A preliminary evaluation of metadata records machine translation

Jiangping Chen
Department of Library and Information Sciences, University of North Texas, Denton, Texas

Ren Ding
School of Information Management, Wuhan University, Wuhan, P.R. China

Shan Jiang
The Wuhan Branch of the National Science Library, Chinese Academy of Sciences, Wuhan, P.R. China

Ryan Knudson
Department of Library and Information Sciences, University of North Texas, Denton, Texas

Abstract

Purpose of this paper

The purpose of this study is to evaluate freely available machine translation (MT) services' performance in translating metadata records.

Design/methodology/approach

Randomly selected metadata records were translated from English into Chinese using Google, Bing, and SYSTRAN Machine Translation (MT) systems. These translations were then evaluated using a five point scale for both Fluency and Adequacy. Missing Count (words not translated) and Incorrect Count (words incorrectly translated) were also recorded.

Findings

Concerning both Fluency and Adequacy, Google and Bing's translations of more than 70% of test data received scores equal to or greater than three, representative of 'non-native Chinese' and 'much coverage,' respectively. SYSTRAN scored lowest in both measures. However, these differences were not statistically significant. A Pearson correlation analysis demonstrated a strong relationship (*r=.86)* between Fluency and Adequacy. Missing Count and Incorrect Count strongly correlated with Fluency and Adequacy.

Research limitations/implications

This study was conducted in a specific domain with a small sample size. It is necessary to conduct the evaluation with a larger, more representative test dataset. Also, other language pairs should be evaluated applying similar technologies.

Originality/value

Most existing digital collections can be accessed in English alone. Few digital collections in the United States support multilingual information access (MLIA) that enables users of differing languages to search, browse, recognize and use information in the collections. Human translation is one solution, but it is neither time nor cost effective for most libraries. This study serves as a first step to understand the performance of current MT systems and to design effective and efficient MLIA services for digital collections.

Keywords

Multilingual information access, Cross-language information retrieval, Machine translation, Digital libraries, Globalization

Paper type

Research paper

Introduction

Digital collections contain digital objects in different formats to serve a defined community or set of communities (Schwartz, 2001; Borgman**,** 1999; Arms**,** 2000). Digital collections are broadly defined in this paper and include library online catalogs and collections in various digital libraries. Libraries and museums in the U.S. have developed numerous digital collections in order to preserve scientific, cultural, and heritage materials and to provide convenient access for their users by organizing the objects and representing them through metadata. However, most of these collections can only be accessed in English. Very few digital collections in the United States support multilingual information access (MLIA) that enables users to search, browse, recognize and use information from multilingual digital objects (Gonzalo & Peters, 2004). In the increasingly global knowledge society, libraries and museums seek new ways of engaging communities, providing information access, and disseminating information (Pastore, 2009). Ensuring effective and efficient multi-lingual information access to the metadata for existing collections is a significant first step in a comprehensive globalization strategy for libraries and museums.

Metadata Records Translation (MRT) is the process of converting metadata records describing objects in a digital collection from one language into other languages. It is the necessary first step toward MLIA for a digital collection. To date, most of this work has been done through painstaking and costly human effort. For example, the Children's Digital Library has established a network of translators to carry out the task of metadata translation (Hutchinson *et al.*, 2005). However, manual metadata records translation is not an option for many libraries and museums because of the tremendous time and cost involved. Machine translation (MT), which automatically translates information from one language to one or more languages, or strategies combining MT with human efforts are possible solutions. But there appear to be minimal instances where MT has been implemented, evaluated, or adapted for use with digital collections.

Machine translation (MT) has been a field within Artificial Intelligence for more than 50 years. MT aims to automatically convert text or speech from one language to one or more other languages. The desired translation is one that expresses the exact meaning in the source text with correct syntax (Manning & Schutze, 1999, p.464). MT is difficult because the process involves interpretation of the meaning in the original language and its expression in the target language using correct terminology and syntax. Significant progress has been made in MT in recent years, especially with the dramatic funding support from federal government agencies such as DARPA (US Defense Advanced Research Projects Agency) and NSF (the National Science Foundation), and the large search-engine companies such as Google, Microsoft, and Yahoo!. MT has been widely used in translating queries in various experimental cross-language information retrieval (CLIR) systems with fairly good retrieval performance (Sakai *et al.*, 2008; Chen and Bao, 2009a; He and Wu, 2010). These language tools have been widely used by Web users including librarians (Yates, 2006; Notess, 2008; Chen and Bao, 2009a).

However, MT has not been applied to digital libraries. The 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. have applied MT for cross-language search or metadata records translation (Chen and Bao, 2009b). 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.

What is the performance of current MT systems to metadata records translation? Can libraries and digital libraries apply freely available MT systems or services to translate metadata records so that MLIA can be provided to the users? Little research has been conducted in this area. In the next section, we will discuss related research in order to provide multilingual information access.

