College of Information

UNT UNIVERSITY OF NORTH TEXAS Discover the power of ideas

Multilingual Information Access to Digital Collections

Jiangping Chen
Http://coolt.lis.unt.edu/
Jiangping.chen@unt.edu
April 20, 2016

Self Introduction

- An Associate Professor at Department of Library and Information Sciences in College of Information, University of North Texas.
- I am teaching and conducting research on Intelligent Information Access, Digital Libraries, Database Design, Information Architecture, and Information System Design and Analysis.
- I supervise LIS Graduate Academic Certificate in Digital Content Management and directs the UNT Intelligent Information Access Lab.
- I am the Editor-in-Chief of The Electronic Library(TEL), an international peer-reviewed journal for the application of technology in information environments.
- My Home Page: http://coolt.lis.unt.edu/

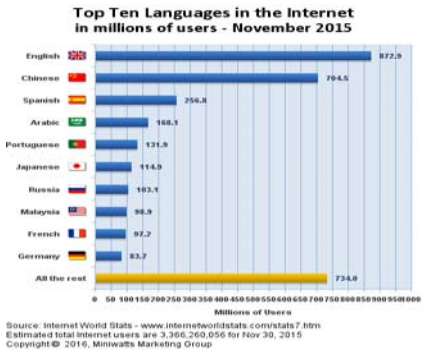
Presentation Outline

- Background
- Multilingual Information Access and Machine Translation
- The MRT Project – Evaluating Machine Translation on Metadata Records
- The MLIA4DC Project – Adding Cross-language Search Service to Digital Collections
- My Book On MLIA
- Summary

Background

- Information creators and users are multilingual
- Internet is multilingual
- The need of the users to access information in many languages
  - For economic development
  - For knowledge sharing/cultural exchange
  - For learning
  - For national security

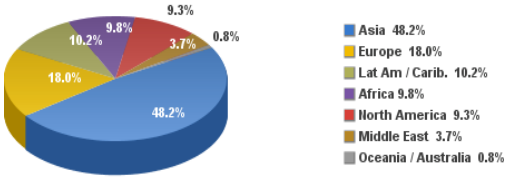
Top Ten Internet Languages



How about Libraries and Librarians?

- You are needed by people who speak different languages that you don't understand!
- The digital information you organized is needed by people in the world who don't understand its language!

Internet Users in the World by Regions November 2015



Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 3,366,261,156 Internet users on November 30, 2015

Your digital libraries are used by people worldwide!

What is Multilingual Information Access (MLIA)?

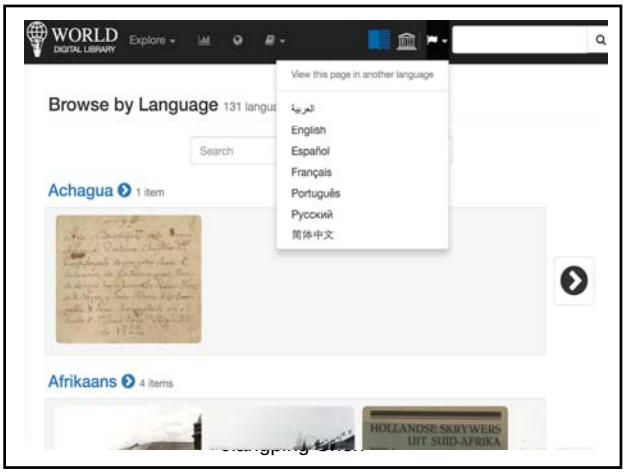
- Solutions to facilitate universal information access by overcoming language barriers
- An extension of Cross-Language Information Retrieval (CLIR)
- Includes but not limited to:
  - CLIR/MLIR: retrieve information in two or more languages
  - CLQA/MLQA – find answers from texts in different languages
  - Bilingual or multilingual browsing and/or result presentation

MLIA is Difficult

- MLIA involves translation
  - Translation is difficult even for professional translators
  - Machine translation is considered the most difficult natural language processing problem
- MLIA includes Information retrieval, which is challenging itself
  - Understanding users' information needs is not always easy
  - Even the most sophisticated IR algorithm cannot satisfy every user

Translation for MLIA

- Translation is necessary to realize MLIA
- We can implement MLIA with manual translation
  - World Digital Library (https://www.wdl.org/en/) presents digital objects in seven UN official languages
  - International Children's Digital Library (http://en.childrenslibrary.org/) collects books in many languages and allows children and their caregivers to search and browse the library in 5 languages
- However, many digital libraries can't afford to apply manual translation!

