

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

Jiangping Chen

[Http://max.lis.unt.edu/](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

Language Tools - Windows Internet Explorer

File Edit View Favorites Tools Help

"every inf

http://www.google.com/language_tools?hl=en

Language Tools

Web [Images](#) [Videos](#) [Maps](#) [News](#) [Shopping](#) [Gmail](#) [more](#) [Search settings](#) | [Sign in](#)

Google™

Language Tools

[About Google](#)

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: Search pages written in:

Tip: Use [advanced search](#) to restrict your search by language and country without translating your search phrase.

Translate text

»

Translate a web page

»

Use the Google Interface in Your Language

Internet 100%

start

Inbox - ... C:\Jang... r.is.unt... Language... ASIST20... Search Desktop

12:41 PM

green energy - Google Translate - Windows Internet Explorer

File Edit View Favorites Tools Help

green eni

http://translate.google.com/translate_s?hl=en&q=green+energy&sl=en&tl=zh-CN

green energy - Google Translate

Web Images Videos Maps News Shopping Gmail more

Google translate Home Text and Web Translated Search Tools

Translated Search

Search for: green energy Translated to: 绿色能源 - [Not quite right? Edit](#)

My language: English Search pages written in: Chinese (Simplified)

Translate and Search

Translated results from Simplified Chinese web pages

Results 1 - 10 of about 4,520,000 for 绿色能源.

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.canengyuan.com/ - 69k - [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:秸秆,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.canengyuan.com/ - 69k - [网页快照](#)

[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](#)

[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;二是化害为利,同改善环境相结合,充分利用城市垃圾淤泥等废物中所蕴藏的 ...

baike.baidu.com/view/295338.htm - 24k - [网页快照](#)

start

Internet 100%

5:12 PM

- 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

NTCIR-8 - Windows Internet Explorer

File Edit View Favorites Tools Help

NTCIR

http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-en.html

Live Search

NTCIR-8

NTCIR (NII Test Collection for IR Systems) Project | [NTCIR](#) | [CONTACT INFORMATION](#) | [NII](#)

NTCIR

NTCIR-8 HOME

NTCIR-8 MEETING

- TASK DESCRIPTION**
- TASK INFORMATION**
 - [ACLIA](#)
 - [GeoTime](#)
 - [MOAT](#)
 - [PAT-MN](#)
 - [PAT-MT](#)
 - [PILOT TASK](#)
- HOW TO PARTICIPATE**
- DATA**
- IMPORTANT**

The 8th NTCIR Workshop (2009/2010)

Evaluation of Information Access Technologies:
Information Retrieval, Question Answering, and and Cross-Lingual
Information Access

May 2009 -June 2010

Final Meeting: June15-18, 2010, NII, Tokyo, Japan

[Japanese]

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 \(Community QA\)](#)

Done

Internet 100%

start

Inbox - ... C:\Jiang... ir.is.unt... ASIST20... NTCIR-8... Search Desktop

1:02 PM

Welcome to Cross Language Evaluation Forum - Windows Internet Explorer

File Edit View Favorites Tools Help

CLEF

http://clef-campaign.org/

Welcome to Cross Language Evaluation F...

Live Search

Home RSS Print Page Tools

Cross Language Evaluation Forum



Home

Coordination

CLEF 2009

CLEF 2008

CLEF 2007

CLEF 2006

CLEF 2005

CLEF 2004

CLEF 2003

CLEF 2002

CLEF 2001

CLEF 2000

Publications

Links

Archives

Contact

Steering Committee

Sponsors

Français

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. 5706.

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](#)

http://www.springer.com/computer/artificial/book/978-3-642-04446-5

Internet 100%

start

Inbox - ...

C:\lang...

ir.lis.unt...

ASIST20...

Welcom...