Related Research

MT and CLIR are the most relevant research areas to multilingual information access (MLIA). In the past decade, progress in MT research has been driven by several benchmarks funded by US federal and European agencies. The NIST Open MT evaluations [1], the DARPA TIDES and GALE projects [2], and the Workshop on Statistical Machine Translation [3] have provided well-founded experimental frameworks to compare and contrast the performance of MT systems, yielding impressive improvements in translation quality. Performance of various MT strategies has been examined using different evaluation metrics and evaluation tools (Lavie *et al.*, 2004; Vilar *et al.*, 2007; Przybocki *et al.*, 2008; Callison-Burch et al, 2008). However, in these evaluations, MT systems are typically trained and evaluated on translations of news stories, web text or parliament proceedings. It remains unclear how effective current MT technologies are for translating other data genres such as metadata records. Furthermore, these benchmarks evaluate translation quality intrinsically by comparing automatic translations to references produced by humans. In contrast, metadata records translation enables an extrinsic evaluation of MT by directly focusing on its usefulness for a specific application.

MLIA is considered an extension of CLIR (Cross-Language Information Retrieval) which provides users with access to information that is in a different language from the language users use to express their queries (Chen, 2006). In other words, CLIR concerns both translation and retrieval. Research in CLIR has been significantly advanced by three major evaluation forums: a Cross-Language Information Retrieval Track at TREC (Text Retrieval Conference)[4] from 1997-2002; the Cross-Language Evaluation Forum (CLEF) [5] which evaluates many European languages; and the NTCIR Asian Language Evaluation [6] that covers Chinese, Japanese, and Korean. These forums provide CLIR researchers and system developers with infrastructures for developing algorithms, testing systems, and sharing resources. Automatic translation of the queries or documents is a necessary step in CLIR. Various translation strategies for translating queries have been explored (Chen, 2006; Sakai *et al.*, 2008; Oard *et al.*, 2008). CLIR for Web search has been available since 2005 when Yahoo launched a CLIR search interface option for German and French sites (Sterling, 2007). In May 2007, Google launched a “Translated Search” feature as part of its Google Language Tools [7]. The launch of the cross-language search interfaces by Yahoo and Google signified the transition from CLIR research to practice. These services were welcomed by Web users because they provided access to information written in foreign languages (Chen and Bao, 2009a). However, CLIR and MT technologies and services, such as those provided by Google and others, have not been applied to enable searching of specific digital collections in the United States.

Metadata records translation is the necessary and important first step in providing CLIR or MLIA to digital collections. 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, 2009b). Even though manual metadata records translation (the use of human translators) can be conducted through collaborating with organizations in other countries, it is expensive and time-consuming.

Researchers are also experimenting with translation processes that integrate MT with human efforts. For example, the Collaborative Translation model attempts to use MT as a bridge to facilitate monolingual users translating in real applications (Bederson, 2009).

In this paper, we report a study evaluating machine translation (MT) performance in translating metadata records. We used MT services provided by Google, Bing, and SYSTRAN to translate 48 metadata records from English to Chinese. The next section will describe the research design of this study.

Research Design

This study has three main purposes: (1) to achieve a general understanding of the MT performance of current MT services that are freely available on the Internet; (2) to investigate and compare evaluation metrics for metadata records translation; and (3) to identify strategies for metadata records translation for digital libraries.

Test Data

The test data were acquired from The Portal to Texas History [8]. The Portal to Texas History is a digital library that provides a variety of materials about Texas history to a worldwide audience. Its users consist of over 115,000 people worldwide every month. For our study, 1000 records were acquired from the UNT

Libraries. We randomly selected 48 out of the 1,000 records as the original metadata records for the machine translators. The metadata records are in Dublin Core format. One of the records is shown in Table I. These sample metadata records describe images, texts, or other digital objects in the digital library.

Table I. A Sample of the Metadata Records

The Original Metadata Record	The Filtered Metadata Record
ID: metapth46004 **publisher:** Abilene Christian College **description:** Catalog describes the governance, history, course offerings, and campus life of Abilene Christian College in Abilene, Texas. **format:** 60 p. : ill. ; 23 cm. **language:** eng **format:** text **type:** text_book **creator:** Abilene Christian College **coverage:** United States - Texas - Taylor County - Abilene **coverage:** new-sou **date:** 1969-03 **title:** Catalog of Abilene Christian College, 1969-1970 **title:** A catalog of general information and courses of instruction, Abilene Christian College, Abilene, Texas, 1969-1970 **title:** Bulletin, Abilene Christian College, Volume 53, Number 3, March 1969 identifier: oclc: 36047647 **subject:** Education - Colleges and Universities **subject:** catalogues **subject:** Abilene Christian College --Curricula--Periodicals. **subject:** Abilene Christian University -- Curricula -- Periodicals **coverage:** 1969-1970 **identifier:** ark: ark:/67531/metapth46004 **identifier:** http://texashistory.unt.edu/ark:/67531/metapth46004/	**publisher:** Abilene Christian College **description:** Catalog describes the governance, history, course offerings, and campus life of Abilene Christian College in Abilene, Texas. **creator:** Abilene Christian College **coverage:** United States - Texas - Taylor County - Abilene **coverage:** new-sou **title:** Catalog of Abilene Christian College, 1969-1970 **title:** A catalog of general information and courses of instruction, Abilene Christian College, Abilene, Texas, 1969-1970 **title:** Bulletin, Abilene Christian College, Volume 53, Number 3, March 1969 identifier: oclc: 36047647 **subject:** Education - Colleges and Universities **subject:** catalogues **subject:** Abilene Christian College --Curricula--Periodicals. **subject:** Abilene Christian University -- Curricula -- Periodicals

