

Metadata Records Translation And Evaluation for Multilingual Information Access

Jiangping Chen

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

Jiangping.chen@unt.edu

September 2011

Presentation Outline

- About Me
- Current research project – The MRT Project
 - Background
 - Research Design and Research plan
 - Current Progress
- Collaboration Ideas

Dr. Jiangping Chen

- An Associate Professor at the Department of Library and Information Sciences, College of Information in University of North Texas. **Profesora Asociada del Departamento de Bibliotecas e Información Científica (LIS) de la Universidad de North Texas (UNT).**
- Teaching database design, information architecture, and doctoral level core course on Information Science. **Enseña cursos de diseño de banco de datos, arquitectura de información y ciencias de Información**
- Research interests include natural language processing, information retrieval, and information systems. **Se interesa en los siguientes temas de investigación procesamiento de lenguajes naturales, recuperación de información y sistemas de información**

The Metadata Records Translation (MRT) Project: Background

Proyecto de Traducción de Datos Bibliográficos (MRT): Historia

- Multilingual Information Access (MLIA) Acceso multilingüe a información (MLIA)
 - An extension of Cross-Language Information Retrieval (CLIR). Una extensión de recuperación de información cruzando idiomas
 - To facilitate universal information access by removing language barriers. Para facilitar acceso universal a la información eliminando la barrera del idioma

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 infc" Live Search

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: English Search pages written in: Spanish

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

Translate text

Spanish » English

Translate a web page

Spanish » English

Use the Google Interface in Your Language

Internet 100% 12:41 PM

start **Inbox - ...** C:\Jiang... ir.lis.unt... Languag... ASIST20... Search Desktop MN 12:41 PM

green energy - Google Translate - Windows Internet Explorer

File Edit View Favorites Tools Help

green enei Favorites Maps

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

Live Search Tools

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](#)

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: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](#)

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](#)

Internet 100% 5:12 PM

start Search Desktop 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

File Edit View Favorites Tools Help

NTCIR Favorites Maps Spaces

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

NTCIR-8

Live Search

Page Tools

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

**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 Cross-Lingual
Information Access**

May 2009 -June 2010

Final Meeting: June 15-18, 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%



Inbox - ...

C:\Jiang...

ir.lis.unt...

ASIST20...

NTCIR-8...

Search Desktop



1:02 PM



File Edit View Favorites Tools Help

CLEF Favorites



<http://clef-campaign.org/>



Live Search



Welcome to Cross Language Evaluation F...



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](#)

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

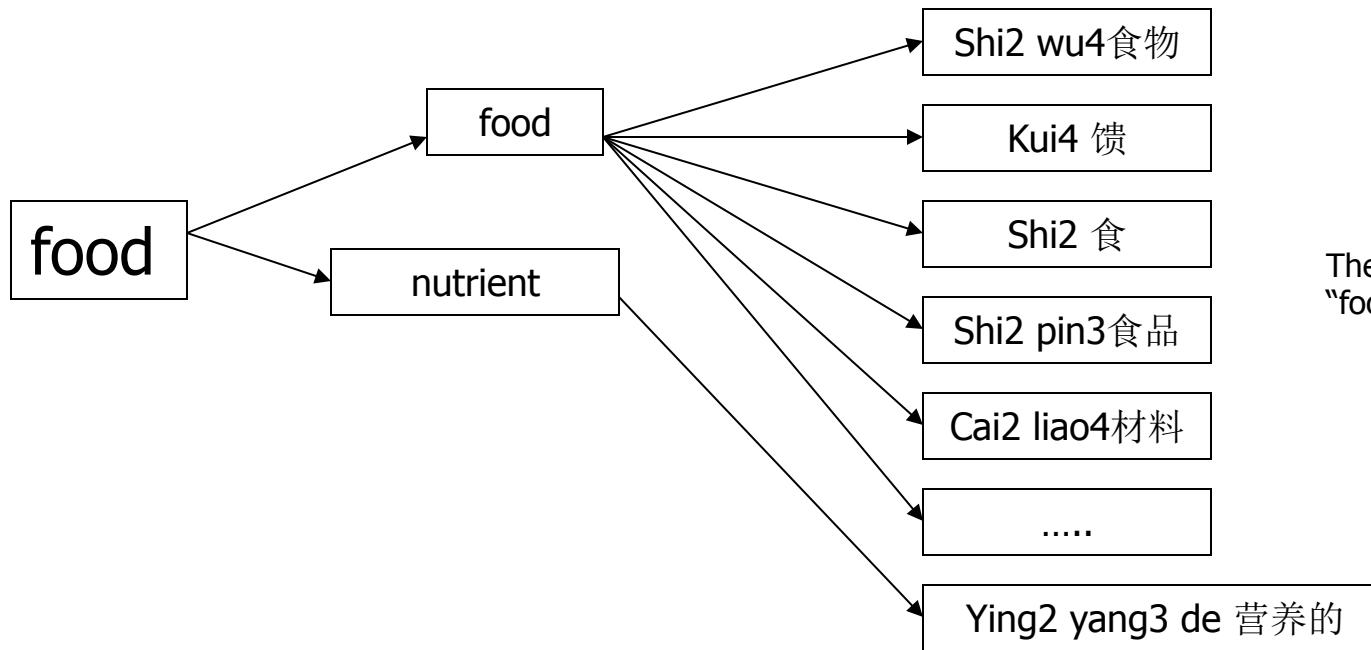
there are 10 translations, two different senses for “food”
- Dictionary related problems
 - Unbalanced number of Chinese equivalents for an English term
 - Coverage of the dictionary
 - Some entries are not appropriate for retrieval purpose