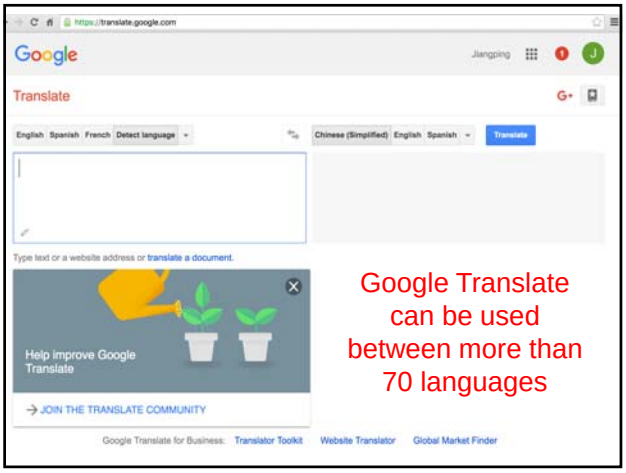


Current U.S. Multilingual Digital Libraries Share the Following Characteristics

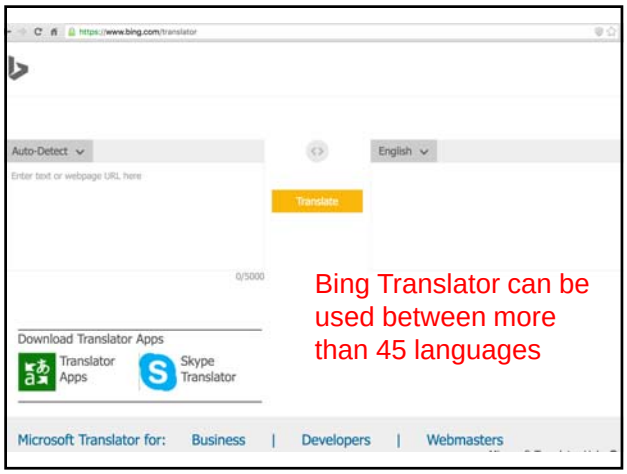
- They have been funded by various sources, especially from the federal government;
- They are the products of collaboration. People from different countries work together to produce multilingual metadata for the digital collections;
- They serve a broader or global user community in which users speak different languages;
- They Do Not employ cross-language information retrieval techniques or machine translation.

MLIA Research

- Fortunately, significant progress has been made in MLIA and Machine Translation (MT)
  - MLIA Evaluation Forums: TREC, NTCIR, CLEF
  - Statistical MT achieved significant progress
  - Big companies, such as Google and Microsoft, provide free online MT services
- Google had launched a cross-language search (2007 – 2011), which built upon many years of research in Machine translation (MT) and Cross-Language Information Retrieval (CLIR)



Google Translate can be used between more than 70 languages



Bing Translator can be used between more than 45 languages

MLIA to Digital Collections (1)

- Digital library users are multilingual
- Many digital collections are multilingual
- MLIA research has been conducted for years, but real applications of the research achievements to digital libraries are rare
- MT systems are producing promising results

MLIA to Digital Collections (2)

- We need to explore the following questions:
  - How effective are current MT technologies for translating digital metadata records?
  - How useful are current Cross-language Information Retrieval (CLIR) and MT technologies for digital libraries?
  - Can information professionals build MLIA services automatically for digital collections based on current CLIR and MT technologies?

The Metadata Records Translation (MRT) Project

- A two-year research project funded by IMLS (Institute of Museum and Library Services (http://www.imls.gov/) and UNT
- A collaborative project among four units in three countries
- The goals of MRT project
  1. to understand to what extent current MT technologies generate adequate translation for metadata records
  2. to explore effective metadata records translation strategies for digital collections

MRT Project Activities

1. Metadata records extraction
   – Extracted 2010 records from UNT Library Catalog and the Portal to Texas History
2. Machine translation of the extracted records to Chinese and Spanish using MT services provided by Google, Bing, and Yahoo!
3. Determination of evaluation measures
4. Creating reference translation and recruiting evaluators
5. Design and implement Multi-Engine MT (MEMT) – 3 different MEMT strategies developed
6. Evaluation using an online multilingual platform built by the project team to assess MT performance of all 6 systems/strategies
   – 1st round: evaluate MT performance of the three online MT services
   – 2nd round: evaluate MT performance of the 3 MEMT strategies

The Six Elements of Metadata Records

Element	Definition	Example
Title	Title of the object to be represented	"The Tulia Herald (Tulia, Tex), Vol. 9, No. 28, Ed. 1, Friday, July 12, 1918 …"
Creator	Author, owner, or generator of the object	"O'Bryan, Barnett"
Subject	Terms that describe the subjects of the object	"Business, Economics and Finance - Communications – Newspapers…"
Description	Short summary or abstract of the object	"Weekly newspaper from Tulia, Texas that includes local, state and national news …."
Publisher	Name and/or address of the publisher	"Engleman, J.S."
Coverage	Geographical coverage, or types of objects	"United States - Texas - Swisher County - Tulia ## new-sou"

