Harvesting Translingual Vocabulary Mappings for Multilingual Digital Libraries

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@sherlock.berkeley.edu

Fredric Gey
UC Data Archive & Technical Assistance (UC DATA)
University of California, Berkeley
Berkeley, California, USA, 94720
gey@ucdata.berkeley.edu

Aitao Chen
School of Information Management and Systems
University of California, Berkeley
Berkeley, California, USA, 94720-4600
aitao@sims.berkeley.edu

ABSTRACT

This paper presents a method of information harvesting and consolidation to support the multilingual information requirements for cross-language information retrieval within digital library systems. We describe a way to create both customized bilingual dictionaries and multilingual query mappings from a source language to many target languages. We will describe a multilingual conceptual mapping resource with broad coverage (over 100 written languages can be supported) that is truly multilingual as opposed to bilingual parings usually derived from machine translation. This resource is derived from the 10+ million title online library catalog of the University of California. It is created statistically via maximum likelihood associations from word and phrases in book titles of many languages to human assigned subject headings in English. The 150,000 subject headings can form interlingua mappings between pairs of languages or from one language to several languages. While our current demonstration prototype maps between ten languages (English, Arabic, Chinese, French, German, Italian, Japanese, Portuguese, Russian, Spanish), extensions to additional languages are straightforward. We also describe how this resource is being expanded for languages where linguistic coverage is limited in our initial database, by automatically harvesting new information from international online library catalogs using the Z39.50 networked library search protocol.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.7 [**Digital Libraries**]: Systems issues; H.5.2 [**User Interfaces**]: Natural Language

General Terms

Algorithms, Design

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL '02 July 13-17, 2002, Portland, Oregon USA.
Copyright 2002 ACM 1-58113-513-0/02/0007 ...$5.00.

Keywords

Cross-Lingual Information Retrieval, Controlled Vocabularies, Entry Vocabulary Indexes

1. INTRODUCTION

As digital libraries expand in scope and content to include resources in a variety of languages from international sources, there is an increasing need for multilingual information access to those resources. Research programs, such as those sponsored by the DARPA TIDES (The Translingual Information Detection, Extraction and Summarization program), have been developing new methods to accelerate information interchange across language barriers. Although automatic machine translation between pairs of languages is moderately well-developed between English and the world's major languages (including Chinese, French, German, Italian, Japanese, Portuguese and Spanish) the same facilities do not exist between other pairs of these languages (for example between German and Japanese) and are even more rare when dealing with less common languages such as those of the Indian Subcontinent.

The resource requirements for commercial quality machine translation are significant – it has been estimated that a high quality general bilingual dictionary of at least 250,000 words is a minimal resource for a good machine translation system. The emerging field of statistical machine translation [3] utilizes entirely different resources – parallel corpora which can train statistical decoding algorithms for automatic transfer between languages. The major parallel corpora which have been utilized thus far have come from political bodies in developed countries which have a requirement to use humans to translate between official languages of the body, vis the Canadian Hansards (English, French) [11] and the official documents of the United Nations (Arabic [16], Chinese, English, French and Russian). When one steps outside these languages, parallel resources are difficult, if not impossible, to obtain. Even with these languages, one is faced with the development of sentence alignment algorithms which account for variations in sentence length, word order and word length between the two languages.

More recently the vast information content of the WWW has been looked to as a source for parallel corpora. Web pages in German or Japanese, for example, may have analogous pages in English on the same site which have been translated from the original language page. Algorithms and software must be developed to mine these web pages and to extract parallel text fragments (sentences, paragraphs, documents) which can serve the same purposes as the parallel reservoirs of literature published by socio-political

entities [15, 12]. Web resources of this type are less likely to be developed with the same degree of attention to detail as translations done by professional translators for official government purposes. That is to say, from a statistical point of view, they are noisy channels. In addition, if the desire is to go beyond the world's mainstream languages to, say Arabic, Farsi, or any Indian subcontinent language, one finds extremely limited parallel resources and is faced with an anarchy of character sets and font representations used by web sites.

Finally, cross-language information retrieval cannot, in general, be directly performed across multiple groups of languages. In the multilingual retrieval evaluation of CLEF (Cross-Language Evaluation Forum [13, 14]), access from non-English queries to other non-English documents is usually performed by translating twice, from query source language to English and then from English to target languages. Thus, English is used as an intermediate 'pivot language' (Dutch has also been used in this fashion in CLEF [7].)