Search Desktop

1:03 PM

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)

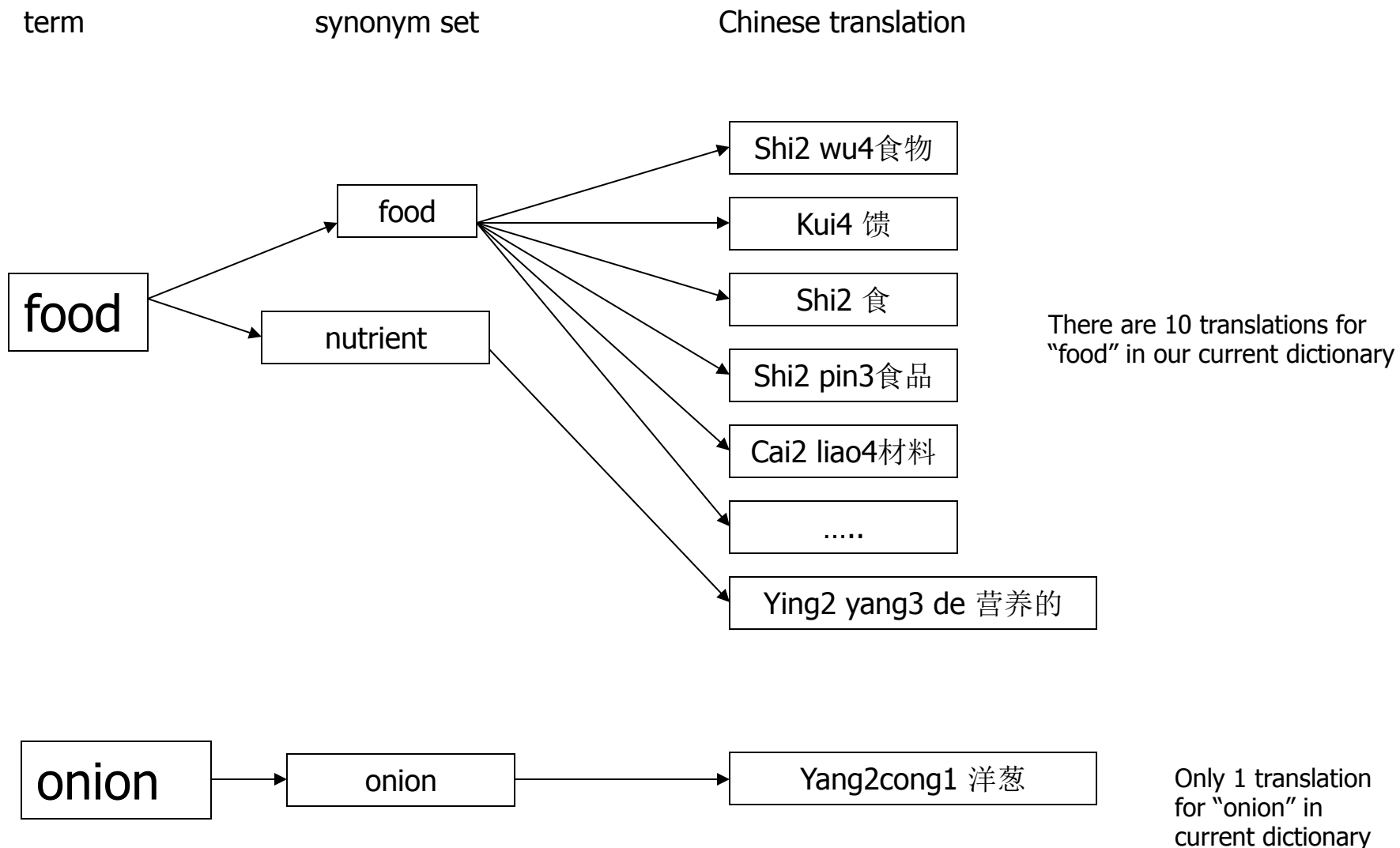
- Translation ambiguity
 - word sense ambiguity
 - Multiple Chinese translations for a term

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



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

Library Name	URL	Languages
Meeting of Frontiers	http://frontiers.loc.gov/intldl/mtfhtml/mfsplash.html	English/Russian
France in America	http://international.loc.gov/intldl/fiahtml/fiahome.html	English/French
Parallel Histories	http://international.loc.gov/intldl/eshtml/	English/Spanish
International Children's Digital Library	http://www.icdlbooks.org/	Digital Objects in 11 languages. Users can do the keyword search in 51 languages.
The Perseus Digital Library	http://www.perseus.tufts.edu	Greek, English, Latin

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.

Metadata Records Translation

[Home](#)

[Project Information](#)

[Advisory Group](#)

[Project Team](#)

[Publications](#)

[Presentations](#)

[Useful Resources](#)

[MT Evaluation](#)

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.