Evaluation Measures

- Individual MT system evaluation:

Adequacy
Fluency

- Comparative evaluation:

Best
Worst

Adequacy Scale	Principles for Judgment	Fluency Scale	Principles for Judgment
All (5 points)	Completely match the meaning of the reference translations. All parts are correctly translated	**Flawless (5 points)**	Translated text fully conforms to rules of the language and is consistent with the evaluator's use of native language
Most (4 points)	Most parts are correctly translated	**Good (4 points)**	Translated text conforms to rules of language to some extent and is partly consistent with the evaluator's use of native language
Much (3 points)	Half or more is correctly translated, but fewer than Most	**Non-native (3 points)**	Translated text is understandable but not consistent with the evaluator's use of native language
Little (2 points)	Less than half are correctly translated, some important concepts are not correctly translated	**Disfluent (2 points)**	Translated text is barely understandable
None (1 point)	Totally different in meaning from the references	**Incomprehensible (1 point)**	Translated text is totally beyond understanding

Conclusions (1)

1. To what extent do current MT technologies generate adequate translation for metadata records?
   - adequate translations are correct and understandable translations.
   - For Chinese, two of the online MT systems achieved average scores above 3 on adequacy and fluency for the whole record, as well as for the 6 elements.
   - For Spanish, all three online MT systems achieved average scores above 3.6 on adequacy and fluency for the whole record, as well as for all 6 elements.
   - We are more confident that for Spanish the current MT systems can generate adequate translation for metadata records.

Conclusions (2)

2. What metadata elements have the potential to provide multilingual information access to metadata records based on current MT technologies?
   - MT systems could perform better on some metadata elements than others.
   - Metadata elements, such as Subject and Creator are promising elements for users to access the digital objects in digital collections in a different language from English.
   - In other words, digital library developers can implement cross-language search of their English collections by translating these access points into other languages using low-cost MT services.

Conclusions (3)

3. What can be a more effective and yet inexpensive metadata records translation strategy for digital collections?

Answer: MEMT, which combines MT results from multiple MT systems with additional manually generated reference translations as training materials, produces better MT results

- We conducted MEMT by combining MT results from Google+Bing+Yahoo!, half (1005) multilingual metadata records (in English, Chinese, Spanish), used an MT platform called Moses (http://www.statmt.org/moses/) to generate MT results for the other half (1005) metadata records.
- The evaluation showed that the MEMT results were significantly better than any of the three online MT systems

Other Findings

- Machine translation can be applied to translate certain access points such as Subject, Creator, and Title, but the translation of Description of digital objects is still challenging;
- Manual translation is indeed difficult and time consuming, and hard to guarantee quality;
- Crowdsourcing for manual translation? - Not a good idea.

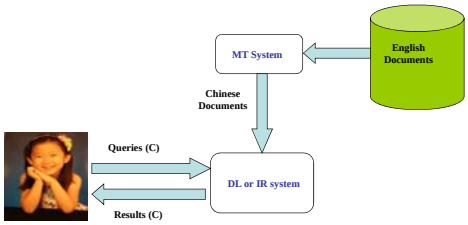
The Effective and Efficient MLIA For Digital Collections (MLIA4DC) Project

- A three-year collaborative research project funded by IMLS (Institute of Museum and Library Services: http://www.imls.gov/) and UNT
- A continuation of the MRT project
- The objectives
  1. Experiment document translation based CLIR model with metadata records;
  2. Apply and compare different MT strategies on metadata records CLIR; and
  3. Provide guidelines to digital libraries for implementing MLIA service.

The MLIA4DC Project

- Research Question: What is the effective and efficient strategy to implement MLIA for digital collections?
  - How effective are current online MT services for document translation based CLIR?
  - How effective are different multi-engine MT (MEMT) strategies for document translation based CLIR?
  - What is the performance of online MT services on query based CLIR?
  - What are the cost of implementing CLIR with each model?
  - What are the needed resources for implementing CLIR with each model?

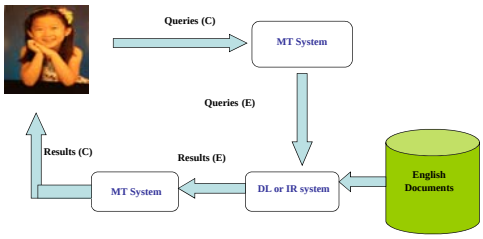
Document Translation Based Cross-Language Information Retrieval (CLIR)



Sample query:白内障有哪些新药？

What are the newest medicines/treatments for cataract?