2. USING ONLINE LIBRARY CATALOGS AS TRANSLINGUAL VOCABULARY RESOURCES

Our search for a new source of multilingual resources derives from a background of library research. The world's great research and university libraries have book content which spans the world's languages. For example in the University of California's electronic catalog MELVYL[1] nearly half its 13 million title collection is non-English. As an example we might submit the query "find subject Islamic Fundamentalism AND Language Arabic" and obtain screens of results as in Figure 1.

What does this imply? That a single query to this library catalog yields 130 Arabic language samples coded with the topic "Islamic Fundamentalism". The implications for multilingual information access are enormous. Indeed, if we further submit another query "Find subject Islamic Fundamentalism AND Language not (English or Arabic)" we would obtain another 55 book records covering nine additional languages (Bengali 9 book records, French 13, German 9, Hebrew 3, Indonesian 3, Malay 2, Russian 4, Turkish 4 and Urdu 2). The University of California library catalog is but one of more than 1000 remotely searchable library catalogs worldwide. Many (if not most) of these catalogs are searchable using the international standard search and retrieval protocol Z39.50 [1]. For example, COPAC, the library catalog of the United Kingdom and Ireland's academic libraries[2] contains over 9 million records with 44,321 in Arabic. We describe the method for "harvesting" such resources in a later section.

There are other advantages to exploiting library catalogs for multilingual tasks. Library catalog databases are structured according to international standards for metadata, such as the MARC format which has been in use for more than 30 years (as opposed to Web pages which possess no standard format). The data are tagged with rich metadata (while Web pages have limited or non-existent metadata). The data content is identified according to well-defined rules and controlled vocabularies such as AACR2 (Anglo-American Cataloging Rules, 2nd Edition) and the Library of Congress Subject Headings (LCSH). In contrast, Web pages rarely have their content identified. Moreover, while electronic online library catalogs are limited in size (for example 13 million items at the University of California) versus the billions of web pages, they may contain the only extant resources in specialized languages. For all the attention paid to mining web pages for parallel texts, they can glean few resources outside the mainstream languages – for example parallel web pages of Arabic and English are currently almost non-existent, partly because there was, until recently, no standard character representation in use for Arabic on the WWW.

[1] http://www.dbs.cdlib.org/?CSdb=cat

[2] http://www.copac.ac.uk/copac

3. ENTRY VOCABULARY INDEXES TO MAP LIBRARY CATALOG DATA

For the past several years, our research group has been developing what we term Entry Vocabulary Indexes (EVIs). Entry Vocabulary Indexes provide statistical mappings between words and phrases in documents and subject categories or topical classifications which have been assigned by humans to the documents. By creating these statistical associations, user vocabulary ('entry vocabulary') can be mapped to the controlled vocabulary terms assigned by human indexers to characterize the document content.

The basic method is founded on work done a decade ago with library classification [9, 8]. This method relies on four elements:

1. A training set of documents that have been indexed using the vocabulary. This training set must be of sufficient size to provide adequate statistical correlation between controlled vocabulary words and words in the text of the documents.

2. NLP methods such as part of speech taggers, dictionary lookups, etc. are used, whenever possible, to identify noun phrases in the language of the text. If not available for a given language, individual words (or segmented sets of characters for languages without orthographic separation of words) are used instead.

3. Software and algorithms to generate a probabilistic mapping between the words and phrases extracted from documents and the controlled vocabulary used in the collection.

4. Software to provide search capabilities for the generated mappings. This software takes words or phrases in natural languages and, using the mappings, produces a ranked list of the most highly associated terms in the controlled vocabulary.

To obtain the collection of documents for the first element usually requires that a large database be acquired, or that network accessible resources be "mined" to obtain large sets of appropriate records. In a later section we discuss how the Z39.50 protocol is used to derive such samples from online library catalogs.

The mapping currently used for our Entry Vocabulary Indexes is based on a maximum likelihood weighting associated with each term (word or phrase) and each classification. In effect we construct two-way contingency table for each pair of word/phrase terms t and classifications C as shown in table 1, where a is the number of document titles/abstracts containing the word or phrase and classified by the classification; b is the number of document titles/abstracts containing the word or phrase but not the classified by the classification; c is the number of titles/abstracts not containing the word or phrase but is classified by the classification; and d is the number of document titles/abstracts neither containing the word or phrase nor being classified by the classification.

