Integrating Machine Translation with Digital Collections for Multilingual Information Access

Jiangping Chen
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203
Jiangping.Chen@unt.edu

Olajumoke Agozu
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203
Olajumoke.Agozu@unt.edu

Wenqian Zhao
Southern Nazarene University
4115 North College
Bethany, Oklahoma 73008
wzhao@snu.edu

Cheng Chieh Lien
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203
guitarrex@gmail.com

Ryan Knudson
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203
Ryan.Knudson@unt.edu

Ying Zhang
Carnegie Mellon University
Silicon Valley
Moffett Field, CA, 94035
Joy.Zhang@sv.cmu.edu

Abstract

This paper describes the role of machine translation (MT) for multilingual information access, a service that is desired by digital libraries that wish to provide cross-cultural access to their collections. To understand the performance of MT, we have developed HeMT: an integrated multilingual evaluation platform (http://txcdk-v10.unt.edu/HeMT/) to facilitate human evaluation of machine translation. The results of human evaluation using HeMT on three online MT services are reported. Challenges and benefits of crowdsourcing and collaboration based on our experience are discussed. Additionally, we present the analysis of the translation errors and propose Multi-engine MT strategies to improve translation performance.

1 Introduction

The U.S. Government and many institutions have significant investment in digital libraries and digital collections. Digital Libraries, such as the ACM Digital Library (http://dl.acm.org/), and the International Children's Digital Libraries (http://en.childrenslibrary.org/) are accessed by many users worldwide. However, very few digital collections in the United States support multilingual information access (MLIA) that enables non-English users to search, browse, recognize, and use information from available digital objects. In the increasingly globalized digital knowledge society, libraries and museums need to design and implement effective and efficient MLIA for their digital collections in order to serve broader user groups and to share information with a global community.

This paper first describes how MT can be integrated with information retrieval systems such as search engines to implement MLIA. The quality of MT is of crucial importance to the performance of information retrieval across languages. We then present our evaluation of three online MT services on 2000 metadata records using HeMT (http://txcdk-v10.unt.edu/HeMT/), an evaluation platform we developed in-house. Finally, we analyze the results of the evaluation and propose future work on multi-engine MT to improve translation performance on digital metadata records.

2 Multilingual Information Access and Machine Translation

Multilingual Information Access (MLIA) is a broad term referring to technologies that enable users to retrieve and use information from multilingual collections. The key of MLIA is to

gain access to information in unfamiliar languages. Research on MLIA initially started from exploration on Cross-Language Information Retrieval (CLIR), which has applied three translation strategies: query translation which translates users' queries into the language of the documents; document translation which translates the whole document collection into the language of the users; and an interlingua approach which converts queries and documents into an intermediate language (Oard and Diekema, 1999). Translation is one of the most important steps for CLIR and MLIA. Using human translators, the Library of Congress has created a number of bilingual digital libraries in collaboration with libraries in other countries (Chen and Bao, 2009). Even though manual metadata records translation can be conducted through collaborating with organizations in other countries, it is expensive and time-consuming.

Although MLIA technologies such as CLIR, Cross-language Question Answering and Cross-Language Information Extraction have been actively explored by researchers since the mid-1990s, none of the technologies have been widely applied to existing digital libraries to enable multilingual information access (Gay et al., 2005; Chen and Ruiz 2009). Digital library and museum communities do not trust the performance of current MT systems. To our knowledge, none of the existing bilingual or multilingual digital collections in the U.S. apply MT for either cross-language search or metadata records translation (Chen and Bao, 2009). Yates (2006) evaluated Babel Fish, an MT system launched in late 1997 on the Internet, and concluded that Babel Fish was not appropriate for most users in law libraries due to the errors in the translation.

3 HeMT: A Multilingual MT Evaluation Platform for Digital Metadata Records

MT technologies have made great progress in recent years with the dramatic funding support from governments and large companies such as Google, Microsoft, and Yahoo! MT has been widely used in translating queries in various experimental CLIR systems with fairly good retrieval performance (Sakai et al., 2008; He and Wu, 2010). It is necessary to systematically assess how current MT technologies perform in translating digital metadata records.

We developed a web-based platform HeMT (Human Evaluation of Machine Translation: http://txcdk-v10.unt.edu/HeMT/) to allow crowdsourcing of high quality manual translations and evaluation of machine translations. HeMT is designed to be used by three types of users: translators (who produce reference translations), MT evaluators (who perform evaluation), and reviewers (who review reference translations and monitor the evaluation process). These users interact with the six functional modules of HeMT, as illustrated in Figure 1.