Recent News

The announcement of this project at UNT Website

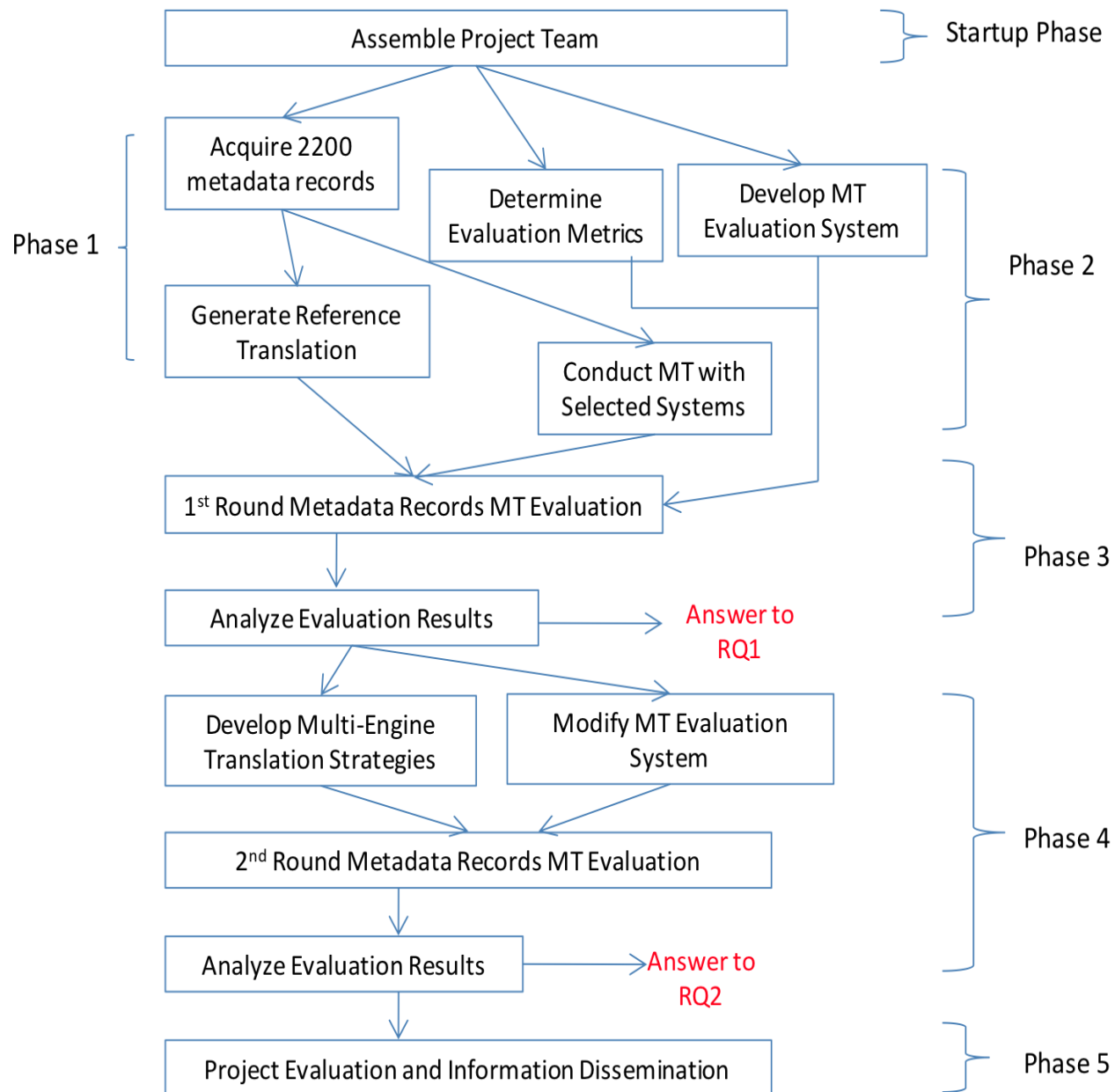
Sponsors/Participating Institutions:



Dr. Jiangping Chen, Department of Library and Information Sciences, College of Information
Discovery Park, Room E297J, 1155 Union Circle #311068, Denton, TX 76203-5017
Phone: (940) 369-8393 Fax: (940) 565-3101 Email: jiangping.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