Examples of unbalanced translations

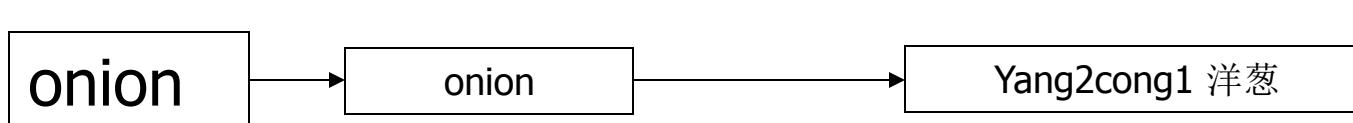
term

synonym set

Chinese translation



There are 10 translations for "food" in our current dictionary



Only 1 translation
for "onion" in
current dictionary

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

MLIA for Digital Libraries

MLIA para Bibliotecas Digitales

- Many digital collections are multilingual **Muchas colecciones digitales son multilingües**
- MLIA research has been conducted for years, but real applications of the research results to digital libraries are rare **Investigaciones MLIA se han realizado por años, pero aplicaciones de los resultados de las investigaciones son raras.**
- Machine translation systems are producing promising results **Las traducciones automáticas estan produciendo resultados prometedores**

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.

The Metadata Records Translation (MRT) Project Proyecto de Traducción de Datos Bibliográficos (MRT)

- A collaborative project among three organizations – UNT, Wuhan University in China, and UAEM. **Un proyecto en colaboración con tres organizaciones UNT, Universidad Wuhan en China y UAEM.**
- Two-year project funded by IMLS National Leadership grant and UNT **Proyecto de dos años patrocinado por IMLS y UNT**
- Project goals **Metas del proyecto**
 - Understand to what extent current machine translation technologies generate adequate translation for metadata records, and **Entender hasta qué punto las tecnologías de traducción automáticas generan traducciones adecuadas para datos bibliográficos, e**
 - Identify the most effective metadata records translation strategies for digital collections. **identificar las estrategias más efectivas en la traducción de datos bibliográficos para colecciones digitales**

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:



武汉大学



Dr. Jiangping Chen, Department of Library and Information Sciences, College of Information

