An Investigation of Effective and Efficient Multilingual Information Access to Digital Collections

Jiangping Chen
University of North Texas
1155 Union circle #311068
Denton, Texas 76203, USA
Jiangping.Chen@unt.edu

Ryan Knudson
University of North Texas
1155 Union circle #311068
Denton, Texas 76203, USA
RyanKnudson@yahoo.com

Min Namgoong
University of North Texas
1155 Union circle #311068
Denton, Texas 76203, USA
MinNamgoong@My.unt.edu

ABSTRACT
This paper presents the background, research design, and current progress of a new project on exploring the application of various machine translation strategies working toward multilingual information access for digital collections.

Keywords Multilingual Information Access, machine translation, digital collections, metadata records.

INTRODUCTION
Digital collections refer to library online catalogs and collections in various digital libraries. Libraries and museums in the U.S. have developed numerous digital collections to preserve scientific, cultural, and heritage materials. However, very few of them support multilingual information access (MLIA), which enables users to search, browse, recognize and use information in languages other than English (Budzise-Weaver, Chen, & Mitchell, 2012). The advancement of machine translation (MT) technologies provides an alternative for libraries and museums wishing to implement MLIA for their digital collections.

This study aims to determine the effectiveness and efficiency of applying MT technologies to enable MLIA for digital collections. It is a continuation of the MRT Project (http://txcdk-iaa.unt.edu/MRT/), which evaluated the performance of MT on metadata records using measures such as adequacy and fluency (Chen, Azogu, & Zhao, 2012). The MRT project demonstrated that MT could generate understandable Chinese and Spanish translations of English metadata records (Chen, Azogu, & Knudson, 2014). However, it did not assess MT performance in the context of cross-language information retrieval (CLIR).

RELATED LITERATURE
Multilingual Information Access is the extension of Cross-Language Information Retrieval (CLIR), a subfield of traditional Information Retrieval (IR). CLIR provides users with access to information that is in a different language from the language used to express queries. Research in CLIR has been significantly advanced by three major evaluation forums: a Cross-Language Information Retrieval Track at TREC (Text Retrieval Conference: http://trec.nist.gov/) from 1997-2002; the Cross-Language Evaluation Forum (CLEF: http://clef-campaign.org/) which evaluates many European languages; and the NTCIR Asian Language Evaluation (http://research.nii.ac.jp/ntcir/) that covers Chinese, Japanese, and Korean.

Automatic translation of the queries or documents is a necessary step in CLIR. Various translation strategies for translating queries have been explored (Sakai, Kando, et al, 2008; Oard, He, & Wang, 2008). Researchers examined integrating MT with other CLIR techniques such as relevance feedback and data fusion to boost the performance of query-based CLIR (Wu & He, 2012). However, it is unclear whether the MT-based MLIA could yield satisfactory results for digital collection users. Furthermore, there are few investigations of models that digital library developers can follow to implement MLIA.

CLIR for web search has been available since 2005 when Yahoo! launched a CLIR search interface option for German and French sites (Sterling, 2007). In May 2007, Google launched a “Translated Search” feature as part of its Google Language Tools. These services were welcomed by Web users because they provided access to information written in foreign languages (Chen & Bao, 2009). However, CLIR and MT technologies and services, such as those provided by Google, have not been applied to enable searching in specific digital collections in the U.S.

METHODOLOGY
The specific objectives of this project are to:
• Evaluate an MLIA service model based on MT for digital collections;
• Develop an open-source multi-engine MT (MEMT) strategy that can be used to translate metadata records;
• Develop a multilingual (English, Chinese, and Spanish) corpus of metadata records for MT and CLIR research; and
• Evaluate the effectiveness and efficiency of the MLIA service model to determine which MT strategy achieves the best MLIA performance.

The following research question will be answered: What is the effectiveness and efficiency of applying MT to metadata records for MLIA services?

The MLIA Model
To answer the research question, we will employ the MLIA model in Figure 1 for our investigation: Metadata records as documents will be translated into target languages (Chinese and Spanish) by different MT systems prior to search. Then information retrieval experiments will be conducted with queries written in these two languages. We will evaluate the performance of different MT systems and analyze methods of MT regarding retrieval of a given set of queries from a collection of English metadata records.

