Multilingual Information Access for Digital Libraries – The Metadata Records Translation Project

Jiangping Chen
Http://max.lis.unt.edu/
Jiangping.chen@unt.edu
July 2011
Presentation Outline

• About Me
• Current research project – The MRT Project
  – Background
  – Research Design and Research plan
  – Current Progress
• Reflection and thoughts about future research
About Me

• An Associate Professor at the Department of Library and Information Sciences, College of Information in University of North Texas

• Received bachelor’s degree in Information Science from Wuhan University and master’s degree from the Library of Chinese Academy of Sciences

• Worked in two different information organizations in China for 7 years before I went abroad for my doctorate in Syracuse University
Multilingual Information Access (MLIA) for Digital Libraries

• The concept
• Background
• Current project: the Metadata Records Translation Project
What is Multilingual Information Access (MLIA)?

- An extension of Cross-Language Information Retrieval (CLIR)
- To facilitate universal information access by removing language barriers
- Includes but not limited to:
  - CLIR, CLQA, Cross-Language Summarization,
  - Bilingual or multilingual searching, browsing, and presentation, post-retrieval processing
Why MLIA?

• Information on the Web
  ▪ multimedia, multilingual, distributed

• Information Creators and Users
  ▪ with diverse cultural background, interests, languages, information literacy skills

• Information is Power
  • Every information its user
  • Every user his/her information
Why MLIA?

- The need to access information in many languages
  - For economic development
  - For knowledge sharing/cultural exchange
  - For learning
  - For national security
- In practice, Google has provided cross-language search service since 2007
Search across languages

Type a search phrase in your own language to easily find pages in another language. We'll translate the results for you to read.

Search for: 
My language: English  Search pages written in: Spanish

Translate and Search

Tip: Use advanced search to restrict your search by language and country without translating your search phrase.

Translate text

Spanish  ☞  English  Translate

Translate a web page

http://

Spanish  ☞  English  Translate

Use the Google Interface in Your Language
Translated results from Simplified Chinese web pages

English translation

**Green Energy Information Network**
Not only green energy, including renewable energy: solar, wind energy, hydropower, biomass, ocean energy, etc.; also includes the application of technology turning waste into treasure: straw, garbage, and other new energy sources. China's information network to provide green energy, green energy, information, green energy laws...
www.canonyguan.com/ - 69k - Cached

**Green Energy - Baidu Wikipedia**
"Green" energy has two meanings: First, the use of modern technology, development of clean, non-polluting new energy, like solar, wind energy, tidal energy, etc.; second of harm into, combined with the improvement of the environment, make full use of municipal solid waste sludge and other waste reservoir of...
baike.baidu.com/view/295338.htm - 24k - Cached

Original Simplified Chinese - Hide Simplified Chinese results

**绿色能源资讯网**
绿色能源不仅包括可再生能源: 太阳能、风能、水能、生物质能、海洋能等, 还包括应用科技变废为宝的: 稻秆、垃圾等新型能源. 中国绿色能源资讯网提供绿色能源资讯. **绿色能源法律 ...**
www.canonyguan.com/ - 69k - 网页快照

**绿色能源 - 百度百科**
“绿色”能源有两层含义：一是利用现代技术开发干净、无污染新能源，如太阳能、风能、潮汐能等；二是变害为益，同改善环境相结合，充分利用城市垃圾污泥等废料中所蕴藏的 ...
baike.baidu.com/view/295338.htm - 24k - 网页快照
• Google’s cross-language search is built upon many years of research in Machine translation (MT) and Cross-Language Information Retrieval (CLIR)

• CLIR Evaluation Fora
  - TREC (before year 2000)
  - NTCIR
  - CLEF
The 8th NTCIR Workshop (2009/2010)
May 2009 - June 2010
Final Meeting: June 15-18, 2010, NII, Tokyo, Japan

Participation is invited from anyone interested in research on information access technologies and evaluation of them, such as retrieval of various document genres, cross-lingual information retrieval of Asian languages, question answering and cross-lingual information access. NTCIR Workshops are periodical event which are held once per one and half years.