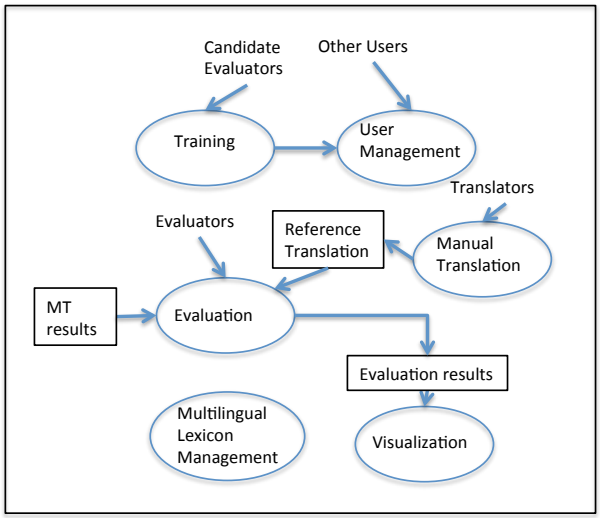

Figure 1: HeMT Structure

The six functional modules include: (1) User Management. This enables all users to register, login, and edit their profiles. It also requests reviewers to approve or deny new users; (2) Multilingual Lexicon Management. The module allows reviewers to generate parallel multilingual terms / phrases so that HeMT can create multilingual webpages. HeMT is language-independent, and new languages can be conveniently added into the system; (3) Manual Translation. This module provides the interface for the creation and storage of reference translations for evaluation. Translators can add or edit translations of English metadata records It also allows reviewers to approve, edit, or deny translators' translations; (4) User Training. This module, designed to reduce possible divergent assessments (Callison-Burch et al., 2010; Lavie,

2010) and ensure standardized evaluations, prepares participants as Evaluators. We have developed training lessons in the three languages involved in the experiments: Chinese, Spanish, and English. The lessons describe the background of the project, procedures for registration and performing evaluations, evaluation measures, and how to handle specific evaluation instances or problems. At the end of the training, an evaluator has to take a 15-question quiz before he/she can register with the system. (5) Evaluation. This is the main module. Once logged in, an evaluator is presented with: (a) the MT results from three online MT services of a metadata record; (b) two reference translations for that record; and (c) pull-down options and textboxes that allow the evaluator to judge the adequacy and fluency of the MT results, as well as which MT system provides the best translations. More information about evaluation measures is presented below; (6) Visualization. Reviewers belonging to the research team will be able to check the evaluation results on the fly. This module presents results of the evaluation in graphs. The evaluation results will be presented in both numbers and color bars for Chinese and Spanish translation respectively. Figure 2 is a screen shot of a page presenting the average adequacy and fluency scores of 3 English-Chinese MT systems.