MT System Selection

The next step involved the selection of MT systems for the evaluation. There are several well-known, freely available MT systems on the Internet, such as Google, Bing, Yahoo, Wordlingo, and SYSTRAN. Our preliminary test showed that the Chinese translation results produced by Yahoo, Wordlingo and SYSTRAN were nearly identical. Therefore, we chose to use Google, Bing, and SYSTRAN online translation systems in the following test. Before utilizing the MT systems for the task at hand, we pre-processed the metadata records so that only elements appropriate for MT were translated. These elements included publisher, description, creator, title, and subject, as presented in the second column of Table I.

We chose to use the human MT evaluation metrics: Fluency and Adequacy (LDC, 2005) to evaluate the quality of the MT services. Fluency refers to the degree to which the target is well formed according to the rules of a particular language (LDC, 2005), in this case - Chinese. Adequacy refers to the degree to which information present in the original is represented in the translation. The evaluation assigns a score to each metadata record on a scale from 1 – 5 based on the coding scheme in Table II.

Table II. Fluency and Adequacy Measures

Scale	Fluency	Adequacy
5	Flawless Chinese	All
4	Good Chinese	Most
3	Non-native Chinese	Much
2	Disfluent Chinese	Little
1	Incomprehensible	None

In addition to evaluating Fluency and Adequacy, we also asked the evaluators to record Missing Count (the number of words/phrases that are not translated) and Incorrect Count (the number of words/phrases that are incorrectly translated) for each translation. These two measures were considered more objective measures for MT quality. Together they provided a measure of Lexicon Error in the translation (Yates, 2006).

Human Evaluators

Two human evaluators (A and B) assessed the translation results of each of the systems for the 48 original records. Both of the evaluators are bilingual in English and Chinese. They are graduate students already having completed a master's degree. An average of 10 minutes was spent on assessing the Fluency, Adequacy, Missing Count, and Incorrect Count of each of the Chinese metadata records translated by the three machine translators. The next section is the report on the evaluation results of these 48 metadata records.

Evaluation Results

The evaluations of the MT translation of the 48 metadata records were analyzed in order to accomplish the purposes of this study. We present the results on: 1) Inter-coder reliability; 2) MT performance of the three systems; and 3) Associations among the four measures.

Inter-coder Reliability

To measure the reliability of the inter-coder analysis, we treated the evaluation as a coding process. The distribution of our sample data was a perfect normal distribution. We therefore applied Krippendorff's Alpha-Reliability (Krippendorf, 2007) to measure the inter-coder reliability. Table III. reports the results. Without surprise, the Alpha-Reliability was very low, which indicated the two evaluators had wide disagreement in their assignment of scores to the sample data, e.g., the disparity of the Fluency judgment on Google Translation (α= 0.039).

Table III. The Result of Alpha – Reliability

	Fluency			Adequacy		
System	Bing	Google	SYSTRAN	Bing	Google	SYSTRAN
Krippendorff's α	0.175	0.039	0.135	0.211	0.589	0.357

However, when we examined the actual results for Google on Fluency, we found that the two evaluators did not display as dramatic a difference in terms of cumulative percentage. Their frequency distribution (see Figure 1) was quite similar. Each evaluator assigned 3 or above to Google on Fluency for more than 85% of the test records.

Comparing the assessments of evaluators A and B, we found that both assigned the Fluency and Adequacy with values of 3 or 4 to the majority of the translations produced by the three MT systems. The proportion of high scores in evaluator B's assessment was lower than that of A's assessment, which indicated that evaluator B is more inclined to give a lower score to machine translation results.

