Metadata Records Machine Translation Combining Multi Engine Outputs With Limited Parallel Data

Brenda Reyes Ayala
Department of Information Science, University of North Texas, 1155 Union circle #311068, Denton, TX 76203, USA. E-mail: Brenda.Reyes@unt.edu

Ryan Knudson
Department of Information Science, University of North Texas, 1155 Union circle #311068, Denton, TX 76203, USA. E-mail: RyanKnudson@yahoo.com

Jiangping Chen
Department of Information Science, University of North Texas, 1155 Union circle #311068, Denton, TX 76203, USA. E-mail: Jiangping.Chen@unt.edu

Gaohui Cao
School of Information Management, Central China Normal University, No. 152, Luoyu Road, Wuhan, Hubei 430079, P.R. China. E-mail: ghcao@mail.ccnu.edu.cn

Xinyue Wang
Intelligent Information Access Lab, University of North Texas, 1155 Union circle #311068, Denton, TX 76203, USA. E-mail: XinyueWang@my.unt.edu

One way to facilitate Multilingual Information Access (MLIA) for digital libraries is to generate multilingual metadata records by applying Machine Translation (MT) techniques. Current online MT services are available and affordable, but are not always effective for creating multilingual metadata records. In this study, we implemented 3 different MT strategies and evaluated their performance when translating English metadata records to Chinese and Spanish. These strategies included combining MT results from 3 online MT systems (Google, Bing, and Yahoo!) with and without additional linguistic resources, such as manually-generated parallel corpora, and metadata records in the two target languages obtained from international partners. The opensource statistical MT platform Moses was applied to design and implement the three translation strategies. Human evaluation of the MT results using adequacy and fluency demonstrated that two of the strategies produced higher quality translations than individual online MT systems for both languages. Especially, adding small, manually-generated parallel corpora of metadata records significantly improved translation performance. Our study suggested an effective and efficient MT approach for providing multilingual services for digital collections.

Introduction

Digital metadata, or metadata records, are bibliographical information generated to describe a digital object such as a document, an e-book, an image, or a combination of digital materials. For example, the ACM Digital Library organizes its digital objects through metadata records that contain elements such as *author*, *abstract*, *cited by*, *references*, and *index terms*. Each paper in the digital library can be identified and accessed through search terms contained in its metadata records.

Digital libraries are looking for new ways to expand their user groups and provide new services to broader communities (Diekema, 2012; Purday, 2012). One method of achieving this is to provide Multilingual Information Access (MLIA), which enables users to search, browse, discover, and use information in their own native languages (Peters, Braschler, & Clough, 2012, p. 5; Chen, 2016, p. 15). MLIA was considered important because it could allow the use of valuable digital resources by those who do not understand

Received July 12, 2016; revised May 9, 2017; accepted June 7, 2017

2017 ASIS&T • Published online 19 September 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23925

the original languages of the digital objects. These users could be immigrants, foreign travelers, or students. Most digital collections that allow MLIA have done so by applying human translation of the objects to their metadata records. For example, digital libraries such as the International Children's Digital Library (International Children's Digital Library Foundation, n.d.) and the World Digital Library (Library of Congress, n.d.) have implemented MLIA using human translation. However, human translation is expensive and time-consuming, and these factors may prohibit smaller libraries with less funding from implementing MLIA for their collections.

There were a considerable number of studies and research projects that explored multilingual metadata and alternative approaches to MLIA, such as describing digital images using bilingual taxonomies (Ménard, 2012), mapping vocabulary in different languages (Matusiak, Meng, Barczyk, & Shih, 2015), or combining machine translation with domain-specific lexicons (Jones, Fantino, Newman, & Zhang, 2008). Machine translation (MT), which automatically translates the metadata records from one language to other languages, has also been used in Cross-Language Information Retrieval (CLIR) experiments (Sakai et al., 2008; Oard, He, & Wang, 2008). While many online MT systems are available for use, they are not always sufficient for producing quality MT metadata records to facilitate MLIA in digital collections. Chen, Ding, Jiang, and Knudson (2012) experimented with three online MT systems that translated metadata records and found that MT performance of these records could stand improvement.

The advancement of MT technologies, especially the release of open-source MT platforms built on up-to-date MT approaches, allows digital library communities to develop their own MT systems without going through lengthy system development. The purpose of this study was to explore methodologies for developing effective and efficient MT systems for translating English metadata records into other languages.

Related Literature

There are various approaches to MT, such as direct translation, rule-based MT, and corpus-based MT, each with relative strengths and weaknesses (Tripathi & Sarkhel, 2010). With the advancement of computing technologies and availability of large-scale comparative corpora and parallel corpora, corpus-based MT has been widely explored. Corpus-based MT can be further categorized into example-based MT, statistical MT, and deep-learning-based MT (Zhang & Zhong, 2016, p. 92). In the past decades, statistical MT has been explored extensively with funding from the US government and the European Union. As a result, large Internet companies such as Google and Microsoft have launched online MT services based on the statistical MT approach. These services have helped web users to overcome language barriers, and to understand and use more information resources (Gaspari, 2004; Chen & Bao, 2009). Furthermore, some MT research groups have released their MT systems as open-source, which allows the public and other MT researchers to conduct research without building an MT system from scratch. Moses, the software used in this study, is one such open-source platform for statistical MT (Koehn et al., 2007).