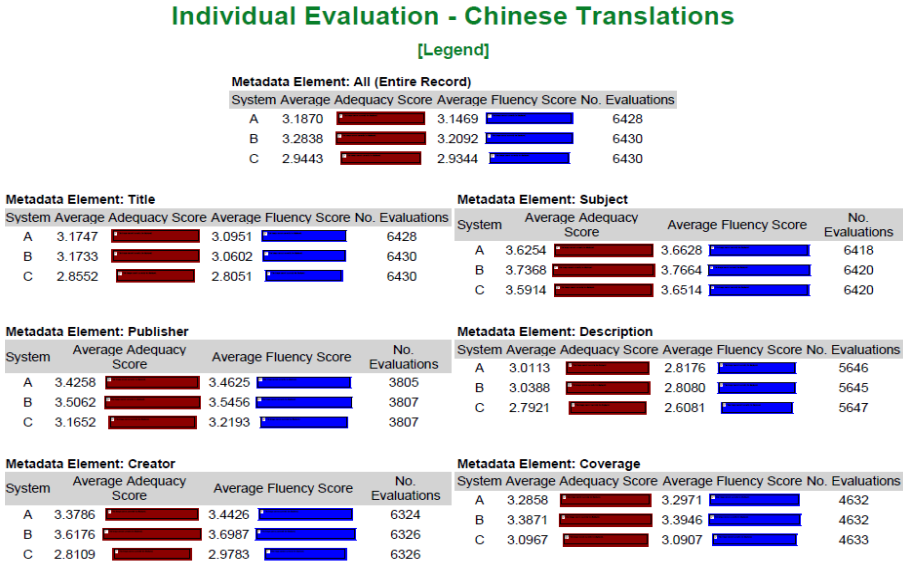

Figure 2: HeMT Page Visualizing Evaluation Results

4 Evaluation Methodologies

4.1 Metadata Records

In order to evaluate the performance of current MT technologies on metadata records, we extracted 2000 metadata records from two digital collections: the UNT Catalog (http://iii.library.unt.edu/), and the Portal to Texas History (http://texashistory.unt.edu/). Metadata records are valuable information generated by librarians to provide access points to an object such as a book, a video, or an image. For example, a metadata record for a book may specify the book's title, author, abstract, publisher and publication date. The elements included in a metadata record can be many. For our purposes, we only kept what we considered the six most valuable elements for translation and evaluation. Table 1 presents the six elements and a sample metadata record for a digital object extracted from the Portal to Texas History. Not all records contain all 6 elements, but each record has at least the first three elements: title, creator, and subject.

Two thousand English metadata records with the six elements were sent to three online MT service (Google, Bing, and Yahoo). They were translated into Simplified Chinese and Spanish. Simultaneously, several translators who are native

speakers of Mandarin or Spanish were recruited to generate reference translations for each metadata record using HeMT.

Evaluators are recruited internationally and nationally. They are only required to speak the language to be evaluated. As mentioned above, each evaluator has to pass a quiz before they can qualify for the job.

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"

Table 1. The Six Elements of Metadata Records

4.2 Evaluation Measures

We employ Adequacy and Fluency as measures to evaluate the MT performance on individual records and their six elements (LDC, 2005). Table 2 summarizes our principles for judging translations using these two measures.

Additionally, we asked the evaluators to compare the three systems and identify the best and worst translation for each element and the whole record. In other words, machine translated records and their elements are assessed (1) individually using Adequacy and Fluency, for how accurately and completely meanings are expressed in comparison to equivalent human manual translations; and (2) comparatively, a comparison of the performance of three machine translation systems to determine which is perceived as the best, and which is perceived as the worst.

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

Table 2: Principals for Judgment of Translations

4.3 Evaluation Process

Six Chinese evaluators were recruited from Chinese Universities. Five of them were masters' students in Library and Information Science programs. One was an undergraduate in the same program.

The evaluation was conducted in early Spring 2012. Interested candidates were asked to take the online training on HeMT. Those who passed the quiz were allowed to register into the HeMT system. When an evaluator logs in, she/he can check the number of records evaluated. When she/he choose to evaluate a new metadata record, the evaluator will be presented with a randomly selected record and required to evaluate the translation results one by one. After the evaluator finishes assessing the three MT results for each record as a whole and its individual elements, HeMT will present the comparative evaluation page for assessment.

HeMT also allows evaluators to comment on their evaluation and issues encountered during the process. Evaluators can reference training pages at any time during the evaluation and are also required to take a survey about the system and evaluation process once they are done with the evaluations.

5 Results and Analysis

At the time of this writing, the evaluation on Chinese translation has been completed and that on Spanish is still ongoing. Below we report the results on Chinese translation and our preliminary analysis of these results.

5.1 Inter-evaluator Reliability

Considering MT evaluation as a coding process, we assess the inter-evaluator reliability applying the approach by Callison-Burch et al (Callison-Burch, et al, 2007). We obtained the Kappa Coefficients for the evaluation of the Chinese translations, which are presented in Table 3.

Kappa Coefficient is commonly used to measure the agreement between two or more raters. The results in Table 3 show that the inter-evaluator reliability was low. For whole record evaluation, Kappa efficient is 0.20 on Adequacy, which means only slight agreement was achieved among the 6 evaluators. Some elements were even worse, such as the Title (0.03 on Adequacy) and the Description (around 0.05). Some elements, such as Publisher (0.53) and Coverage (0.34), were much higher.