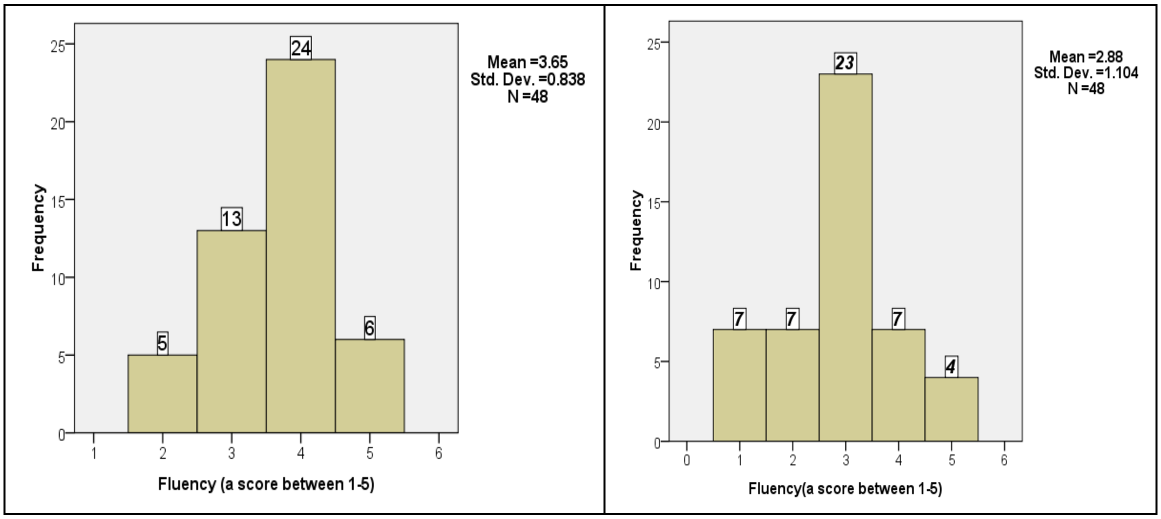

Figure 1. The Histograms of Evaluator A (Left) and Evaluator B (Right) on Fluency for Google Translation

Machine Translation Performance

One of the main purposes of this study was to achieve a general understanding of the performance of existing MT services that are freely available on the Internet. We did not perform normalization of the evaluation scores, as done by other human MT evaluations (Collison-Burch *et al.,* 2007). We believed reporting in the original scale was more straightforward for the readers.

We combined the evaluation results of the two evaluators and used their average as the final score for each record for the four measures respectively. Table IV presents the mean and standard deviation of the three systems. It shows all three systems achieved a mean score above 3.0. SYSTRAN received the lowest scores among the three systems.

Table IV. Descriptive Statistics

MT System	**Fluency**		**Adequacy**	
	Mean	**Std. Deviation**	**Mean**	**Std. Deviation**
Bing translation	3.30	0.84	3.30	0.80
Google translation	3.26	0.88	3.29	0.84
SYSTRAN translation	3.08	0.96	2.97	0.98

Comparison of Fluency Scores

Figure 2 is the frequency distribution of fluency scores of the three systems. Both Bing and Google's translations of more than 70% of test data received a score equal to or higher than 3, which coincides with "non-native Chinese".

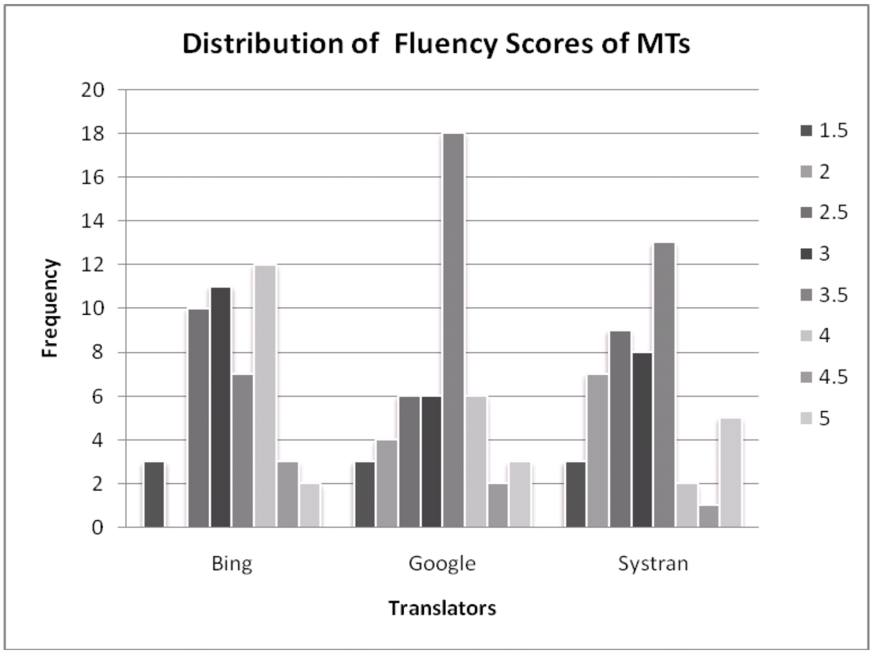

Figure 2. Frequency Distribution of Fluency Scores

Comparison of Adequacy Scores

Figure 3 is the frequency distribution of Adequacy scores of the three systems. Bing and Google's translations of more than 70% of test data received a score equal to or higher than 3, which coincides with "Much Coverage". Again, SYSTRAN received the lowest scores among the three systems.