One of the strategies for MT is termed Multi-Engine Machine Translation, or MEMT. This is the practice of utilizing more than one MT processor and combining the translation results into a single output or a ranked list of best candidate outputs, which has proven effective (Heafield & Lavie, 2010; Roxas et al., 2008; Nirenburg & Frederlcing, 1994). MEMT enables researchers to exploit the strengths of certain MT approaches while minimizing their weaknesses in an attempt to produce higher-quality translations. This is particularly useful when source language documents are not from a small, limited domain. Ren and Shi (2002) used MEMT by combining four MT processors to improve the success rate and quality of dialog MT. In this study, we applied the idea of MEMT by combining results from multiple online MT services.

We also evaluated MT results on metadata records from the MEMT strategies and compared them with those from online MT services. The topic of metadata evaluation has received considerable attention (Hillman, 2008; Robertson, 2005). But much of the literature on metadata evaluation only considered monolingual metadata. Our evaluation, however, focused on the quality of MT, not the quality of the metadata. Evaluation of MT has long been considered an important task, and has been addressed in numerous studies and publications (Hovy, King, & Popescu-Belis, 2002; Guzmán, Joty, Màrquez, & Nakov, 2015). Also, there have been a number of MT evaluation forums, workshops, or campaigns, such as those funded by the US government and the European Union. In these evaluations, MT systems were typically trained and evaluated on translations of news stories, web text, or parliamentary proceedings. It remains unclear how effective current MT technologies are when translating other types of data such as metadata records.

In a review of usability studies of digital libraries, Chowdhury, Landoni, and Gibb (2006) found that language and cultural issues impact the usability of a digital library. Metadata records are different from most full-text because they contain different elements, some of which are segments or short phrases. For example, a *subject* element may consist of one or more words. In many respects, metadata records can be considered structured or semistructured data. The nature of metadata records may make them difficult to translate, much like natural dialogue (Ren & Shi, 2002).

MT results can be evaluated manually or automatically. Two measures called *fluency* and *adequacy* have been frequently used for various human evaluations of MT (Linguistic Data Consortium, 2005; Lommel et al., 2014). They have been used in some MT evaluations such as OpenMT, TIDES, and some of the Statistical MT workshops (Callison-Burch, Fordyce, Koehn, Monz, & Schroeder, 2007). Other measures, such as a multidimensional quality metric based on an operational definition of accuracy and fluency were proposed (Lommel et al., 2014). Automatic measures such as BLEU and METEOR have been widely used in MT evaluation (Callison-Burch et al., 2007)

In this study, we conducted MEMT experiments and human evaluation of these results using fluency and adequacy measures. The next section describes our design of the MEMT experiments and subsequent human evaluation of the MT results.

Methods

This study consisted of multiple stages in order to investigate effective MT of metadata records using MEMT strategies. The specific research question was:

> Which MEMT strategy can achieve better performance on metadata records?

To answer the above research question, we first extracted 2,010 English metadata records at random from two digital collections. The records' contents were then preprocessed so only six elements were kept for each record for MT. These records were then translated into Spanish and Chinese by three online MT services: Google Translate, Bing Translator, and Yahoo! BabelFish. After that, we conducted MT experiments using Moses and applied three different MEMT strategies (Moses Statistical Machine Translation System Version 3.0, 2016). The translation quality of all six sets of results was then evaluated in terms of adequacy and fluency (Linguistic Data Consortium, 2005) by evaluators who were native speakers of Spanish or Chinese. The following sections describe each stage in more detail.

Metadata Records Extraction, MT Using Free Online MT Services, and Reference Translation Generation

We randomly extracted 2,010 nonduplicate metadata records from two different digital collections: The UNT Library Catalog (University of North Texas Libraries, n.d.), and the Portal to Texas History (University of North Texas Libraries, 2016). These two digital collections used different metadata formats: the UNT Library Catalog used the MARC format, while the Portal to Texas History used the Dublin Core format. We wrote a computer program and converted the UNT Library Catalog records from MARC format to Dublin Core format.

Each metadata record contained more than 10 elements for each digital object. For our purposes, we selected and kept only six elements, including *title*, *creator*, *subject*, *description*, *publisher*, and *coverage*, due to their relevance to users' search and retrieval behavior (Chen, 2016, p. 91). Furthermore, some of these elements also provided challenges to MT systems that would be interesting to explore. For example, creator, coverage, and publisher usually contain named entities including organizational, geographical, or personal names. In this study we did not provide special treatment to named entities. However, we were interested in understanding to what degree the online MT systems could correctly translate named entities. Not every record contained all six elements, but each contained at least title, creator, and subject. A sample metadata record that has been processed is presented in Table 1.