Metadata Element	Measure	Kappa Coefficient
Whole Record	Adequacy	0.1965
	Fluency	0.1236
Title	Adequacy	0.0310
	Fluency	0.0334
Subject	Adequacy	0.0838
	Fluency	0.0827
Publisher	Adequacy	0.5284
	Fluency	0.5078
Description	Adequacy	0.0494
	Fluency	0.0985
Creator	Adequacy	0.2722
	Fluency	0.2086
Coverage	Adequacy	0.3389
	Fluency	0.3214

Table 3: Inter-evaluator Reliability on Chinese Translation

Metadata Element	Measure	Kappa Coefficient
Whole Record	Best	0.2805
	Worst	0.3596
Title	Best	0.2619
	Worst	0.4186
Subject	Best	0.2580
	Worst	0.3627
Publisher	Best	0.6218
	Worst	0.6375
Description	Best	0.2358
	Worst	0.3544
Creator	Best	0.4328
	Worst	0.5003
Coverage	Best	0.2800
	Worst	0.3516

Table 4: Inter-evaluator Reliability on Comparative Evaluation on Chinese Translation

Interestingly, Kappa Coefficient of the other two measures – best system and worst system were much more consistent among the six evaluators. Table 4 presents the Kappa Coefficient for the whole record as well as for the individual elements.

5.2 Adequacy, Fluency, and Best System

We have evaluated the Chinese and Spanish translations of three online MT systems. These systems were freely available at the time of our translation (December 2011). The results in Adequacy and Fluency are presented in Table 5. We used A, B, and C to label the three systems in the following tables.

Metadata Element	MT System	Number of Evals.	Average Score	
			Ade-quacy	Fluen-cy
Whole Record	A	6428	3.19	3.15
	B	6430	3.28	3.21
	C	6430	2.94	2.93
Title	A	6428	3.17	3.10
	B	6430	3.17	3.06
	C	6430	2.86	2.81
Subject	A	6418	3.63	3.66
	B	6420	3.74	3.77
	C	6420	3.59	3.65
Publisher	A	3805	3.43	3.46
	B	3807	3.51	3.55
	C	3807	3.17	3.22
Description	A	5646	3.01	2.82
	B	5645	3.04	2.81
	C	5647	2.79	2.61
Creator	A	6324	3.38	3.44
	B	6326	3.62	3.70
	C	6326	2.81	2.98
Coverage	A	4632	3.29	3.30
	B	4632	3.39	3.39
	C	4633	3.10	3.09

Table 5: Results for Individual Evaluation of MT systems (Chinese)

Table 5 shows System A and System B received a mean score above 3 on both Adequacy and Fluency for the whole metadata records. As for individual elements, Description receives the lowest score on Adequacy and Fluency. Subject receives the highest score. All three systems were judged above 3.5 on Subject, which is one of the most important access points for retrieval.

Table 6 presents the results of the comparative evaluation. System A was consistently chosen as the best system. Note HeMT used a random process to present the order of the three systems to the evaluators. In other words, System A can be presented to the evaluators as System A, B, or C. This design is to avoid possible bias on scoring a system based on its presenting order instead of performance.

Metadata Element	MT System	Number of Hits	Percentage
Whole Record	A	3224	50.15%
	B	2327	36.20%
	C	878	13.66%
Title	A	2691	41.86%
	B	2541	39.52%
	C	1197	18.62%
Subject	A	2781	43.32%
	B	1981	30.86%
	C	1657	25.81%
Publisher	A	1746	45.87%
	B	1364	35.84%
	C	696	18.29%
Description	A	2460	43.57%
	B	2226	39.43%
	C	960	17.00%
Creator	A	3387	53.55%
	B	2062	32.60%
	C	876	13.85%
Coverage	A	1957	42.25%
	B	1658	35.79%
	C	1017	21.96%

Table 6: Results for Comparative Evaluation of MT systems (Chinese) – Best Rating

5.3 Automatic Evaluation Results

We also calculated the Bleu and Meteor scores of the three systems. See the results in Table 7 (For Chinese translations) and Table 8 (For Spanish Translations). The Spanish scores are based on the evaluation being conducted so far.

System	**A**	**B**	**C**
Chinese Translations - BLEU Scores			
coverage	37.64	41.02	40.13
creator	23.99	31.85	28.65
description	22.60	19.03	19.24
publisher	24.93	24.73	27.60
subject	37.64	36.49	39.90
title	25.26	26.33	23.59
Chinese Translations - METEOR Scores			
coverage	29.54	32.47	29.04
creator	28.74	34.16	30.90
description	25.05	25.12	23.35
publisher	27.45	27.50	27.98
subject	32.92	33.01	32.70
title	26.99	28.17	25.44

Table 7: BLEU and METEOR Scores for Chinese MT Translations

System	A	B	C
Spanish Translations - BLEU Scores			
coverage	59.85	56.31	42.59
creator	77.48	67.03	69.06
description	47.32	47.26	34.76
publisher	68.39	53.50	57.72
subject	44.00	43.66	39.09
title	57.55	54.59	40.68
Spanish Translations - METEOR Scores			
coverage	37.82	38.85	29.85
creator	47.17	43.11	41.73
description	31.85	33.35	27.68
publisher	33.68	30.01	34.19
subject	30.20	34.25	27.03
title	36.19	36.08	27.69

Table 8: BLEU and METEOR Scores for Spanish MT Translations

Note that the BLEU scores of the Spanish translations are much higher than those of the Chinese translations. Also, the METEOR scores of the Spanish translations are higher than those of the Chinese translations, but not as dramatically as with BLEU.