**TASK DESCRIPTION**

**TASK INFORMATION:** ACLIA - GeoTime - MOAT - Patent Mining - Patent Translation - Pilot Task (Community QA)
The Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

CLEF 2009 Workshop
30 September - 2 October, Corfu, Greece, in conjunction with ECDL2009
CLEF2009 Working Notes are now online

CLEF 2008 Post Workshop Proceedings
Evaluating Systems for Multilingual and Multimodal Information Access
Lecture Notes in Computer Science Vol. 5708.
NOW AVAILABLE

To be included on the CLEF mailing list and for further information, contact:
Carol Peters (carol.peters@isti.cnr.it)

CLEF 2010 will take the form of an independent, peer-reviewed Conference organised in conjunction with a set of Evaluation Labs

CLEF is an activity of the TrebleCLEF Coordination Action

All text is available under the terms of the Creative Commons Licence.
MLIA Challenges and Opportunities (1)

• Translation: can MT systems be trusted?
• User related: Where are the users?
• Application related: What is the effectiveness of its applications?
MLIA Challenges and Opportunities (2)

• Translation ambiguity
  – word sense ambiguity
  – Multiple Chinese translations for a term
    there are 10 translations, two different senses for “food”

• Dictionary related problems
  – Unbalanced number of Chinese equivalents for an English term
  – Coverage of the dictionary
  – Some entries are not appropriate for retrieval purpose
Machine Translation

- Machine translation is needed before and after retrieval
  - MLIR approaches based on different translation methods
  - The results need to be translated into the language of the users
  - Effective and efficient machine translation remains the most challenging problem for CLIA.
- Lost in translation
Examples of unbalanced translations

<table>
<thead>
<tr>
<th>term</th>
<th>synonym set</th>
<th>Chinese translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>food</td>
<td>Shi2 wu4食物</td>
</tr>
<tr>
<td></td>
<td>nutrient</td>
<td>Kui4 餐</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shi2 食</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shi2 pin3食品</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cai2 liao4材料</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ying2 yang3 de 营养的</td>
</tr>
</tbody>
</table>

There are 10 translations for “food” in our current dictionary.

<table>
<thead>
<tr>
<th>term</th>
<th>synonym set</th>
<th>Chinese translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>onion</td>
<td>onion</td>
<td>Yang2cong1 洋葱</td>
</tr>
</tbody>
</table>

Only 1 translation for “onion” in current dictionary.

12/22/11  Jiangping Chen
MLIA for Digital Libraries

• Many digital collections are multilingual
• MLIA research has been conducted for years, but real applications of the research results to digital libraries are rare
• MT systems are producing promising results
Digital Library Services

- Search
  - Basic search
  - Advanced search
- Browsing
- Virtual Reference
- Personalized service
- Social Computing
- Bilingual or multilingual Information Access
  - Cross-Language Search?
## Bilingual or Multilingual Digital Libraries in the United States

<table>
<thead>
<tr>
<th>Library Name</th>
<th>URL</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting of Frontiers</td>
<td><a href="http://frontiers.loc.gov/intldl/mtfhtml/mfsplash.html">http://frontiers.loc.gov/intldl/mtfhtml/mfsplash.html</a></td>
<td>English/Russian</td>
</tr>
<tr>
<td>France in America</td>
<td><a href="http://international.loc.gov/intldl/fiahtml/fiahome.html">http://international.loc.gov/intldl/fiahtml/fiahome.html</a></td>
<td>English/French</td>
</tr>
<tr>
<td>Parallel Histories</td>
<td><a href="http://international.loc.gov/intldl/eshtml/">http://international.loc.gov/intldl/eshtml/</a></td>
<td>English/Spanish</td>
</tr>
<tr>
<td>The Perseus Digital Library</td>
<td><a href="http://www.perseus.tufts.edu">http://www.perseus.tufts.edu</a></td>
<td>Greek, English, Latin</td>
</tr>
</tbody>
</table>
The Five Digital Libraries Share the Following Characteristics

• They have been funded by various funding agencies, especially from the federal government;
• They are the products of collaboration. People from different countries work together to produce the bilingual or multilingual 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 for DL: Questions for Discussion

• How useful are current MLIA technologies for digital libraries?
• What are the costs and benefits for DLs to provide multilingual access?
  – Follow traditional Information System development principles
• Are there other solutions to language barriers for information access?
  – With Web 2.0, you may have somebody translate a document for you. Can you count on that?
The Metadata Records Translation (MRT) Project

- A collaborative project among four units in three countries – UNT college of Information, UNT Libraries, IM School, and UAEM in Mexico
- Two-year project funded by IMLS national leadership grant and UNT
- The goals include: (1) understand to what extent current MT technologies generate adequate translation for metadata records and (2) identify the most effective metadata records translation strategies for digital collections.
Welcome!