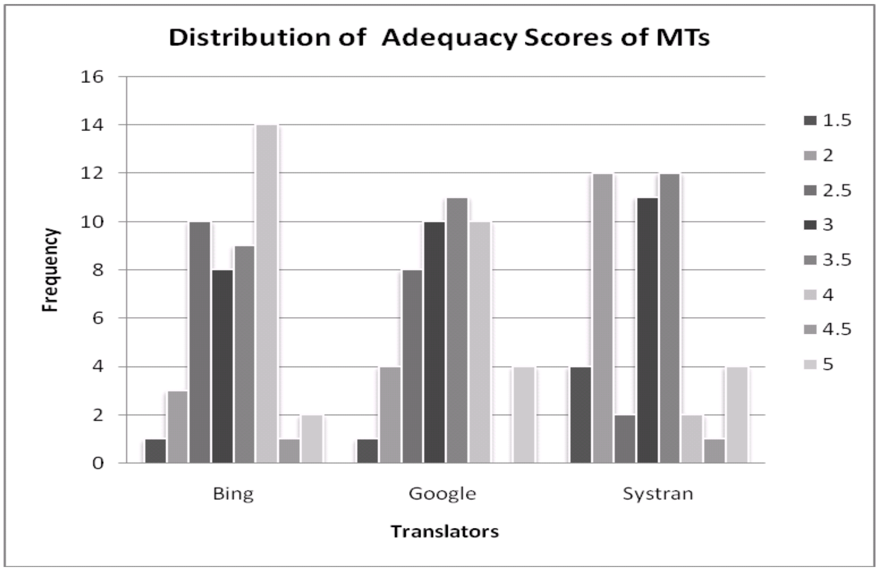

Figure 3. Frequency Distribution of Adequacy Scores

Are the differences among the three systems significant in terms of Fluency and Adequacy? A one-way ANOVA analysis of Fluency and Adequacy shows there are no significant differences among the three systems with respect to Fluency or Adequacy.

Missing Count and Incorrect Count

We then examined the results for Missing Count and Incorrect Count. On average, the three MT systems made 3-5 translation mistakes, as illustrated in Table V. Among the three systems, Systran made more translation mistakes than Bing and Google. Bing had the fewest average Incorrect Count, while Google had the fewest average Missing Count.

Table V. Results of Missing Count and Incorrect Count

	Bing Translation		**Google Translation**		**Systran Translation**	
Measure	Total Mistakes	Average	Total Mistakes	Average	Total Mistakes	Average
Missing Count	259	2.70	268	2.80	335	3.49
Incorrect Count	122	1.27	103	1.07	128	1.33
Total	381	3.97	371	3.69	463	4.8

Table VI presents two examples of Missing Count and Incorrect Count. These examples were translated by Bing. The results show that translation mistakes, even though only a few, adversely affect the meaning of the translated texts. In the associate analysis below, we found that both Missing Count and Incorrect Count have a strong relationship with Fluency or Adequacy, which is not surprising.

Table VI. Examples for Missing Translation and Incorrect Translation

Example 1：

Description: Black and white photograph of two faculty members and one student at Eastfield College registration. Man with glasses is history professor, Tim Hughes, the man without glasses is psychology Professor, Adolf. The professors are seated. The student is standing beside.

Bing Translation Result：

描述：两个教员和一位学生在大学Eastfield注册的黑白照片。戴眼镜的人是历史系教授，Tim Hughes,没有玻璃的男子是心理学教授，Adolf。教授们都坐着。站学生。

Incorrect Translation：“glasses” was falsely translated into“玻璃”

Missing Translation：“The student is standing beside ”missing“站在旁边”

Example 2：

Description: Black and white photograph of Dr. Jan LeCroy, Chancellor of the Dallas County Community College District blowing out candles on his birthday cake. Pattie Powell is seated. The lady holding the cake is unidentified as is the gentlemen in the background.

Bing Translation Result：

描述：博士Jan LeCroy，达拉斯县社区学院区校长， 吹出他的生日蛋糕上的蜡烛的黑白照片。 坐在选择鲍威尔。 这位女士担任蛋糕和这位在后台一样，是未经确认的。

Incorrect Translation：“Pattie Powell” was falsely translated into “选择鲍威尔”; “holding” was falsely translated into “担任”；“blowing” was falsely translated into “吹出”

Missing Translation：“gentleman” was not translated

Correlation Among the Measures

We also conducted correlation analysis among the four measures: Fluency, Adequacy, Missing Count, and Incorrect Count. Table VII shows the correlation analysis results.

Table VII. Pearson Correlation among the Four Measures

	Fluency	Adequacy	Missing Count	Incorrect Count
Fluency	1	.860**	-.537**	-.460**
Adequacy	.860**	1	-.507**	-.489**
Missing Count	-.537**	-.507**	1	.328**
Incorrect Count	-.460**	-.489**	.328**	1

The above results verify the strong relationship (*r = .86)* between Fluency and Adequacy, as demonstrated in the literature on human evaluation of MT (Collison-Burch *et al.*, 2008).

Both Missing Count and Incorrect Count have a strong relationship with Fluency or Adequacy. Is it possible that reducing the number of missing terms and incorrect terms in the metadata records translation would improve Fluency and Adequacy scores? This will be investigated further in a future study.