	C	$\neg C$
t	a	b
$\neg t$	c	d

Table 1: Contingency table from words/phrases to classification

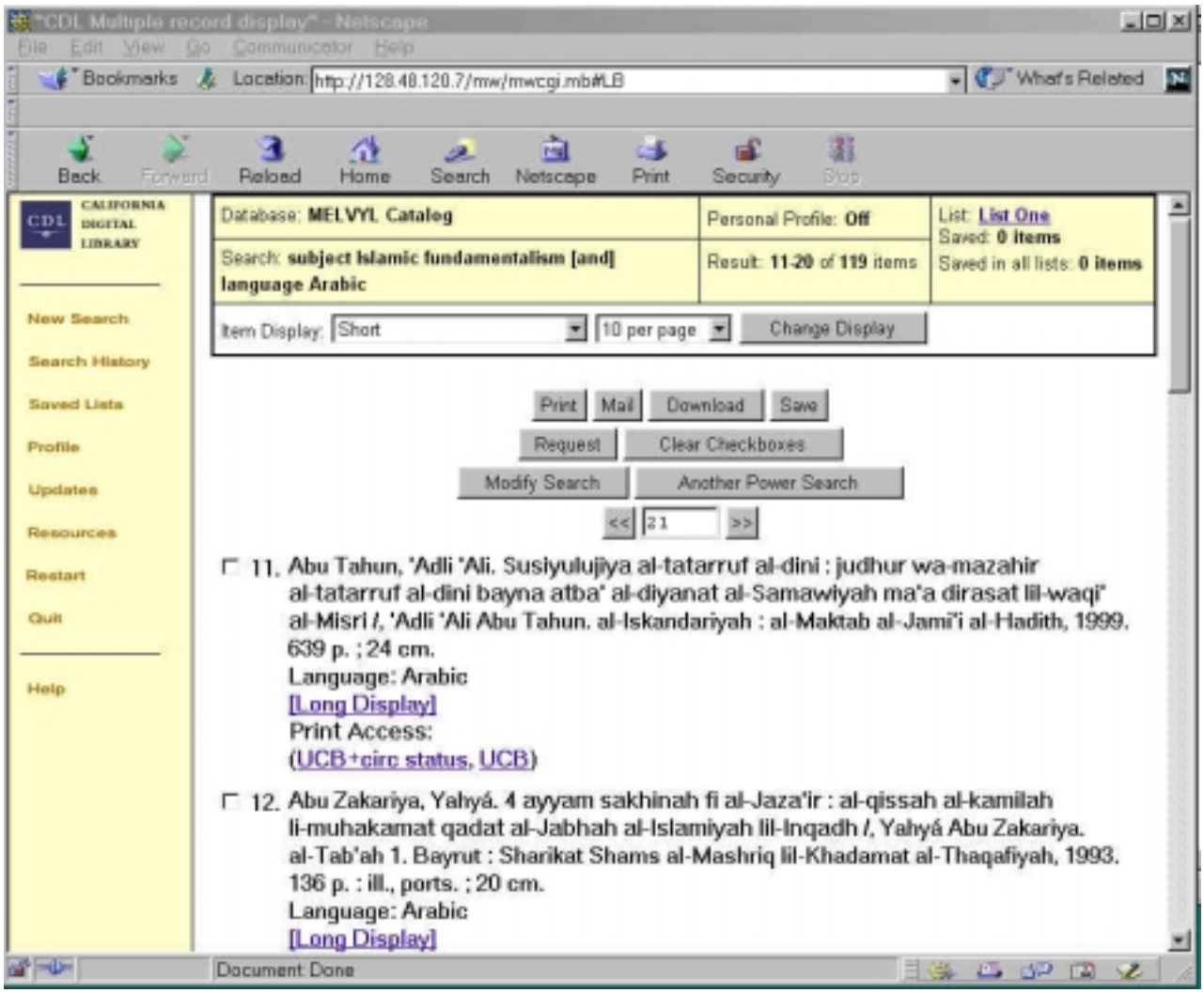

Figure 1: Library Catalog Query "Islamic Fundamentalism AND Language Arabic"

The association score, $W(C,t)$, between a word/phrase t and an classification C, is computed following Dunning [5].

$$W(C,t) = 2[logL(p_1,a,a+b) + logL(p_2,c,c+d) - \quad (1)$$
$$= logL(p,a,a+b) - logL(p,c,c+d)] \quad (2)$$

where

$$logL(p,n,k) = klog(p) + (n-k)log(1-p) \quad (3)$$

and $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, and $p = \frac{a+c}{a+b+c+d}$.