"Enabling Multilingual Information Access to Digital Collections: An Investigation of Metadata Records Translation" is a two-year research project sponsored by the Institute of Museum and Library Services (IMLS) and the University of North Texas (UNT). This project represents a collaboration of four entities: The Department of Library and Information Sciences in the College of Information at UNT; the UNT Libraries Digital Projects Unit (DPU); the School of Information Management at Wuhan University, China; and the Autonomous University of the State of Mexico (UAEM) in Mexico. It aims to evaluate the extent to which current machine translation technologies generate adequate translation for metadata records, and to identify the most effective metadata records translation strategies for digital collections.

Sponsors/Participating Institutions:

Dr. Jianping Chen, Department of Library and Information Sciences, College of Information
Discovery Park, Room E287J, 1155 Union Circle #311068, Denton, TX 76203-5017
Phone: (940) 369-8393   Fax (940) 565-3101   Email: jianping.chen@unt.edu
MRT: Research Plan

• Extraction of sample data sets from two different digital collections
• Development of the evaluation system
• Manual and machine translation for evaluation
• Usability testing of the evaluation system
• Human evaluation of machine translation (MT)
• Exploring Applicable MT strategies
• Second round of human evaluation of MT translation
Assemble Project Team

Acquire 2200 metadata records

Determine Evaluation Metrics

Develop MT Evaluation System

Generate Reference Translation

Conduct MT with Selected Systems

1st Round Metadata Records MT Evaluation

Analyze Evaluation Results

Answer to RQ1

Develop Multi-Engine Translation Strategies

Modify MT Evaluation System

2nd Round Metadata Records MT Evaluation

Analyze Evaluation Results

Answer to RQ2

Project Evaluation and Information Dissemination

Phase 1

Phase 2

Phase 3

Phase 4

Phase 5

Startup Phase
Project Deliverables

- A machine translation evaluation model appropriate for metadata records;
- An open-source machine translation evaluation system;
- Multi-engine machine translation algorithms for metadata records translation;
- Recommendations and guidelines for digital collection developers on selecting appropriate machine translation strategies;
- A test body of content to be used by the machine translation research communities;
- Publicly available machine translation evaluation results;
- A comprehensive project Web site.
HeMT Design Principles

• This is a prototype system: The major functionalities should be implemented first. Other functions can be added later;

• Easy to use: the system will be used by evaluators from the US, China, and Mexico. So it is important to keep the system easy to understand, fast uploaded, images should be minimized;

• A language independent system structure is desired for the backbone database. However, the system should provide interfaces (webpages) in three languages: English, Simplified Chinese, and Spanish.
Site Map of the Evaluation System
HeMT Users

- **Translators:** conduct manual translation of metadata records;
- **Reviewers:** The research team and Spanish language consultant will review the manual translation, and generate a multilingual terminology for the system;
- **Evaluators:** conduct human evaluation of the system.
We are still Working on it!

A database-driven Web system will be developed to facilitate human evaluation of machine translation.

We will soon invite Web users to participate in this exciting task. Please check this page often, or you can write to Jiangping Chen if you are interested in participation.

Dr. Jiangping Chen, Department of Library and Information Sciences, College of Information
Discovery Park, Room E297J, 1155 Union Circle #311088, Denton, TX 76203-5017
Phone: (940) 369-8393  Fax (940) 565-3101  Email: jiangping.chen@unt.edu
Registration for HeMT

Personal Information

First Name: sdfg
Last Name: dfg
Expected Role in this Project: Evaluator
Education: College Student
Native Language: Chinese
2nd Language: None
You are currently a(an): student

Contact Information

Please provide your phone number and mailing address. The information will be used for awards.