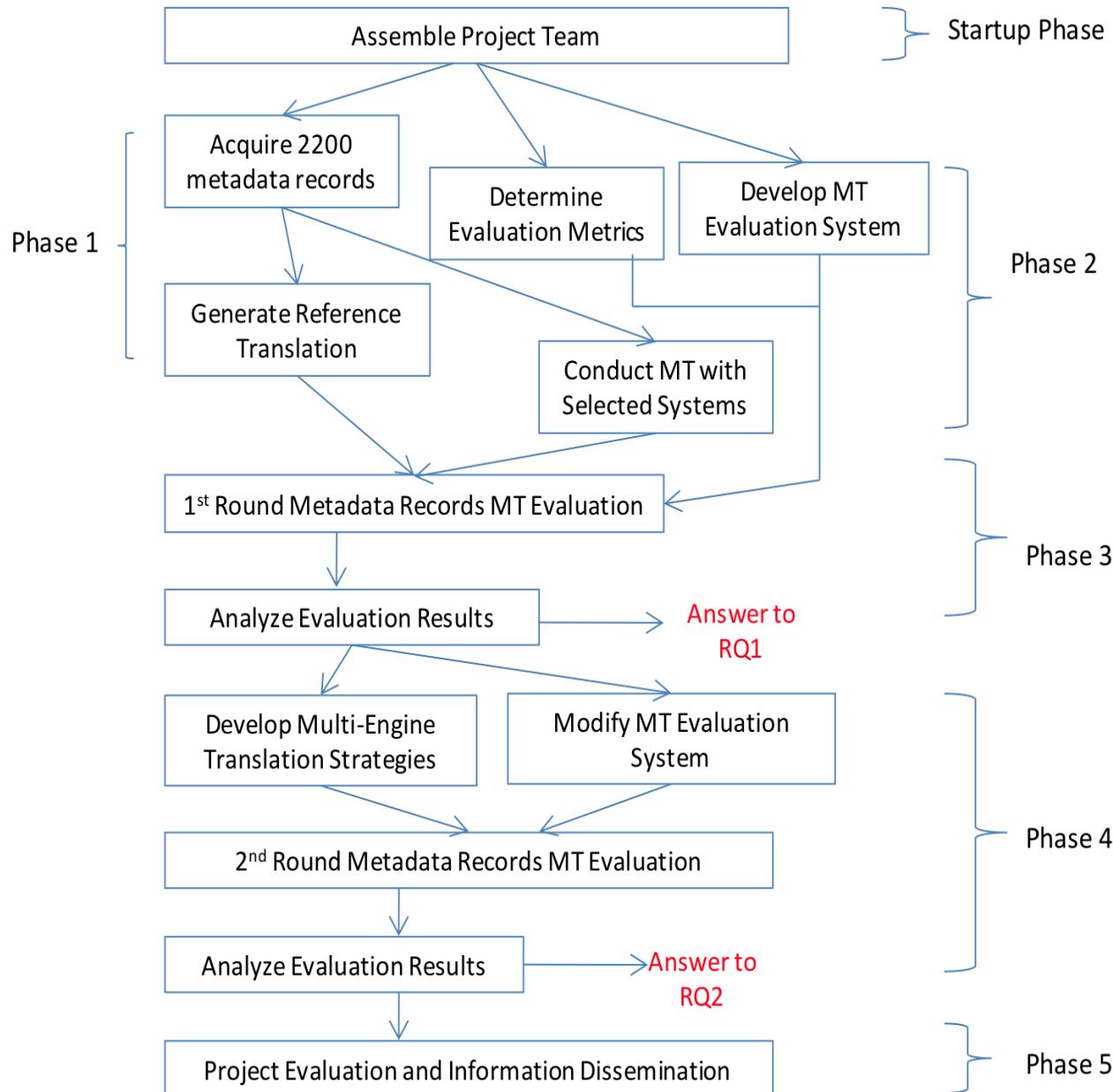
Discovery Park, Room E297J, 1155 Union Circle #311068, Denton, TX 76203-5017

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

MRT: Research Plan

MRT: Plan de Investigación

- Extract sample data sets from two different digital collections.
Extraer una muestra de datos de dos colecciones digitales
- Develop and test an evaluation system. Desarrollar y probar un sistema de evaluación
- Conduct manual and machine translation Conducir traducciones manuales y automáticas
- Conduct human evaluation of machine translation (MT) Evaluación humana de traducciones automáticas
 - I need your help here! Necesito su ayuda para esto.
- Explore applicable MT strategies Explorar estrategias aplicables a MT
- Conduct second round of human evaluation of MT translation Segunda rueda de evaluaciones humanas de las traducciones automáticas



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: 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.

