface. As for improving precision of search results, our preliminary laboratory test has shown the need of using filters to better control the translation and expansion functionalities and suggests that Controlled Vocabularies of SHs can provide us with these filters. To avoid that such a filtering is too selective, linguistic variations of the terms should also be considered. A further improvement will be the treatment of proper names, in order to distinguish those cases where a proper name that is ambiguous with a common word should not be translated across languages from those where the translation is required. A similar special treatment in translation and thesaurus expansion is necessary for multi words and compound names (especially important in German).

Furthermore, different definitions of relevance should be explored and analyzed and their results should be compared so as to reach a clearer picture of MUSIL performance. These experiments will be useful also to calibrate the ranking that is now based on a vector space model. This model gives good results in the default mode but should be adjusted for the results of the translate and expand functionalities.

5. References

L. Dini, D. Liebwald, L. Mommers, W. Peters, E. Schweighofer, and W. Voermans. 2005. Cross-lingual legal information retrieval using a wordnet architecture. In *proceedings of ICAIL '05*, pages 163–167.

Paolo Dongilli, Enrico Franconi, and Sergio Tessaris. 2004. Semantics driven support for query formulation. In *Proceedings of the 2004 International Workshop on Description Logics (DL 2004)*. CEUR Electronic Workshop Proceedings, `http://ceur-ws.org/Vol-104/`.

Patrice Landry. 2004. Multilingual subject access: The linking approach of MACS. *Cataloging & Classification Quarterly*, 37(3/4):177–191.

Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors. 2003. *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum (CLEF 2002)*, volume 2785 of *Lecture Notes in Computer Science*. Springer.

E. Stamatatos S. Michos and N. Fakotakis. 1999. Supporting multilinguaglity in library automation systems. *Applied Artificial Intelligence*, 13(7):679–704.

Jean Tague-Sutcliffe. 1997. The pragmatics of information retrieval experimentation, revised. In Karen Spaarck Jones and Peter Willett, editors, *Readings in Information Retrieval*, pages 205–216. Morgan Kaufmann.

Charlotte Wien. 2000. Sample sizes and composition: Their effect on recall and precision in ir experiments with opacs. *Cataloging & Classification Quarterly*, 29(4).