In the third phase, we are developing a multilingual corpus generator (MLCG) and will use it to create a multilingual corpus of 5,000 metadata records for the purpose of training our MT system. MLCG (http://txcdk.iaa.unt.edu/MLCG/) is adapted from the HeMT system, which was used in the MRT Project for evaluating metadata records translation (Chen, Azogu, & Zhao, 2012). Two or more Chinese and Spanish translators will be hired to use MLCG to generate the multilingual corpus through a post-editing approach.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Test Data Preparation</td>
<td>Acquire one million metadata records from two digital collections; Data pre-processing; Conduct MT with 2 selected systems;</td>
</tr>
<tr>
<td>2. Retrieval Environment Construction</td>
<td>Select and install retrieval system; customize the IR system; upload metadata records</td>
</tr>
<tr>
<td>3. Multilingual Corpus Generation</td>
<td>Select representative metadata records; Develop generation system; Generate multilingual corpus in 3 languages.</td>
</tr>
<tr>
<td>4. Multi-engine Machine Translation (MEMT)</td>
<td>Use Moses – an open-source MT system to implement 3 MEMT approaches</td>
</tr>
<tr>
<td>5. Cross-language information retrieval (CLIR) experiments</td>
<td>Develop multilingual test topics and relevance judgment; Conduct CLIR experiments to test retrieval performance of MT systems including the MEMT systems.</td>
</tr>
<tr>
<td>6. Evaluation</td>
<td>Analyze the results of the CLIR experiments and answer the research question.</td>
</tr>
</tbody>
</table>

Table 1. The Project Phases

The core of this study is to identify an MT approach that is effective and efficient. From our previous study in the MRT project, we found that combining multiple MT results could achieve better performance on metadata records translation than any single system. In this project, we will conduct CLIR experiments to corroborate that finding. In phase 5, we plan to conduct six CLIR experiments, as illustrated in Table 2.

For all the CLIR experiments, we will develop a set of test queries and their relevance judgments for our 1 Million test metadata records. We plan to develop 30 test topics that represent realistic needs of users for the said test metadata records. The development will reference possible resources such as TREC test queries, query logs from the University of North Texas catalog (http://iii.library.unt.edu/), as well as a literature review on characteristics of user queries.

For each test topic, in addition to the English baseline, five experiments as listed in Table 2 (with IDs 2 to 6) will be conducted for both target languages. We will hire participants to create queries using the topics, search the
retrieval system in Chinese or Spanish, and assess the retrieval results for up to 10 test topics. Each topic will be tested by at least two participants in each target language.

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English Baseline</td>
<td>Perform English monolingual retrieval as baseline</td>
</tr>
<tr>
<td>2</td>
<td>Online System A</td>
<td>Perform CLIR using online MT system A</td>
</tr>
<tr>
<td>3</td>
<td>Online System B</td>
<td>Perform CLIR using online MT system B</td>
</tr>
<tr>
<td>4</td>
<td>MEMT1 - A+B</td>
<td>Perform CLIR using the result from the first MEMT approach, which combines in Moses the MT results from both A and B</td>
</tr>
<tr>
<td>5</td>
<td>MEMT2 - A+B+MLCorpus</td>
<td>Perform CLIR using the results of the 2nd MEMT approach, which adds the multilingual corpus generated in Phase 3 to MEMT1</td>
</tr>
<tr>
<td>6</td>
<td>MEMT3 - A+B+MLCorpus +MonoCorpus</td>
<td>Perform CLIR using the results of the 3rd MEMT approach, which adds Chinese and Spanish metadata records to MEMT2 to train the MT system</td>
</tr>
</tbody>
</table>

Table 2. CLIR Experiments

We will be able to build a relevance judgment file for the test topics applying a TREC-like method, i.e., assemble relevant records from all the runs, and then use the relevance judgments to calculate the precision and recall (Voorhees, 2000) of each CLIR run. Next, we will compare each CLIR run with the original English run to understand the percentage of relevant records found in the original English run that also appear in each CLIR run. Also, we will analyze the efficiency and the cost of implementing the MLIA model. The result of the analysis will answer our research question: What is the effectiveness and efficiency of applying MT to metadata records for MLIA services?

SUMMARY

CLIR is a well-studied field in information retrieval (IR). However, its application to digital collections has received little exploration. This study will perform CLIR experiments on a collection extracted from two digital libraries. The purpose is to understand whether machine translation technologies could be applied to metadata records in order to provide multilingual information access services for digital collections. So far we have completed the work in the first two phases: One million metadata records have been extracted and indexed by an open-source IR system. We are currently developing the multilingual platform to generate a multilingual metadata records corpus that will be translated by machine translation (Phase 3).

ACKNOWLEDGMENTS

This project is supported by the Institute of Museum and Library Services (IMLS) under the National Leadership Grant LG-06-13-0187-13.

REFERENCES