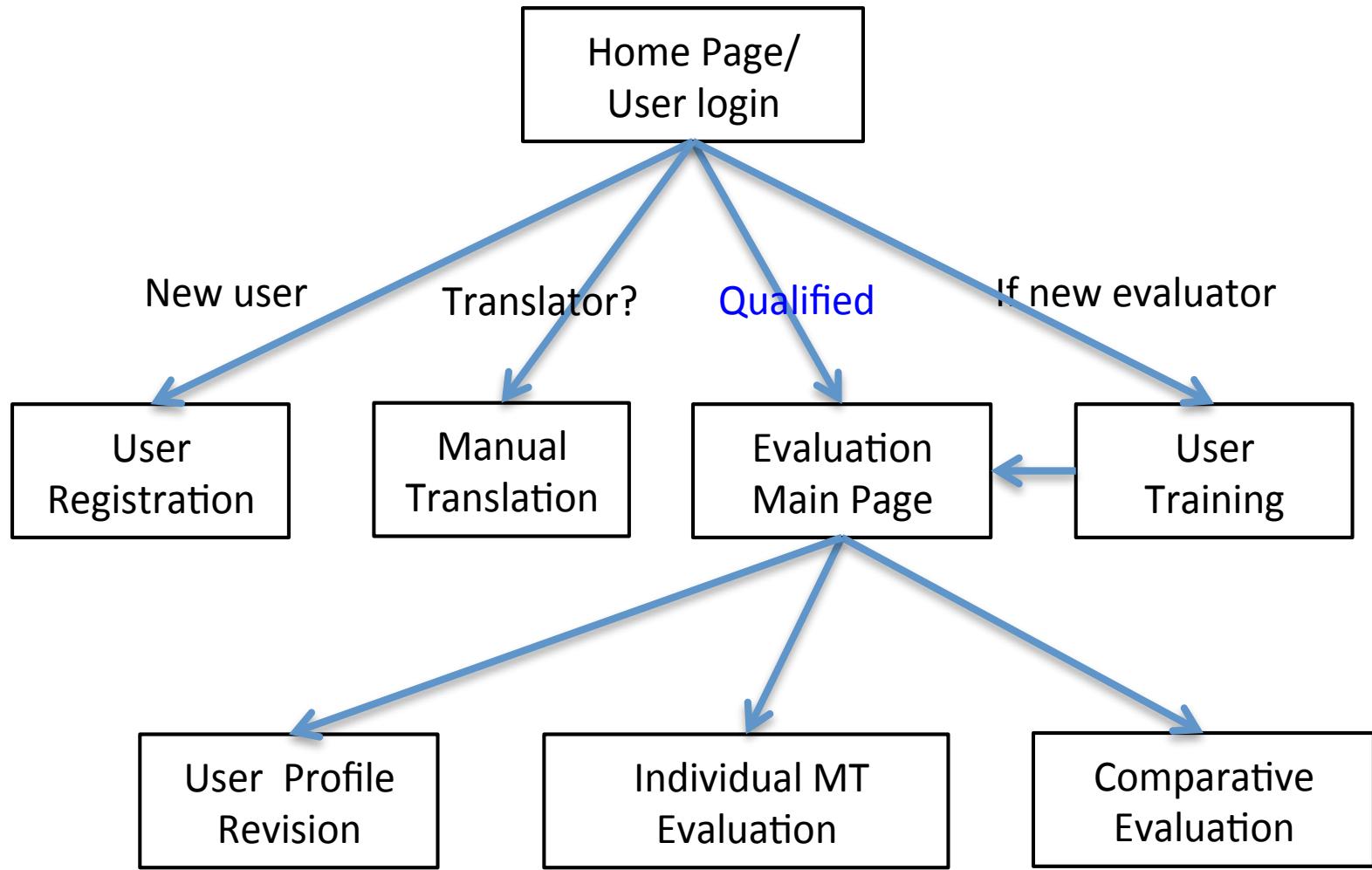


Dr. Jiangping Chen, Department of Library and Information Sciences, College of Information
Discovery Park, Room E297J, 1155 Union Circle #311068, Denton, TX 76203-5017
Phone: (940) 369-8393 Fax: (940) 565-3101 Email: jiangping.chen@unt.edu

The Current Evaluation Site
(Will be in Spanish in the Future)

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

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

HeMT: Training Lessons for Evaluators

Home

Registration

Evaluations

Survey

Self Test

Welcome!

This project, jointly sponsored by the [Institute for Museum and Library Services](#) and the [University of North Texas](#), includes participants from all over the world. We hope you enjoy carrying out evaluations of machine translations. To find out more about this project, please go to the project's [homepage](#).

Your Role

You will be an Evaluator for this project to carry out individual and comparative assessments of English to Chinese / Spanish translations of metadata records.

These training lessons are designed to assist you understand the process of joining the project as an Evaluator, and carrying out machine translation evaluations. Please study ALL materials presented here carefully. The lessons have **three (3)** modules; [Registering](#), [Carrying out Evaluations](#), and [Taking the post-evaluation Survey](#). A short self-test at the end of the modules is for you to evaluate your understanding of the materials presented and our expectations of you as an Evaluator.

Ready? Please go to the [Registration](#) page to start the lessons.

Your cooperation is greatly appreciated. Thank you.

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

I Need Your Help! Necesito su ayuda!

- To recruit 10 -15 UAEM students for participation
Para reclutar la participación de 10-15 estudiantes de UAEM
- Students will go to the evaluation website (will be in Spanish) Los estudiantes iran a la página web de evaluación (que estará en Español)
 - Take a training lesson Tomaran una lección de entrenamiento
 - Register Se registrarán
 - Conduct evaluation Conducirán evaluaciones
 - Students will be awarded in cash or gifts Los estudiantes recibirán premios en efectivo o regalos

Please Contact Me!

Por favor póngase en contacto conmigo

- You can email me: jiangping.chen@unt.edu
- Me puede escribir a: jiangping.chen@unt.edu
- We can have face to face meeting as I will stay in UAEM September 28-September 30. [Nos podemos reunir en persona cuando estoy en UAEM del 28 al 30 de septiembre.](#)
- Thank you! Gracias!



Gracias!