Further details on the method used may be found in our paper in HLT-2001 [6].

4. PROTOTYPE TRANSLINGUAL VOCABULARY MAPPING SOFTWARE

Thanks to a special arrangement with the California Digital Library (CDL)[3], we obtained a private copy of the University of California's MELVYL catalog database which contains 10,091,737 records, of which 4,626,793 were non-English. We utilized the entry vocabulary index methodology to create mappings from Library of Congress Subject Headings to words found in book titles in the following nine languages: Arabic, Chinese, French, German, Italian, Japanese, Portuguese, Russian and Spanish. Figure 2 shows an example mapping from the LCSH term "Islamic Fundamentalism" to the highest ranked three words in these languages.

The message "not in the training set" means that for there were no books on that subject in that language.

This example shows a single use of the resource. We have also developed the reverse mappings from non-English languages to Library of Congress Subject Headings. If this feature is used, one can enter a German word such as 'wirtschaftspolitik' and the system will return the LCSH heading 'economic policy', which is an exact English translation of the word. We have attached a keyboard feature which supports the Cyrillic alphabet, so a Russian word such as перевод can be entered and the system will return the subject heading 'Translating and Interpreting'. The query word Übersetzung in a German query would return the same LCSH heading, leading to the possible equivalence of перевод to Übersetzung in a bi-lingual Russian-German lexicon. In this way the assigned LCSH heading functions as an interlingual between the Russian

[3]http://www.cdlib.org/

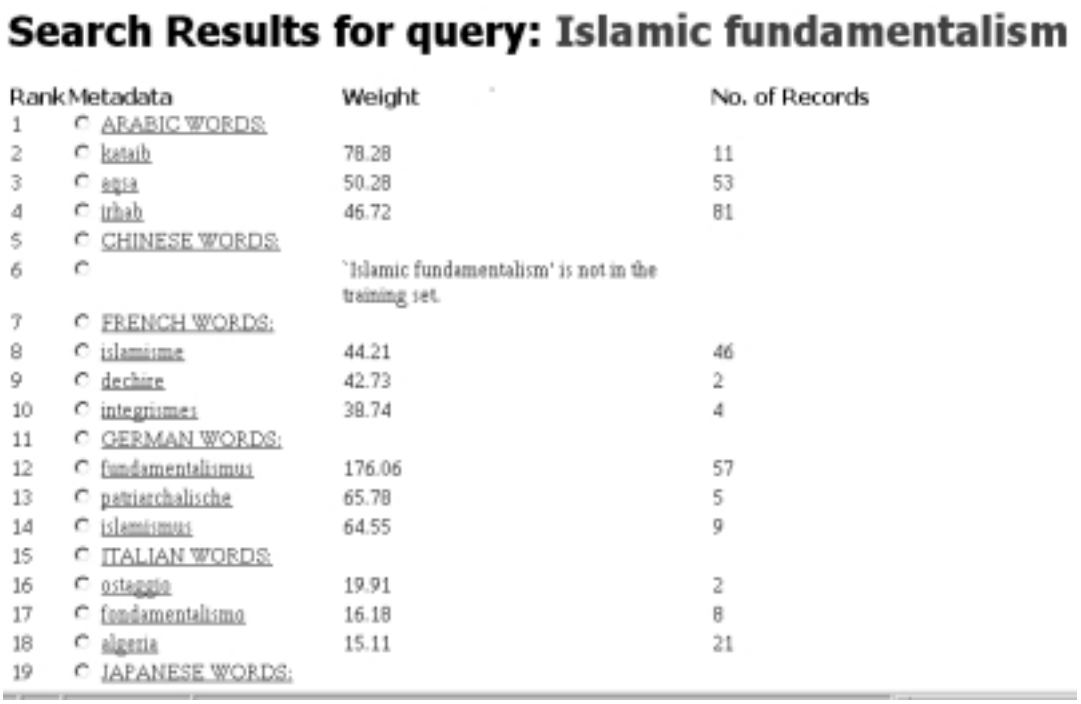

Figure 2: Mapping Subject Heading "Islamic Fundamentalism" to several languages

and German languages.