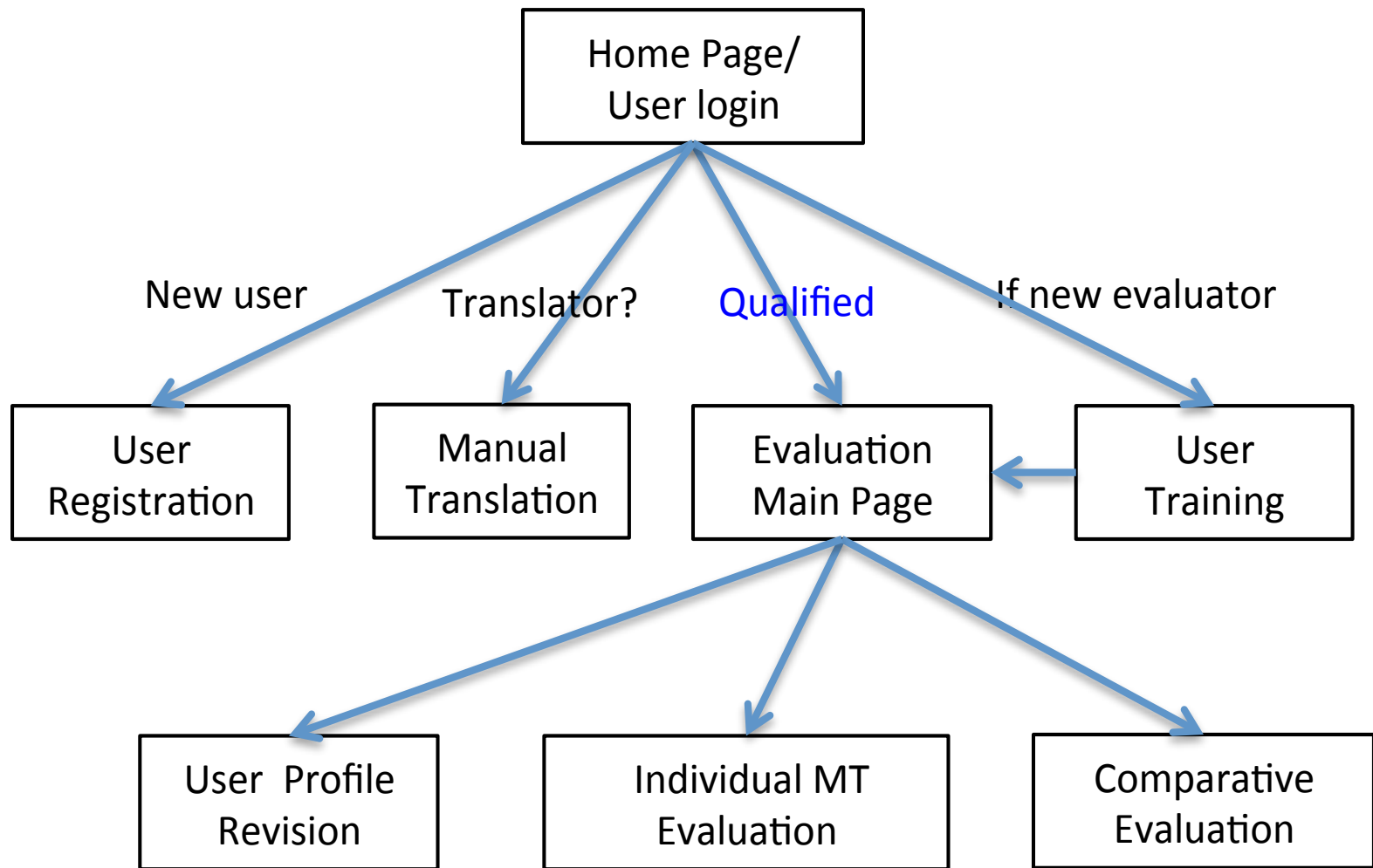


Project Deliverables

- 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

HeMT: Human Evaluation of Machine Translation

Sign In to MyHEMT

Email address:

Password:

Not a User? [Register](#) Now!

Forgot your [Password?](#)

Want to know more about this project?

Please visit [MRT](#) homepage.

