



Information Access Across Languages

Jiangping Chen

School of Library and Information Sciences

University North Texas

<http://max.lis.unt.edu>

April 13, 2005



What is CLIR and CLQA ?

- Cross language information retrieval (CLIR) enables users to search and use information that is in a different language from their queries.

A Scenario:

Users submit Queries or Questions in **Spanish** or **Chinese**

CLIR system returns retrieved **English** documents

- Cross language question answering (CLQA) systems can locate and return answers to questions asked in a different language from the document collection

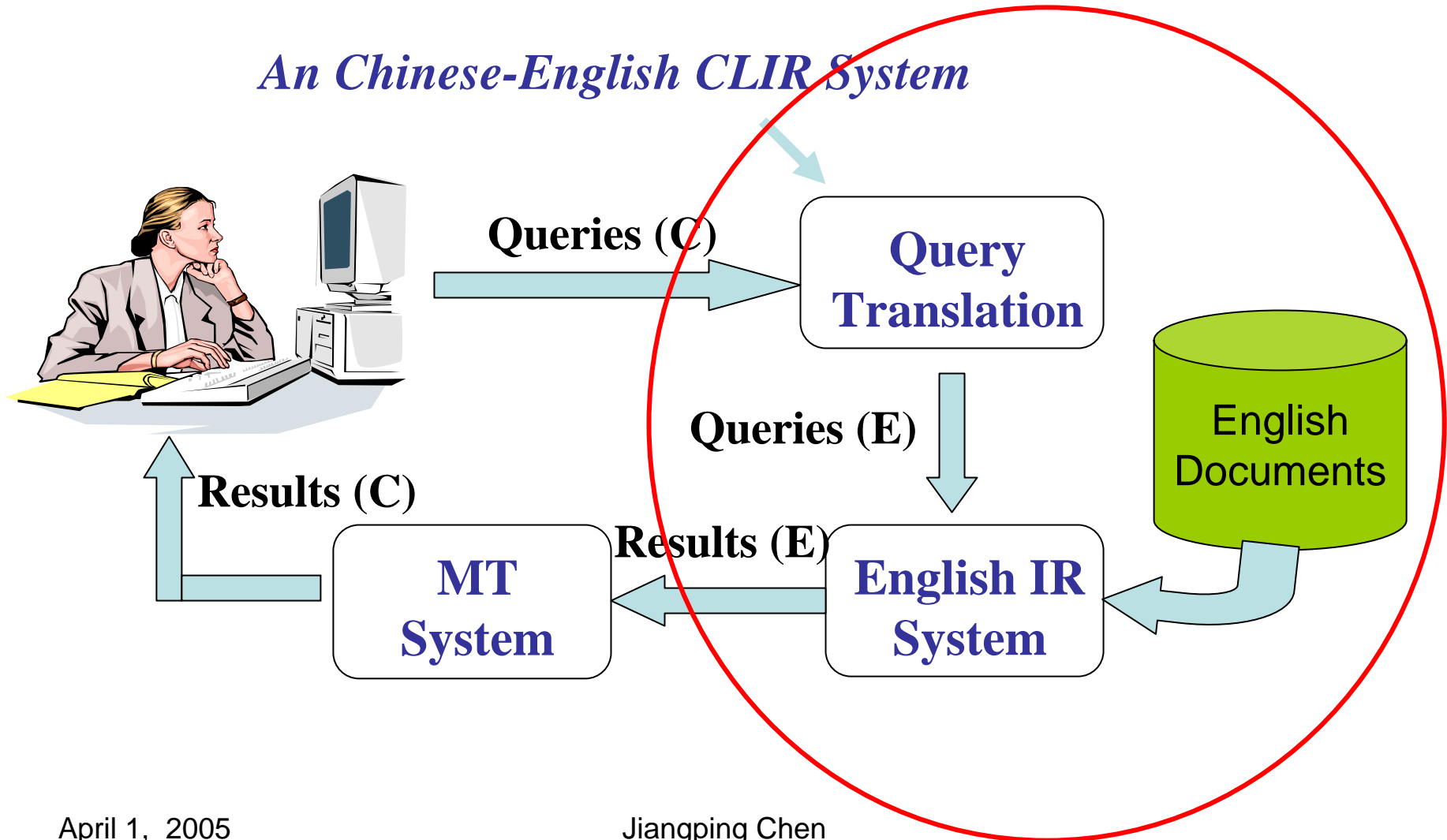
A Scenario:

Users submit Natural Language Questions in **Spanish**

CLIR system finds answers in **English** document collection and returns the answers in **Spanish**

CLIR System Principle

An Chinese-English CLIR System



Importance of CLIR and CLQA

- CLIR and CLQA research are becoming more and more important for global information exchange and knowledge sharing.
 - National Security
 - detect relevant information in documents written in foreign languages
 - Foreign Patent Information Access
 - locate and make use of inventions for new product development for global business
 - Medical Information Access for Patients
 - improve medical information access for minorities

Funding Opportunities

- **NSF** (Natural Science Foundation)
 - Information and Intelligent Systems (NSF05-551)
<http://www.nsf.gov/pubs/2005/nsf05551/nsf05551.htm>
- **NIH** (Natural Institutes of Health)
 - Planning Grant for Integrated Advanced Information Management Systems (PAR-05-064)
<http://grants.nih.gov/grants/guide/pa-files/PAR-05-064.html>
- **DARPA** (Defense Advanced Research Projects Agency)
GLOBAL AUTONOMOUS LANGUAGE EXPLOITATION (GALE) (BAA 05-28)
<http://www.fedgrants.gov/Applicants/DOD/DARPA/CMO/BAA05-28/Grant.html>

Areas of Interests in CLIR and CLQA

- New theories and strategies for CLIR and CLQA in general
 - Semantic
 - Statistical
- Domain specific CLIR and CLQA
 - Medical information
 - Law information
 - Patent
- User-centered CLIR and CLQA
 - Who are the potential users?
 - User needs analysis

My Research on CLIR, QA, and CLQA

- Strategies

- Apply **Natural Language Processing** to process both queries and documents
 - CLIR: **A Lexical Knowledge Base (LKB) Approach**
 - CLQA: **LKB + Machine Learning**
- Make use of available linguistic resources and NLP tools on the Web
- Make use of Web as the knowledge resource