We calculated the correlation (Pearson's *r*) among human evaluation and automatic evaluation scores for Chinese translation. Table 9 presents the results. It shows that Adequacy is closely correlated with Fluency. METEOR is more closely related with human evaluation measures than BLEU.

	Adequacy	*Fluency*	*BLEU*
Adequacy	1		
Fluency	0.96	1	
BLEU	0.55	0.61	1
METEOR	0.73	0.82	0.81

Table 9: Correlation Among Evaluation Scores

5.4 Evaluators' Scoring Pattern

We decided to look at individual evaluators' scoring patterns to gain additional insight into the individual evaluators' behavior. The low inter-evaluator Kappa Coefficient indicates that the individual evaluators may have quite different scoring patterns. Figure 3 and Figure 4 show individual Chinese evaluators scoring patterns on Subject and Title elements.

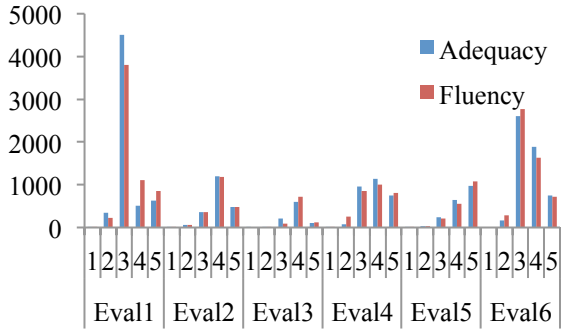

Figure 3: Chinese Individual Evaluator Score Pattern – Subject

Figure 3 shows that the fifth evaluator seems to assign higher scores to most records. Other evaluators assigned more scores of "3" to records. The two measures Adequacy and Fluency are related closely. Figure 4 is the evaluation pattern for the Title element. It shows that the first and the sixth evaluator assigned more scores of "2" than other evaluators.

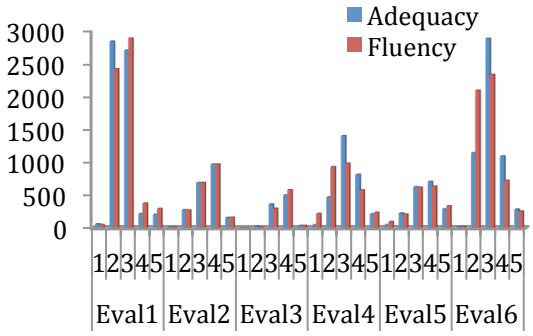

Figure 4: Chinese Individual Evaluator Score Pattern - Title

5.5 Comments from Evaluators

All evaluators have made about 482 comments. Table 10 lists major categories of these comments. It presents some of the existing problems regarding the reference translations, the major problem of MT results, and the issues for comparative translation. For example, most of the comments were on comparative evaluation – two or more systems translated a particular element, such as Creator, the same. HeMT currently only allows the system to choose one best/worst translation. We will have to make changes to handle this in our second round of evaluation.

Category	Examples
Reference Translation	Two reference translation contradict to each other Mistakes in reference translations, reference translations may be for different records
Machine Translation	Specific translation errors, failure to translate person names and location names; personal names should not be translated for Spanish
Comparative Evaluation	Two or three systems provide the same results for certain elements, all systems are bad on translation,

Table 10: Comments from Evaluators

5.6 Translation Errors

We did a preliminary analysis of the Chinese translation errors through identifying the translations that were judged as "1" on Adequacy" or Fluency. Our purpose is to understand the major challenges of current translations. Table 11 shows the top five categories of translation errors.

Rank	Category
1	False translation - not understandable/irrelevant
2	Incomplete translation - with words in original English
3	Poor translation - nonnative expression
4	Missing translation
5	False Translation - person name

Table 11: Translation Error Analysis - Chinese

6 Discussions

The purpose of our work is 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. Our evaluation results showed that 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. The generation of the reference translations for the 2000 records took much more time than we expected. Also, the quality of translation, even being reviewed, is not perfect. The evaluators' comments reflect some of the mistakes present in the reference translations.