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 #311068, Denton, TX 76203-5017

Phone: (940) 369-8393 Fax: (940) 565-3101 Email: jiangping.chen@unt.edu

Evaluation_Registration - Windows Internet Explorer

File Edit View Favorites Tools Help

Search web

http://bxcdk-v10.unt.edu/HeMT/Registration.html

Evaluation_Registration

Registration for HeMT

Personal Information

First Name:

Last Name:

Expected Role in this Project:

Education: (Highest degree obtained)

Native Language:

2nd Language:

You are current a(an):

Contact Information

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

Phone/Cell Phone:

Street Address:

City:

Country:

ZIP code:

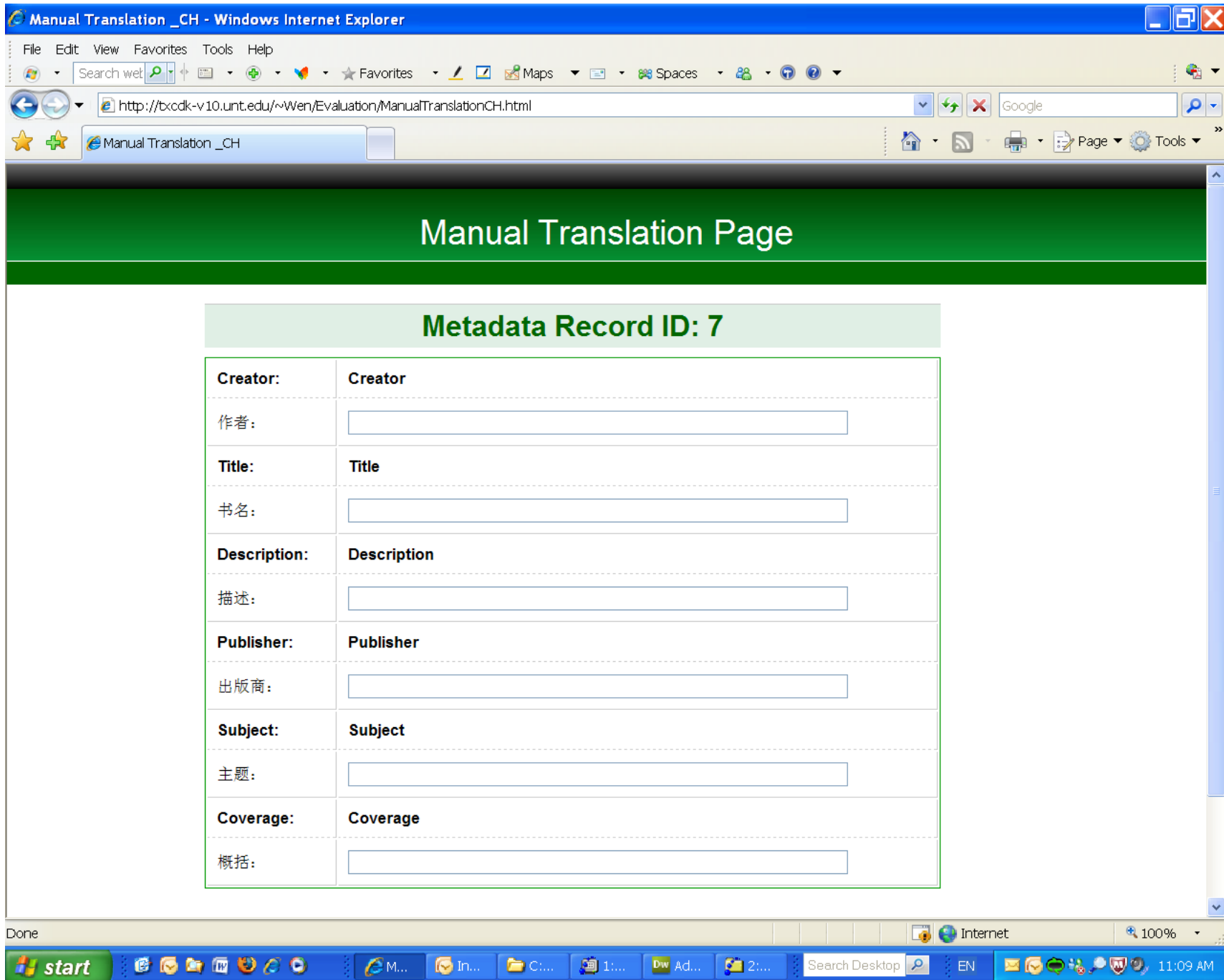
Enter Your Email and Password

Done

Internet 100%

start

Eval... Inbo... C:\J... MRT... bxcd... Search Desktop EN 9:55 AM



Manual Translation - Windows Internet Explorer

File Edit View Favorites Tools Help

Search web

http://bxcdk-v10.unt.edu/HeMT/ManualTranslation.php?userid=jchen06@hotmail.com

HTML form button in PHP

Translated Metadata R... Handling multiple sub... Manual Translation

Home RSS Print Page Tools

Manual Translation

Dear Jiangping, you are translating metadata Record: CATb1022412


You can see [the whole metadata record Here](#)



title:	Prelude a l'apres-midi d'un faune.
书名:	<input type="text"/>
publisher:	Vienna : Universal Edition, 1984.
出版商:	<input type="text"/>