Discussion

MT Performance

Based on our evaluation results, the evaluators considered Bing as the best MT system in terms of Fluency and Adequacy, followed by Google, then SYSTRAN. However, statistically speaking, there was no significant difference among these three MT systems. The evaluators also made the following comments on the three systems:

- Bing has a relatively comprehensive translation function. It performs better than Google and SYSTRAN, especially in translating short sentences with fewer than 10 words. The Chinese translations from Bing most resembled what a native Chinese speaker would consider well formed, although it did not translate accurately in some situations, e.g., translating persons' names, the periodical volume issues, dates, locations, and a few situations in which word order is of a non-native quality.

- The performance of Google translator is medium. Though its translation results have less missing words, Google does not perform well when dealing with the ambiguity of some words and sometimes produces redundant words in the translation.

- SYSTRAN's translation has many problems related to both the structure of the sentences and the selection of appropriate words. It is unable to choose the most appropriate translation for words based on the context. In addition, a small number of domain-specific terms were not recognized.

MT Measures

We may need to change or consider new measures due to the high correlation between fluency and adequacy. Measures such as "Preference", or "human-targeted Translation Edit Rate (HTER)" may be worth testing for metadata records translation as well (NIST, 2010).

Translation Strategies for Metadata Records

In this study, the four evaluation measures displayed a strong correlation between them, which may indicate that, if a system could take a strategy that significantly reduced the number of translation errors, it could improve the fluency and adequacy of machine translation.

Multi-engine machine translation (MEMT), which combines the best 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 (Nirenburg & Frederlcing, 1994). Different approaches have been proposed and experiments conducted to combine results from multiple systems (Nirenburg & Frederlcing, 1994; Tidhar & Kiissner, 2000; Akiba et al, 2002; Callison-Burch & Flournoy, 2001; Nomoto, 2004; Jayaraman & Lavie, 2005; Matusov et al, 2006; Rosti, et al, 2007; Chen et al, 2007). MEMT has the potential to achieve significantly better performance than any single MT system (Callison-Burch, et al, 2008). Evaluation needs to be conducted on a larger scale and in an application-driven fashion for these multi-engine MT models. Although there was no significant difference between the three MT systems under consideration, the types of errors across the three were not identical. MEMT has the potential to be applied to metadata records translation due to the availability of several free MT services.

Conclusion

Based on this small-scale evaluation, two of the MT systems achieved a performance equal to or above “non-native Chinese” without any training on the material. Each system has unique characteristics and strengths, see Discussion: *MT Performance* above, that may complement each other and reduce the number of errors if the translation results could be appropriately combined. Strong correlations can be found among the four measures used in this study. It may indicate that if an MT strategy that significantly decreased the number of missing terms and incorrectly translated terms were employed, improved Fluency and Adequacy would result.

This study was conducted in a specific domain with a small sample size. It is necessary to conduct the evaluation with a larger, more representative test dataset. Also, more in-depth analysis based on a larger sample dataset will be valuable. For example, we are interested in examining translation performance of individual metadata fields, such as Publisher, Creator, and Subject, which may help us to develop more effective MT strategies for these fields.

We recently obtained funding from the Institute of Museum and Library Services (IMLS) [9] to extend this study and to conduct the above analysis. Our project will evaluate the extent to which current machine translation technologies generate adequate translation for metadata records and identify the most effective metadata records translation strategies for digital collections. More evaluators will be recruited and the analysis will also include the examination of the disagreement among the evaluators. We will also explore automatic evaluation measures like BLEU and the performance of current MT systems on other language pairs, and investigate the effectiveness of MEMT on metadata records translation.

In summary, digital libraries can take advantage of several ways to use existing automatic translation systems to implement multilingual information access: (1) Apply query translation based cross-language information retrieval technology to find digital objects, and then use Google or Bing translation interface to dynamically translate the retrieved metadata records; (2) Apply multi-engine machine translation technology to effectively combine results from multiple MT systems, and translate all metadata records into desired languages for the retrieval and display of digital objects; (3) Develop thesauri or multilingual directories specific to the digital libraries, and then use (1) or (2). These will all be investigated in future research.

Acknowledgement

This study is partially supported by the Institute of Museum and Library Services (IMLS) grant LG-06-10-0162-10.

About the Authors

Jiangping Chen is an Associate Professor at the Department of Library and Information Sciences of the University of North Texas (UNT). She earned a Ph.D. from Syracuse University, an M.S. from the Library of Chinese Academy of Sciences, and a B.S. from Wuhan University, China. Jiangping teaches and

conducts research in Multilingual Information Access and Digital Libraries. She can be reached through email at Jiangping.Chen@unt.edu.

Ren Ding is a doctoral student at the School of Information Management, Wuhan University, Wuhan, China. Her research interest focuses on information behavior and information searching. Ren can be reached through email at Dingren114400@126.com.