Because all the human related activities, such as reference translation generation and MT result evaluation, were conducted in a distributed and collaborative mode – translators and evaluators worked at their own locations and paces, and it was a challenge to keep everyone on track and ensure that the work was done on schedule.

Due to the low inter-evaluator Kappa Coefficient values, we cannot assume the reliability of our evaluation results. We are unclear of the causes of the low inter-evaluator reliability. Translation evaluation is a highly subjective activity which poses special challenges to researchers. We developed the training lesson to reduce misunderstanding and to assist evaluators in making consistent judgments. However, it does not help much as reflected by the Kappa Coefficient statistic. Further investigation will be conducted to understand the low inter-evaluator reliability issue.

7 Ongoing Work and Future Directions

In addition to continuing our Spanish evaluation and result analysis, we have planned possible ways to improve MT of metadata records.

Multi-engine machine translation (MEMT), which combines the results from a variety of MT systems working simultaneously on the same text to improve the overall quality, has been a very active area in machine translation research. We are interested in applying MEMT with additional in-domain parallel or monolingual corpora to translate metadata records. Currently we use Moses to test our MEMT approaches. The idea is to generate both Language Models and Translation Models using the translation results of the three MT systems plus in-domain Chinese and Spanish metadata records, as well as half of the reference translations.

As we mentioned before, machine translation is the most important step toward multilingual information access. Our future work includes conducting cross-language information retrieval using MT results on real-world digital collections. Experiments will be conducted with real users to explore effective and efficient solutions to finding useful information from multilingual digital collections.

Acknowledgment

This study is supported by the Institute of Museum and Library Services (IMLS) grant LG-06-10-0162-10. We thank all translators, evaluators, and our Spanish consultant for their hard work and support.

References

Callison-Burch, et al., (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of ACL-2007 Workshop on Statistical Machine Translation*.

Callison-Burch, et al. (2010). Findings of the 2010 workshop on statistical machine translation and metrics for machine translation. In *Proceedings of WMT2010,* July, 2010, Uppsala, Sweden, 17-53.

Chen, J. and Bao, Y. (2009), Information Access Across Languages on the Web: from Search Engines to Digital Libraries. Proceedings of ASIST Annual Conference.

Chen, J. & Ruiz, M. (2009). Towards an integrative approach to cross-language information access for digital libraries. Proceedings of SIGIR 2009 Workshop: Information Access in a Multilingual World, Boston, Massachusetts, USA. July 23, 2009.

Gey, F. C., Kando, N., and Peters, C. (2005). Cross-language information retrieval: the way ahead. Information Processing and Management, 41, 415-431.

He, D. & Wu, D. (2010). Exploring the future integration of machine translation in multilingual information access. Online Proceedings of 2010 iConfernece, February. 3 -6, University of Illinois at Urbana-Champaign.

Lavie, A. (2010). Evaluating the output of machine translation systems. In *Proceedings of AMTA* 2010, Denver, Colorado.

LDC (Linguistic Data Consortium) (2005), "Linguistic Data Annotation Specification:Assessment of Fluency and Adequacy in Translations Revision 1.5", [2010-01-08], Available http://www.ldc.upenn.edu/-Projects/TIDES/Translation/TransAssess04.pdf

Oard, D. (2009). Multilingual Information Access. in (Bates, M. J. ed.) Encyclopedia of Library and Information Sciences, 3rd Ed. Available at: http://terpconnect.umd.edu/~oard/pdf/elis09.pdf.

Oard, D. W., & Diekema, A. R. (1999). Cross-language information retrieval. In M. Williams (Ed.), *Annual Review of Information Science and Technology*, *33* (pp. 223-256).

Ruiz, M.E., Chen, J., Druin, A., Kando, N., Oard, D., and Peters, C. (2008). Multilingual Access to Digital Libraries. In Proceedings of the ASIS&T 2008 Annual Meeting, Columbus, OH. October 2008.

Sakai, T., Kando, N., et al (2008). Overview of the NTCIR-7ACLIA IR4QA task. *Online proceedings of NTCIR-7 Workshop Meeting*. December 16-19, 2008. Tokyo, Japan. Available at: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/01-NTCIR7-OV-IR4QA-SakaiT.pdf.

Yates, S. (2006). Scaling the tower of Babel Fish: an analysis of the machine translation of legal information. Law Library Journal, 98(3), 481-500.