Language coverage and size of our resource is shown in the following table 2 for languages with more than 20,000 records in the catalog:

Language	N Docs	Language	N Docs
German	840,032	Danish	41,517
Spanish	641,025	Hebrew	41,468
French	609,089	Czech	35,432
Russian	341,050	Urdu	30,206
Italian	266,424	Turkish	30,015
Portuguese	149,389	Bulgarian	27,850
Chinese	127,636	Norwegian	26,478
Japanese	110,956	Korean	25,979
Arabic	96,124	Rumanian	25,874
Dutch	90,170	Finnish	25,027
Latin	88,818	Thai	24,693
Polish	81,698	Serbo-croat	24,601
Indonesian	59,445	Greek	23,926
Swedish	53,854	Bengali	23,430
Hungarian	46,330	Catalan	20.392
Hindi	42,886	Tamil	20,232

Table 2: University of California Catalog's Non-English Language Distribution

In addition, there are 106 additional languages with at least 500 catalog records. Note that, unlike web pages, a significant presence of the Latin language is present. Note also that a number of Eastern European languages are represented, as well as four of the major Indian subcontinent languages (Hindi, Urdu, Bengali and Tamil).

Our prototype translingual vocabulary resource can be found at http://otlet.sims.berkeley.edu/mulevm2.html

5. EXTRACTING LANGUAGE RESOURCES FROM ONLINE LIBRARY CATALOGS

A multilingual resource, such as the one described above, can be developed in two ways: 1) aquiring a large multilingual database, such as the MELVYL database, or 2) incrementally extracting information in the desired languages from multiple online catalog databases. For over two decades research, academic, and public libraries have been moving their catalog information to digital form. One requirement for many of these systems was the ability to use the Z39.50 protocol to receive queries from remote users (often via some other online catalog system). Over this period many public and university libraries that were creating or purchasing an online catalog system required that the system support Z39.50. This has resulted in a large installed base of Z39.50 Servers that provide (usually) MARC bibliographic information of the sort used to create the multilingual resource described earlier in this paper.

The Z39.50 Information retrieval protocol is made up of a number of "facilities", such as *Initialization*, *Search*, and *Retrieval*. Each facility is a logical group of services (or in some cases, a single service) that perform various functions in the interaction between a client (origin) and a server (target). The facilities that we will be concerned with in what follows are the Search Facility, the Retrieval Facility, the Explain Facility, and the Browse Facility.

5.1 The Search Facility

The Z39.50 Search facility permits the client to submit a arbitrarily complex query (usually Boolean, although some systems permit ranked searches as well) to a Z39.50 server in a standardized representation. The server, in turn, returns to the client a standardized representation of the search results (e.g., number of items retrieved, diagnostic or error information, a result set identifier, etc.) and optionally some or all of the matching records, also in a standard record syntax. A basic set of search attributes has been defined in the Z39.50 standard, known as BIB-1. The BIB-1 attribute set has

been based largely on the requirements of bibliographic retrieval for online library catalogs and includes standard representations for search elements such as personal and corporate authors, titles, subjects, date, and language (and many other elements largely derived from the elements of the MARC records used in online catalog systems).

5.2 The Retrieval Facility

The Z39.50 Retrieval Facility allows the client to request some number of records derived from those identified by a Search in a specific record syntax. Although different servers support different record syntaxes, for online library catalogs the various MARC record syntaxes or the SUTRS (Simple Unstructured Text Record Syntax) are most common.

5.3 The Explain Facility

The Z39.50 Explain facility permits the client to obtain information about the server implementation, including information on databases supported, attribute sets used (an attribute set specifies the allowable search fields and semantics for a database), diagnostic or error information, record syntaxes and information on defined subsets of record elements that may be requested from the server (called elementsets). The server (optionally) maintains a database of Explain information about itself and may maintain Explain databases for other servers. The explain database appears to the client as any other database, and uses the Z39.50 Search and Retrieval facilities to query and retrieve information from it. There are specific attributes, search terms and record syntaxes defined in the standard for the Explain database to facilitate interoperability among different server implementations.