Phone/Cell: 1234
Phone: 
Street Address: dfgfdsg
City: sdfgd
Country: sdfg
ZIP code: sdfg

Enter Your Email and Password
# Manual Translation Page

**Metadata Record ID: 7**

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Creator</td>
</tr>
<tr>
<td>作者</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Title</td>
</tr>
<tr>
<td>书名</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Description</td>
</tr>
<tr>
<td>描述</td>
<td></td>
</tr>
<tr>
<td>Publisher</td>
<td>Publisher</td>
</tr>
<tr>
<td>出版商</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>Subject</td>
</tr>
<tr>
<td>主题</td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
</tr>
<tr>
<td>概括</td>
<td></td>
</tr>
</tbody>
</table>
Dear Jiangping, you are translating metadata Record: CATb1022412

You can see the whole metadata record here

<table>
<thead>
<tr>
<th>title</th>
<th>Prelude à l'après-midi d'un faune.</th>
</tr>
</thead>
<tbody>
<tr>
<td>书名:</td>
<td></td>
</tr>
<tr>
<td>出版商:</td>
<td></td>
</tr>
</tbody>
</table>

Save and Continue  Skip this Record  Reset

---

Dr. Jiangping Chen, Department of Library and Information Sciences, College of Information
Discovery Park, Room E287J, 1155 Union Circle #311058, Denton, TX 76203-5017
Phone: (940) 369-8393  Fax: (940) 565-3101  Email: jiangping.chen@unt.edu
Evaluation Procedures

• Develop Evaluation System
• Generate manual translation for each record
• Load MT translation into the evaluation system
• Train evaluators
• Conduct evaluators surveys: pre-evaluation, post-evaluation
Training of the Evaluators

- The MRT project and its purpose, etc
- Procedures for evaluators
- Evaluation measures (such as adequacy, fluency, best system, worst system, etc) and the criteria for assign values to these measures
- Comments – why should I provide comments?
- An example to show the evaluation procedures
- Test questions make sure evaluators understand how to do evaluation
### Reference Translation for Metadata Record CATb1039428

<table>
<thead>
<tr>
<th>Title:</th>
<th>Apparel, Korea.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>服饰，韩国。</td>
</tr>
<tr>
<td>Creator:</td>
<td>United States. ## Industry and Trade Administration.</td>
</tr>
<tr>
<td>作者:</td>
<td>美国. ## 商管。</td>
</tr>
<tr>
<td>Subject:</td>
<td>Clothing trade ## Korea.</td>
</tr>
<tr>
<td>主题:</td>
<td>服饰贸易 ## 韩国。</td>
</tr>
<tr>
<td>范围:</td>
<td>封面印刷为英文. ## &quot;国际市场信息系列&quot; ## 十一月. 1979.</td>
</tr>
</tbody>
</table>

Please Evaluate the machine translation results by choosing the following machine translation system:

- **System 1**
- **System 2**
- **System 3**

Please compare the machine translation systems after you have finished the evaluation of all three systems.

**Which system do you think is the best:**

**Which system do you think is the worst:**

**Comment:**

Submit  Reset
You are reviewing translation of Metadata Record: CATb1039428

<table>
<thead>
<tr>
<th>Reference</th>
<th>Measure</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref_1</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>Ref_2</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>MT result</td>
<td>Fluency</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Measure</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref_1</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>Ref_2</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>MT result</td>
<td>Fluency</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Measure</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref_1</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>Ref_2</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>MT result</td>
<td>Fluency</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Measure</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref_1</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>Ref_2</td>
<td>Adequacy</td>
<td></td>
</tr>
<tr>
<td>MT result</td>
<td>Fluency</td>
<td></td>
</tr>
</tbody>
</table>

Please evaluate the translation of the whole record by this system

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comment:

Submit
MRT Project: Characteristics

• Interdisciplinary: computer algorithm, user behavior, information system

• Three languages: English, Chinese, and Spanish
  – Multilingual translation evaluation system
  – Collaboration among researchers in different countries

• Integrate development with research
Reflection of Past Research

• Overall, much room to improve
• Research quality is affected by many factors
  – Time
  – Students/research assistants
  – Tenure pressure
• Research topics: I did what I wanted to do
• Research topic consistency: good to have clearly identified fields.
Thoughts about Future Research

• Understand the strengths, weaknesses, and interests
  – Intelligent information access
  – User behavior
  – Digital libraries

• Collaboration with students, colleagues, and peers

• Combine teaching and research
Future Research

• Multilingual Information Access
  – MRT for Information Access – my next project
  – Theories and models that work for MLIA
  – Culture and information access (information users in different cultural background)

• International collaboration for information professionals in digital environments – research, education, information services and technologies
Thank You!

Any comments, suggestions are welcome! Please Contact: Jiangping.Chen@unt.edu