Shan Jiang is an academic librarian at the Wuhan Branch of the National Science Library, Wuhan, China. He holds a BA in Microelectronics and a MA in Materials Physics and Chemistry from Shan Dong University (China). He works in the Information Research Department and focuses on the research of academic information and policy in the area of Advanced Manufacturing and Materials. Shan can be reached through email at dexter.jiang@gmail.com.

Ryan Charles Knudson is a teaching and research assistant at the College of Information of the University of North Texas (UNT). He holds a BA in English from Texas Woman's University, an MA in linguistics, a graduate academic certificate in TESOL, an MS in information science from UNT, and is currently working toward a Ph.D. in interdisciplinary information science at UNT. He has served for over two years as an editor for academic manuscripts. Ryan can be reached through email at ryanknudson@yahoo.com.

References

Akiba, Y., Watanabe, T. and Sumita, E. (2002), "Using language and translation models to select the best among outputs from multiple mt systems", *International Conference on Computational Linguistics, 19th,* Taipei, Taiwan, 24 August-1 September 2002, Assocation for Computational Linguistics, Available (ACL) http://www.aclweb.org/anthology/C/C02/C02-1076.pdf

Arms, W.Y. (2000), "Automated digital libraries: How effectively can computers be used for the skilled tasks of professional librarianship?", *D-Lib Magazine*, Vol 6 No 7/8. Available (D-LIB) http://www.dlib.org/dlib/july00/arms/07arms.html

Bederson, B. (2009), "Translation by Iterative Collaboration between Monolingual Users", *Google Tech Talk*, 24 September 2009. Available http://www.youtube.com/watch?v=rMj1PnrZphg

Borgman, C. L. (1999), "Discussion: What are Digital Libraries? Competing Visions", *Information Processing and Management,* Vol 35 No 3, pp. 227-243

Callison-Burch, C., and Flournoy, R. S. (2001), "A program for automatically selecting the best output from multiple machine translation engines", *Machine Translation Summit VIII,* Santiago de Compostela, Spain, 18-23 September 2001, pp. 63-66. Available (IAMT) http://www.mt-archive.info/MTS-2001-Callison.pdf

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007), "(Meta-) evaluation of machine translation", *Proceedings of the Second Workshop on Statistical Machine Translation,* Prague, Czech Republic, 2007, Association for Computational Linguistics, pp. 136-158

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008), "Further meta-evaluation of machine translation", *Proceedings of the Third Workshop on Statistical Machine Translation,* Prague, Czech Republic, 2008, Association for Computational Lingustics, pp. 70-106

Chen, J. (2006), "A lexical knowledge base approach for English-Chinese cross-language information retrieval", *Journal of the American Society for Information Science and Technology*, Vol 57 No 2, pp. 233-243

Chen, Y., Eisele, A., Federmann, C., Hasler, E., Jellinghaus, M. and Theison, S. (2007), "Multi-engine machine translation with an open-source decoder for statistical machine translation", Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007, Association for Computational Linguistics, Morristown, New Jersey, USA, pp. 193-196

Chen, J. and Bao, Y. (2009a), "Information access across languages on the web: From search engines to digital libraries", *72nd ASIS&T Annual Conference,* Vancouver, British Columbia, Canada, 6-11 November 2009. American Society for Information Science and Technology. Silver Spring, Maryland. Available (ASIS) http://www.asis.org/Conferences/AM09/open-proceedings/openpage.html

Chen, J. and Bao, Y. (2009b), "Cross-language search: The case of Google Language Tools", *First Monday,* Vol 14 No 3. Available (*First Monday*) http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2335/2116

Gonzalo J, Peters C. (2004), "Comparative evaluation of multilingual information access systems", *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Trondheim, Norway, August 2003, Springer, Berlin, pp. 1-6

He, D. and Wu, D. (2010), "Exploring the further integration of machine translation in multilingual information access", *Online Proceedings of 2010 iConference*, Urbana-Champaign, Illinois, USA, 3-6

February 2010, Available (University of Illinois) http://www.ideals.illinois.edu/bitstream/handle/2142/14916/iConferencev6-wu.pdf?sequence=2

Hutchinson H. B., Rose, A., Bederson, B. B., Weeks, A. C., and Druin, A. (2005), "The International Children's Digital Library: A case study in designing for a multi-lingual, multi-cultural, multi-generational audience", *Information Technology and Libraries,* Vol 24 Num 1, pp. 4-13

Jayaraman, S. and Lavie, A. (2005), "Multi-Engine Machine Translation Guided by Explicit Word Matching", *Conference of the European Association of Machine Translation , 10th*, Budapest, Hungary, 30-31 May 2005, Association for Computational Linguistics, pp. 143-152. Available (EAMT) http://www.mt-archive.info/EAMT-2005-Jayaraman.pdf

Krippendorf K. (2007), "Computing Krippendorff's alpha reliability" Annenberg School for Communication Departmental Papers (2007 ASC), [2010-02-23]. Available http://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Lavie, A., Sagae, K., and Jayaraman, S. (2004), "The Significance of recall in automatic metrics for mt evaluation", *Conference of the Association for Machine Translation in the Americas, 6th,* Washington, D.C., USA, 28 September-2 October 2004, Springer, Berlin, pp. 134-143

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

Manning, C., and Schutze, H. (1999), *Foundation of Statistical Natural Language Processing,* MIT Press, Cambridge, MA.