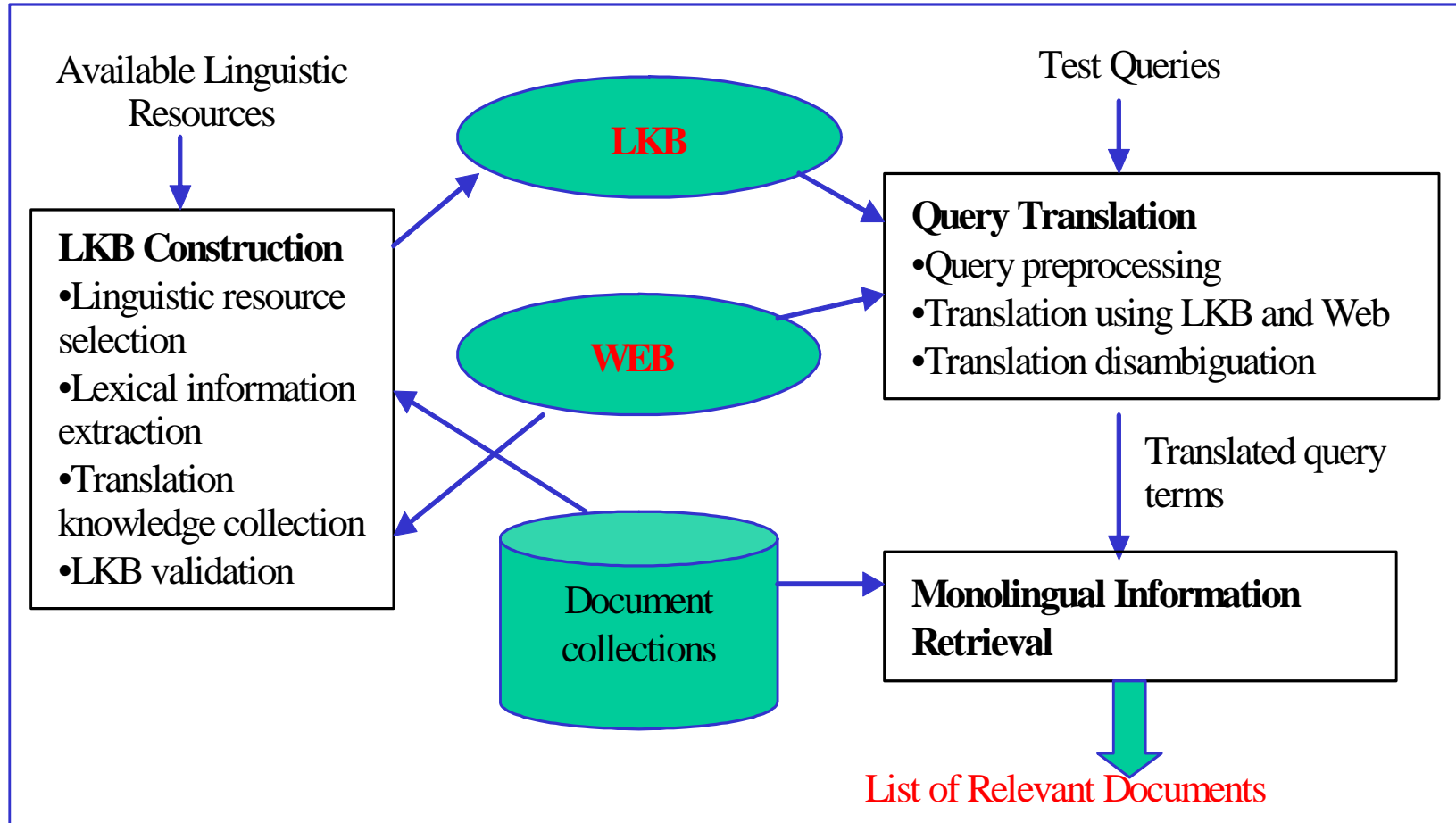
- Research Plan

- Explore system development strategies for both Monolingual and Crosslingual Information Retrieval and Question Answering
 - 2004: English QA, English-Chinese CLIR
 - 2005: CLIR and CLQA between English and Chinese
 - 2006: CLIR and CLQA in Digital Libraries, Add a new language (Spanish?)

My Research on CLIR, QA, and CLQA

- Research Plan (con't)
 - Adding CLIR and CLQA functionalities to digital libraries
 - System strategies that fit into digital library objectives
 - User study for CLIR and CLQA
 - Who are the users?
 - Information need and seeking behavior of potential users
 - Characteristics of user-centered CLIR and CLQA systems
 - CLIR/QA for specific domains or specific users
 - US immigrants
 - Small businesses
 - Children

EagleClir Architecture



Cross Language Question Answering System

- CLQA system components
 1. Question Processing & Translation
 2. Document Annotation
 3. Sentence Retrieval
 4. Web QA
 5. Answer Finding
 6. Answer Determination & Translation

Our NTCIR-5 Participation

- What is NTCIR-5 Workshop?
An Asian languages and English information retrieval evaluation forum sponsored by Japanese government
- Tasks we are going to participate
 - Monolingual Chinese information retrieval
 - Chinese-English cross language information retrieval
 - English-Chinese question answering
 - Chinese-English question answering