5.4 Z39.50 Browse Facility

As the name of this facility implies, it was originally intended to support browsing of the server contents, specifically the items extracted for indexing the databases. The single service in the Browse facility is the Scan service. It is used to scan an ordered list of terms (subject headings, titles, keyword, text terms, etc.) drawn from the database. Most implementations of the Scan service directly access the contents of the indexes on the server and return requested portions of those indexes as an ordered list of terms along with the document frequency for each term.

5.5 Using Z39.50 to Harvest Linguistic Resources

There are several approaches for using Z39.50 to extract records that may be used to build linguistic resources like the multilingual EVM described above. The method used depends on the facilities available on a given Z39.50 server, and on the supported search attributes of the server.

The optimal situation is where the server supports all of the facilities described in the preceding sections. In this case the extraction program can use the explain facility to discover the searchable elements for the server, and may use the scan facility to extract terms from the indexes of the server (see [10] for a discussion of this technique). If the server permits direct search by language, then the complete set of records in a given language can be retrieved from the server and used to build a mapping. If (as with some servers) language can only be used in conjunction with another search element to restrict the resultset to records in that language, then the extraction program may need to use multiple searches to select a topical or other subset of the records in the target language.

Systems that provide this sort of optimal access via Z39.50 include the MELVYL catalog and the COPAC catalog hosted by Manchester Computing in the U.K. In the COPAC catalog, for example, a Z39.50 search for language=arabic returns 44549 records with Arabic titles. Other Z39.50 accessible library catalogs include the catalog of the Library of Congress, as well as servers for the National Libraries of Australia, Poland, and Wales among hundreds of other university and public libraries or consortia of libraries.

The "worst case" scenario for harvesting information to build a multilingual resource is where the server doesn't support search (or limitation of a search) by the language used in the records, and limitations are placed on server such as a limited number of items that may retrieved. Even in this case multiple topical or other searches may be performed and the restriction to the target language may be done on the client side.

For servers that support the OAI (Open Archives Initiative) protocol[4] and do not support Z39.50, it should be possible to use the "ListRecords" verb in OAI to extract all records from a given database, and to then select those in the languages of interest on the client side.

6. ISSUES AND CURRENT DEVELOPMENT

Two major issues remain to be resolved by further development to more completely exploit the potential of this new resource. They are transliteration and back-transliteration for non-European scripted languages, and phrase mapping for non-English languages. You will note from figure 2 that the Arabic words returned in response to the English query are in a Romanized character set, that is they have been transliterated from their original Arabic. In order for them to be useful in search, the words will have to be "back transliterated". Transliteration of the title words has been done by library cataloguers who follow rules set out by the American Library Association and the Library of Congress [2] – however these rules have not, in general, been instantiated in software. Our project will, at least, want to provide "back transliteration" to the original language's alphabet and script in order to make the words recognizable and useful for online search. This latter is problematic for some languages because the prefered ALA/LC transliteration is not reversable without considerable processing (e.g., similar sounding characters or groups of characters in the source language may be mapped to a single phonetically similar transliteration, but it may be very difficult to infer what the original characters were from that transliteration).

Second, the current version of the library language mapping deals with single words only in the target language (although it will search for phrases in English within the Library of Congress Subject Headings). Thus, for example, in figure 2, the Italian word 'fondamentalismo' would properly be replaced/related to the phrase 'Fondamentalismo islamico'. To do this thoroughly would require part-of-speech tagging in all target languages. However, a simple form of statistical phrase identification can be done by examining bigrams (consisting of two adjacent words) using the expected mutual information measure, which computes the deviation from random expectation of finding the two constituent words adjacent in a corpus The authors have previously utilized this technique for statistical segmentation of Chinese text [4].

In addition we are currently undertaking experiments to evaluate our multilingual EVM resource in cross-language information retrieval. A particularly good venue for evaluation is the CLEF collection, utilizing the LCSH to all languages for translation of CLEF queries from English to French, German, Italian and Spanish. Another example of evaluation would be to randomly sample the German collection of titles to simulate less dense languages.