Matusov, E., Ueffing, N. & Ney, H. (2006), "*Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment*", *Conference of the European Chapter of the Association for Computational Linguistics, 11th*, Trento, Italy, 3-7 April 2006, Association of Computational Linguistics, East Stroudsburg, pp. 33-44

Nirenburg, S. and Frederking, R. (1994), "Toward multi-engine machine translation", Workshop on Human Language Technology, Plainsboro, New Jersey, USA, 8-11 March 1994, Association for Computational Linguistics, pp. 147-151. Available (ACL) http://www.aclweb.org/anthology/H/H94/H94-1026.pdf

NIST (2010), "The NIST metrics for machine translation 2010 Challenge: Evaluation plan", Available http://www.nist.gov/itl/iad/mig/upload/NISTMetricsMaTr10EvalPlan.pdf

Nomoto, T. (2004), "Multi-engine machine translation with voted language model", *Annual Meeting-Association for Computational Linguistics, 42nd,* Barcelona, Spain, 21-26 July, 2004, Association for Computational Linguistics, pp. 494-501. Available (ACL) http://www.aclweb.org/anthology/P/P04/P04-1063.pdf

Notess, G. R. (2008), "Multilingual Searching: Search Engine Language Tools", *Information Today, Inc.* , Vol 32 No 3, pp. 40-42. Available http://www.infotoday.com/online/may08/Notess.shtml

Oard, D.W., He, D. and Wang, J. (2008), "User-assisted query translation for interactive cross-language information retrieval", *Information Processing and Management,* Vol 44 No 1, pp. 181-211

Pastore, E. (2009), "The Future of Museums and Libraries: A Discussion Guide",

(IMLS-2009-RES-02). *Institute of Museum and Library Services*. Washington, D.C., Available (IMLS) http://www.imls.gov/pdf/DiscussionGuide.pdf

Przybocki, M., Peterson, K., and Bronsart, S. (2008), "Translation adequacy and preference evaluation tool (TAP-ET)", *International Language Resources and Evaluation Conference, 6th,* Marrakech, Morocco, 28-30 May 2008, European Language Resources Association, pp. 1175-1182. Available (ELRA) http://www.lrec-conf.org/proceedings/lrec2008/pdf/299_paper.pdf

Rosti, A.V.I., Ayan, N.F., Xiang, B., Matsoukas, S., Schwartz, R. and Dorr, B.J. (2007), "Combining outputs from multiple machine translation systems", *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA, 23-27 April 2007, Association for Computational Linguistics, pp. 228-235

Sakai, T., Kando, N., Lin, C.J., Mitamura, T., Shima, H., Ji, D., Chen, K.H. and Nyberg, E. (2008), "Overview of the NTCIR-7ACLIA IR4QA task", *Online Proceedings of NTCIR-7 Workshop Meeting*, Tokyo, Japan, 16-19 December 2008. pp. 77-114. Available (NTCIR) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/01-NTCIR7-OV-IR4QA-SakaiT.pdf

Schwartz, C. (2000), "Digital Libraries: An Overview", *Journal of Academic Librarianship*, Vol 26 No 4, pp. 385-

Sterling, G. (2007), "Google launches 'cross-language information retrieval (CLIR)", Search Engine Land, 24 May 2007, Available http://searchengineland.com/google-launches-cross-language-information-retrieval-clir-11296

Tidhar, D. and Küssner, U. (2000), "Learning to select a good translation", *International Conference on Computational Linguistics, 18th,* Saarbrücken, Germany, 31 July-4 August 2000, Association for Computational Linguistics, pp. 843-849. Available (ACL) http://www.aclweb.org/anthology/C/C00/C00-2122.pdf

Vilar, D., Leusch, G., Ney, H., and Banchs, R. E. (2007), "Human Evaluation of Machine Translation Through Binary System Comparisons", *Second Workshop on Statistical Machine Translation,* Prague, Czech Republic, 23 June 2007, Association for Computational Linguistics, pp. 96-103

Yates, S. (2006), "Scaling the tower of Babel Fish: An analysis of the machine translation of legal information" *Law Library Journal*, Vol 98 Num 3, pp. 481-500

Web sites

[1] http://www.itl.nist.gov/iad/mig/tests/mt/

[2] http://www.darpa.mil/ipto/programs/gale/gale.asp

[3] http://www.statmt.org/wmt09/

[4] http://trec.nist.gov/

[5] http://clef-campaign.org/

[6] http://research.nii.ac.jp/ntcir/

[7] http://www.google.com/language_tools

[8] http://texashistory.unt.edu/

[9] http://www.imls.gov