Query Translation Based Cross-Language Information Retrieval (CLIR)



Sample query: 白内障有哪些新药？

What are the newest medicines/treatments for cataract?

MLIA4DC Project - Activities

1. Test Data Preparation—Acquire 1 million metadata records from two digital collections; MT using two MT systems;
2. Multilingual Corpus Generation—Develop a parallel corpus comprised of English, Simplified Chinese, and Spanish;
3. MEMT Experiments—Use Moses to integrate MT results and linguistic resources to produce new translations;
4. CLIR Experiments—Conduct Experiments based on different MT results from different Systems;
5. Evaluation —Analyze results and measure effectiveness and efficiency of applying MT.

Planned Experiments

I. English Baseline—monolingual retrieval
II. Online System A—CLIR using 1st online MT system
III. Online System B—CLIR using 2nd online MT system
IV. MEMT1—CLIR using MEMT that combines results of Systems A and B
V. MEMT2—CLIR using MEMT1 + the multilingual corpora created manually
VI. MEMT3—CLIR using MEMT2 + monolingual corpus (Chinese and Spanish respectively)

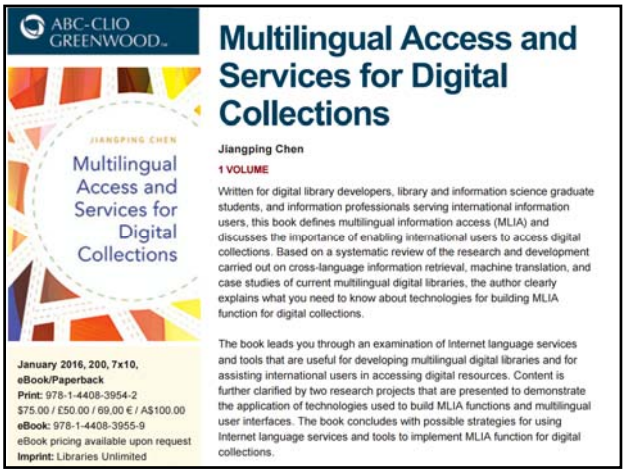
Note: II – VI will be conducted on Chinese-English, and Spanish-English language pairs

Current Progress

- We have done much of the work!
  - ✓ Test Data Preparation—Acquire 1 million metadata records from two digital collections; MT using two MT systems;
  - ✓ Multilingual Corpus Generation—Develop a parallel corpus comprised of English, Simplified Chinese, and Spanish;
- We are working on the following tasks!
  - ❖ MEMT Experiments—Use Moses to integrate MT results and linguistic resources to produce new translations;
  - ❖ CLIR Experiments—Conduct Cross-Language Information Retrieval (CLIR) Experiments based on different MT results;
- We will complete the project by end of this year!
- Please visit the project website (http://iia01.ci.unt.edu/MLIA/) for updates!

My New Book Includes Everything!

- "Multilingual Access and Services for Digital Collections"
  - ISBN: 978-1-4408-3954-2 EISBN: 978-1-4408-3955-9
  - Published in January 2016 by Libraries Unlimited http://www.abc-clio.com/LibrariesUnlimited/product.aspx?pc=A4845P
  - Written for digital library developers, LIS graduate students, and information professionals who are interested in serving international users
  - Seven chapters. The book describes theories, research, and technologies for building multilingual services for digital collections and libraries



Seven Chapters of the Book

1. Introduction
2. Cross-Language Information Retrieval
3. Machine Translation Research and Practice
4. Machine Translation for Digital Collections
5. Multilingual Systems and Interfaces
6. Beyond Retrieval: Knowledge Discovery and Future Directions
7. Related Internet Language Resources and Tools

Summary: Serving International Users (1)

- For reference librarians
  - Get familiar with free online language resources including MT services such as Google Translate or Bing Translator
  - Instruct English users who need to access foreign materials to use online MT services first to get their queries translated into the target language, then to use Google to locate foreign information resources
  - Use online MT services to communicate with users who don't speak English, and
  - Help them use online MT services to get their queries translated into English and then search in English online databases/catalogs

Summary: Serving International Users (2)

- For digital librarians and technical librarians
  - Integrate online MT services with your digital collections/digital libraries to provide query translation based cross-language retrieval services.
  - Develop your own MT systems using free MT platforms such as Moses following the MLIA4DC Project to implement metadata records translation based cross-language retrieval services

Thank You!

Any comments, suggestions are welcome!

Please contact: Jiangping.Chen@unt.edu