[4]http://www.openarchives.org

For example, to simulate the Tamil language would require a German sample of 20,232 titles, and thus the sampling fraction would be $20232/840032 = 0.0241$. The effectiveness of a sampled vocabulary could then be measured against a full vocabulary in cross-language retrieval evaluations.

We are also testing the use of the multilingual EVMs that we have developed as tools to automatically assign controlled vocabulary to documents based on the mappings described above. For this purpose we are using samples of records and then comparing the actual assignments of subject headings by human indexers to the highly ranked subjects suggested by the the EVM when the titles are submitted as a query. We hope to report on this work at the meeting. Some preliminary results (using English language materials only) indicate that exact prediction (where the desired heading is the top-ranked heading from the EVM) only occurs about 12% of the time, but that the "correct" heading is among the top 10 suggested by the EVM over 40% of the time (using a sample of over 100,000 records that were not used in building the EVM).

7. ACKNOWLEDGMENTS

Development of our trans-lingual vocabulary prototype was supported by research grant number N66001-00-1-8911 (Mar 2000-Feb 2003) from the Defense Advanced Research Projects Agency (DARPA) TIDES (Translingual Information Detection Extraction and Summarization) program. Entry Vocabulary Technology was developed under support by the DARPA Information Management program through Contract N66001-97-8541; AO# F477: "Search Support for Unfamiliar Metadata Vocabularies", and by a grant from the Institute of Museum and Library Services (IMLS). Work on Z39.50 harvesting of information was supported by the National Science Foundation and Joint Information Systems Committee(U.K) under the *International Digital Libraries Program* award #IIS-9975164.

8. ADDITIONAL AUTHORS

Additional author: Michael Buckland (School of Information Management and Systems, University of California, Berkeley, Berkeley, California, USA, 94720-4600, buckland@sims.berkeley.edu).

9. REFERENCES

[1] ANSI/NISO. *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)*. American National Standards Institute (also available from the Library of Congress, Z39.50 Maintenance Agency at http://lcweb.loc.gov/z3950/agency), Washington, D.C., 1995.

[2] R. K. Barry, editor. *ALA-LC romanization tables : transliteration schemes for non-Roman scripts approved by the Library of Congress and the American Library Association*. Library of Congress, Washington, 1997.

[3] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estim ation. *Computational linguistics*, 19:263–312, June 1993.

[4] A. Chen, J. He, L. Xu, F. Gey, and J. Meggs. Chinese text retrieval without using a dictionary. In A. D. N. Nicholas J. Belkin and P. Willett, editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Rese arch and Development in Information Retrieval, Philadelphia*, pages 42–49, 1997.

[5] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, March 1993.

[6] F. Gey, M. Buckland, A. Chen, and R. Larson. Entry vocabulary – a technology to enhance digital search. In *Proceedings of HLT2001, First International Conference on Human Language Technology San Diego*, pages 91–95, March 2001.

[7] D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld. Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *Cross Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal*, pages 102–115. Springer, 2001.

[8] R. R. Larson. Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarterly*, 61(2):133–173, 1991.

[9] R. R. Larson. Experiments in automatic library of congress classification. *Journal of the American Society for Information Science*, 43(2):130–148, 1992.

[10] R. R. Larson. Distributed resource discovery: Using Z39.50 to build cross-domain information servers. In *JCDL '01, June 24-28, 2001, Roanoke, Virginia.*, pages 52–53. ACM, 2001.

[11] M. Littman, S. Dumais, and T. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette, editor, *Cross Language Information Retrieval*, pages 51–62. Kluwer, 1998.

[12] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 74–81. ACM, 1999.

[13] C. Peters, editor. *Cross Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop*. Springer Computer Scinece Series LNCS 2069, 2001.

[14] C. Peters, editor. *Working Notes of the CLEF 2001 Workshop 3 September, Darmstadt, Germany*. DELOS Network of Excellence on Digital Libraries Workshop Serise, September 2001.

[15] P. Resnik. Parallel stands: A preliminary investigation into mining parallel text from the web for bilingual text. *AMTA*, 1998.

[16] J. Xu, A. Frazier, and R. Weischedel. Trec 2001 cross-lingual retrieval at bbn. In E. Voorhees and D. K. Harman, editors, *Notebook Proceedings of the TREC 2001 Conference*, pages 122–131, November 2001.