

Multilingual Access to Library Catalogues

Barbara Plank
bplank@unibz.it

Students Symposium, LCT Colloquia
Free University of Bolzano-Bozen

October 26, 2006

Outline

- 1 Introduction and Problem Definition
- 2 The MuSiL system
 - The main tasks of the project
 - Language Functionalities of MuSiL
 - Excursion: Related terms
- 3 Conclusion and future work

Outline

- 1 Introduction and Problem Definition
- 2 The MuSiL system
 - The main tasks of the project
 - Language Functionalities of MuSiL
 - Excursion: Related terms
- 3 Conclusion and future work

Introduction (1/3)

- Free University of Bolzano-Bozen (FUB) is a multilingual university (three official languages), offering several international study programmes (e.g. EM LCT) :-)
- The FUB library:
 - multilingual users
 - books in different languages
- However, the Online Public Access Catalogue (OPAC) of the FUB library lacks multilingual features and has some intrinsic limitations

Introduction (2/3)

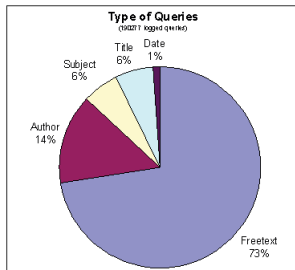
Current limited search capabilities of OPAC

- No multilingual search
- OPAC search bases on just "string matching": e.g. search term "probability":
where [...] stm1.stichwort like 'probability'
 - no linguistic information used
 - books about "probabilities" will not be retrieved, unless the user searches using truncation: "probabilit%" (user often not aware of it)
- Limited to at most 3 search terms

Introduction (3/3)

Study of OPAC logs

- Queries are "duplicated", repeated in 2-3 languages
- 27 % of queries are field search (title, author, subject), 73 % free text search



The study reveals obvious problems

- User has additional effort in translating manually
- User may not find possible relevant books (written or catalogued in a language different from the language of the query terms)

Background on Cataloging Systems

Definition

A **Subject Heading** is a linguistic expression (word or group of words) representing the subject content of a document and used for retrieval in a catalogue, bibliography or index. [1]

Cataloging in the Library of Bolzano

The Library of the FUB uses **language specific subject headings systems** for cataloging the relative bibliographic items, motivated by the fact that bibliographic information is gathered from other national libraries:

- German *Schlagwortnormdatei* (SWD),
- English *Library of Congress Subject Headings* (LCSH),
- Italian *Soggettario Italiano* (SI).

The problem of Multilingual Search (1/2)

A solution was necessary in order to extend OPAC's basic search functions with a multilingual search.

Possible solution approaches

- Exploit complex mappings between SH systems of various languages [2]
- Use Cross-lingual Information Retrieval (CLIR)

We are interested in the latter approach.

The problem of Multilingual Search (2/2)

Cross-lingual Information Retrieval

Can we apply standard CLIR methods to a Library Catalogue to overcome the problem of cross-lingual access?

Possible problems:

- The approach is statistically based and suffers when information is scarce
 - However, this goes into the direction of digital libraries (more text, better performance)
- Standard thesauri in IR created for other purposes
 - Specific thesauri for the library domain.

Outline

1 Introduction and Problem Definition

2 The MuSiL system

- The main tasks of the project
- Language Functionalities of MuSiL
- Excursion: Related terms

3 Conclusion and future work

Introduction to the MuSiL system

The MuSiL project

- **Multilingual Search in Libraries (MuSiL)**
- On-going project on the enhancement of an OPAC search system with multilingual access
- Collaboration between FUB Library (Elisabeth Frasnelli), Faculty of Computer Science KRDB (Raffaella Bernardi, Diego Calvanese, Barbara Plank) and CELI™, Turin (Luca Dini, Paolo Curtoni, Vittorio di Tomaso)

Aim of the project

- To integrate advanced linguistic technologies in a user friendly interface and bridge the gap between the world of free text search and the world of conceptual librarian search

The two main tasks of the project are:

- 1 Integrating Cross-Language Information Retrieval (CLIR) into the library
- 2 Adapting the search engine to the library's needs

1. Integrating CLIR into the library (1/2)

- Integration of the multilingual IR engine DOCDIGGER of CELI into OPAC
- Keep existing OPAC functionalities (e.g. library account management) and extend OPAC's search functionalities with ML search

1. Integrating CLIR into the library (2/2)

Looking at the big picture

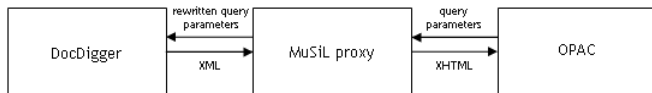


Figure: The three components of the MuSiL system

The Integration

- Extending the Graphical User Interface of OPAC with new features
- Using DocDigger as underlying search engine
- MuSiL proxy to communicate between the two systems

Technologies used: Java, Jakarta Struts, XML, XSLT

2. Adapting the search engine to the library's needs

- Combine **general and library specific thesauri** (cascading approach) and **filters** to better control the expansion functionalities
- 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.
- Detection of **multiwords** to handle translation and expansion
- **Information sparseness**: include further information (abstracts, table of contents)

More details on DOCDIGGER and the library

DOCDIGGER

- An index was created over the whole textual content of the database of the library (title, subtitle, author, SHs, abstracts)
- External information (PDFs with table of contents) were also indexed, whenever available
- Depending on the languages (DE-EN-IT) different lemmatizers have been used.

Language functionalities (1/2)

The search engine is based on advanced linguistic technologies in order to search for:

- **Linguistic variations** of the search term (e.g. orario, orari, ora).
- **Translation** of the query into Italian, German and English.
The system looks for the search term's translations by using a multilingual dictionary.
- **Related terms** i.e. terms semantically related to the query.
The system looks for the search term and conceptually related terms by using a multilingual thesaurus only in documents of the same language of the search terms; We will look into this in more detail;

Language functionalities (2/2)

The system guesses when you have inserted as search term:

- **Proper name** (e.g. Massimo Rossi) and warns you that it has neither translated nor expanded it.
- **Multiword** (e.g. post office) If it has detected a multiword, it will translate it as a whole instead of translating it word-by-word.

Another feature of the system is it's ability to **search in volumes** of encyclopedias, journals.

Excursion: Related terms (1/3)

In the beginning, a **general purpose thesaurus** was used to retrieve terms semantically related to the query:

- Structured configurable thesaurus, property of CELI, based on WordNet relations:

```
<ExpandRel>  
<Relation type="has_hyponym" enabled="true" />  
<Relation type="has_hyperonym" enabledFirstLevel="true" ...  
<Relation type="has_holonym" enabled="true" />  
<Relation type="has_meronym" enabled="true" />  
<Relation type="be_in_state" enabled="true" />  
..
```

A preliminary laboratory study on evaluating the precision of the system showed that MuSiL reached a precision of about 65 % (for German) and around 49 % (for Italian and English).

Clearly, too many non relevant books are also found.

Excursion: Related terms (2/3)

First idea: Filter the content of the general purpose thesaurus

By using the corpus of documents a filter was created that

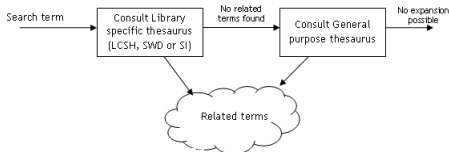
- removed terms from the thesaurus not available in the corpus
- removed terms that are too far away (threshold)

What is the "right" threshold?

Excursion: Related terms (3/3)

Second Idea: Use library specific thesauri (SH systems) combined with general purpose thesauri in a cascading approach.

Search for related terms



- + Exploit domain specific thesauri whenever possible (increase in precision)
- Search term may be not found in the library specific thesauri (user often not aware of SHs)
- + Hence, still providing filtered general thesaurus approach in a second step

Todo: We have to evaluate the filtering of the general thesaurus and the cascading approach to query expansion.