Save and Continue

Skip this Record

Reset



Dr. Jiangping Chen, Department of Library and Information Sciences, College of Information
Discovery Park, Room E297J, 1155 Union Circle #311068, Denton, TX 76203-5017
Phone:  (940) 369-8393  Fax: (940) 565-3101 Email: jiangping.chen@unt.edu

start

S... M... 1:... C... 2:... 3:...

Search Desktop

EN

Internet 100%

4:03 PM

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

Machine Translation Evaluation

[Home](#) / [Logout](#)

Reference Translation for Metadata Record CATb1039428

title:	Apparel, Korea.
标题:	服装. 韩国
creator:	United States. ## Industry and Trade Administration.
作者:	美国. ## 工商管理.
subject:	Clothing trade ## Korea.
主题:	服装贸易 ## 韩国.
publisher:	[Washington] : ## Dept. of Commerce, Industry and Trade Administration, ## 1979.
出版商:	[华盛顿] : ## 商业部, 工商管理, ## 1979.
coverage:	United Kingdom appears in error on cover. ## "International marketing information series." ## Nov. 1979.
范围:	封面错印为英国. ## "国际市场信息系列." ## 十一月. 1979.

Reference Translation for Metadata Record CATb1039428

title:	Apparel, Korea.
标题:	服饰. 朝鲜.
creator:	United States. ## Industry and Trade Administration.
作者:	美国. ## 工业和贸易行政管理.
subject:	Clothing trade ## Korea.
主题:	服装贸易 ## 朝鲜.
publisher:	[Washington] : ## Dept. of Commerce, Industry and Trade Administration, ## 1979.
出版商:	[华盛顿] : ## 商贸工业管理部, ## 1979.
coverage:	United Kingdom appears in error on cover. ## "International marketing information series." ## Nov. 1979.
范围:	封面上联合王国属出版错误. ## "国际市场信息系列." ## 十一月. 1979.

Please Evaluate the machine translation results by choosing the following machine translation system213

[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:



Machine Translation Evaluation

You are reviewing translation of Metadata Record : CATb1039428

标题			
Ref_1	服装, 韩国	Measure	Rating
Ref_2	服装, 朝鲜.	Adequacy	<input type="text"/>
MT result	服装, 韩国.	Fluency	<input type="text"/>

作者			
Ref_1	美国. 韩工商管理.	Measure	Rating
Ref_2	美国. 韩工业和贸易行政管理.	Adequacy	<input type="text"/>
MT result	美国. "工业和贸易管理.	Fluency	<input type="text"/>

主题			
Ref_1	服装贸易 韩 韩国.	Measure	Rating
Ref_2	服装贸易 韩 朝鲜.	Adequacy	<input type="text"/>
MT result	服装贸易 " 韩国.	Fluency	<input type="text"/>

出版商			
Ref_1	[华盛顿]: 韩 商业部, 工商管理, 韩 1979.	Measure	Rating
Ref_2	[华盛顿]: 韩 商贸工业管理部, 韩 1979.	Adequacy	<input type="text"/>
MT result	[美国]: "美国商务部、工业和贸易管理, " 1979年.	Fluency	<input type="text"/>

范围			
Ref_1	封面错印为美国. 韩 "国际信息市场系列." 韩 十一月, 1979.	Measure	Rating
Ref_2	封面上联合国成员国出版错误. 韩 "国际市场信息系列." 韩 十一月, 1979.	Adequacy	<input type="text"/>
MT result	美国出现在封面上的错误. "'国际营销'" - 11-1979年. 信息的系列.	Fluency	<input type="text"/>

Please Evaluate the translation of the whole record by this system

	Adequacy:	<input type="text"/>	Fluency:	<input type="text"/>	
--	-----------	----------------------	----------	----------------------	--

Comment:



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