The MU SiL system interface

The old system and the new system

OPAC
(old system)

[Library Home] Deutsch - Italiano

Keyword








Catalogue: Unibz [MetaOPAC](#) || [Single OP](#)

Search completed by: Keyword: "probability"

Hits: Unibz: 73

BRIEF DESCRIPTION Hits 73

Page 1 of 11 [01 02 03 04 05 06 07 08 09 10 11]

-  B. Statistics and probability. - 2003. - 344 S. : graph. Darst. - 2003
13 - ZG 9260 S783 -3
-  B. A treatise on probability. - 1988. - XXII, 514 S. - 1988
13 - QC 072 K44 -9
-  A modern introduction to probability and statistics : understanding why
and how / F. M. Dekking ... (Springer texts in statistics) - 2005
-  Advances on methodological and applied aspects of probability and
statistics / ed. by N. Balakrishnan - 2002
13 - QH 230 B171
-  Agresti, Alan: An introduction to categorical data analysis / Alan Agresti -
10. print. (Wiley series in probability and sta...) - 1996
13 - SK 830 A277
-  Agresti, Alan: Analysis of ordinal categorical data / Alan Agresti - 14.
print. (Wiley series in probability and mat...) - 1984
13 - SK 830 A277
-  Agresti, Alan: Categorical data analysis / Alan Agresti - 2. ed.,... (Wiley
series in probability and sta...) - 2002
13 - SK 830 A277(2.02)

MuSiL
(new system)

[MuSiL Home] Deutsch - Italiano

Search term

Search term language

Search with Translate Related terms

Catalogue: Unibz

Search term: "probability"
>> [What has been searched? -- Explanations](#)

Total documents: 240 - English: 69 - German: 9 - Italian: 160 - Various: 2

Document's language: English - 69 documents
Page: 1 | > | >>

-  Author and title: Ross, Sheldon M.: Introduction to **probability** models /
Sheldon M. Ross - 8. ed.
Publisher: Academic Press Year: 2003
Relevance: 100% ★★★★★

Document's language: German - 9 documents
Page: 1 | > | >>

-  Author and title: Weichselberger, Kurt : Elementare Grundbegriffe einer
allgemeineren Wahrscheinlichkeitsrechnung / Kurt Weichselberger. Unter
Mitarb. von Thomas Augustin und Anton Wallner. [search "probability" in
the volumes](#)
Publisher: Physica-Verl. Year: 2006
Relevance: 82% ★★★

Document's language: Italian - 160 documents
Page: 1 | > | >>

-  Author and title: Agresti, Alan: An introduction to categorical data
analysis / Alan Agresti - 10. print. (Wiley series in **probability** and
sta...)

Outline

- 1 Introduction and Problem Definition
- 2 The MuSiL system
 - The main tasks of the project
 - Language Functionalities of MuSiL
 - Excursion: Related terms
- 3 Conclusion and future work

Conclusion and future work (1/2)

- Library systems can take advantage of CLIR technologies.
- We have developed jointly with the library and CELI team a bridge that connects their two system for enabling Multilingual Search.
- We are currently planning an evaluation with real users, also to evaluate the newly incorporated features (treatment of proper names, multiwords and the cascading approach to related terms)

Conclusion and future work (2/2)

- Future work consists in the full integration of the Compounder for German (see next talk)
- I did not speak about the ranking of documents, a general TF*IDF measure was applied. We have to find out if it is suitable for the domain of a library, or should be adapted. A difficulty lies in determining the relevance criteria: Given a user query, what are the relevant books? Titles not always indicative for the topic.
- Plan: Go online in December 2006

Thank you.

Bibliography



K. G. Saur

Principles Underlying Subject Heading Languages (SHLs).
UBCIM Publications - New Series, Vol. 21., München, 1999.



Patrice Landry.

Multilingual subject access: The linking approach of MACS.
Cataloging & Classification Quarterly, 37(3/4):177191. 2004.