TABLE 1. A sample metadata record.

Element	Example
Title	"Catalog of Abilene Christian College, 1969–1970"
Creator	"Abilene Christian College"
Subject	"Education - Colleges and Universities"
Description	"Catalog describes the governance, history, course offerings, and campus life of Abilene Christian College in Abilene, Texas. …"
Publisher	"Abilene Christian College"
Coverage	"United States - Texas - Taylor County – Abilene"

Each of the 2,010 English records was submitted to three MT systems—Google Translate (Google, 2016), Microsoft Bing Translator (Microsoft, 2016), and Yahoo! BabelFish (no longer active) through their application program interfaces for Spanish and Chinese translations. Spanish and Chinese were chosen as the target languages for two reasons: 1) they are the most spoken languages on the Internet (Internet World Stats, 2016) and the native languages for many US immigrants; and 2) from a linguistics perspective, these two languages were considered quite different. English and Spanish share the same characters and usually named entities such as personal names don't require translation. In contrast, the basic unit of Chinese is a character called an ideograph. Chinese words can be composed of one or more Chinese characters, and there are no word boundaries in a Chinese sentence. Choosing these two very different languages could better test the generalizability of an MT strategy.

Simultaneously, we recruited eight bilingual native Spanish speakers and four native Chinese speakers to manually translate two reference translations for each of the 2,010 records. A database-driven web application was constructed to facilitate the creation of these reference translations (Chen, 2016, p. 93). The reference translations were used in multiple processes including MEMT system development and human evaluation. For example, half of the reference translations were used to develop the language and translation models in our MEMT approaches.

Moses for MEMT

This study utilized Moses, an open-source platform for statistical MT (Koehn et al., 2007), to implement three MT approaches. We called them MEMT approaches because all of them combined the MT results from the three online MT services to achieve better performance. Statistical machine translation (SMT) has been the predominant MT approach since the 1990s after the seminal work by IBM researchers Brown, Pietra, Pietra, and Mercer (1993). Their work explained the translation of the source language s to a target language t by finding t', the maximum of the products of the

Language Model (LM) *p(t)* and a separate Translation Model (TM) *p(s|t)*, as illustrated in the formula below:

$$t' = \text{argmax} \ \ \text{p(t)} * \text{p(s|t)}$$

Since its launch, the statistical MT platform Moses has been constantly under development and it now offers two types of decoders: phrase-based and tree-based. For our study, we used the phrase-based Beam Search decoder which takes inputs from an LM module and a TM module in order to calculate the most probable target language translation for a source language input string or file (Koehn et al., 2007).

MEMT Approaches

This study tested three different MEMT approaches. The differences between these approaches lay in the training data employed by Moses's LM and TM modules. However, the decoding method for them was the same: All approaches applied Moses's phrase-based decoding method.

We used the following data for our experiments: The original 2,010 English metadata records, three sets of MTs from three different MT systems (6,030 records in total for each target language), and two sets of reference translations in Simplified Chinese and Spanish by native speakers of the target languages (4,020 records total for each language), as described earlier. We divided the English metadata records into two equal parts of 1,005 records: The first part and its 4,020 reference translations in the two target languages (2,010 for each language) are called *Set A*, and the second part and its reference translations are *Set B*. Set A was used in training for two of the MEMT approaches below, and Set B was used for testing and evaluation. As mentioned earlier, each metadata record consisted of multiple elements, such as title, creator, subject, and description. Each of these elements was sent to Moses as a separate sentence to train its translation and language models.

We decided to conduct experiments using three adapted MEMT approaches, which were called MEMT1, MEMT2, and MEMT3, respectively. These three approaches used different resources, and each was built on the other, as described below.

MEMT1

For the first approach, the TM was trained using the original English records and their MTs. We doubled the Microsoft Bing translation results in the TM, which was justified by a previous study finding that Microsoft Bing slightly outperformed Google Translate and performed much better than Yahoo! in adequacy and fluency (Chen et al., 2012), as well as the fact that Microsoft Bing's translations ranked highest of the three in terms of BLEU scores for Chinese (Papineni, Roukos, Ward, & Zhu, 2002; Chen et al., 2012). The LM consisted of only the MT systems' output in Chinese or Spanish. The output of Microsoft Bing Translator's was also doubled in the LM.

MEMT2

For the second approach, we used all the training data as was used in MEMT1, as well as Set A (composed of 1,005 English metadata records and their 2,010 reference Chinese or Spanish translations). In other words, Set A was added to the training data for generating both the TM and the LM for each language. The TMs were created by feeding parallel corpora of metadata records in both the target and source languages—Spanish/English or Chinese/English—to the TM module in Moses, which then generated phrase tables and reordering tables. The LM module in Moses trained Chinese and Spanish language models using manual translations of Set A and MT translations from Google, Bing, and Yahoo! for these two languages. The test data were the English metadata records of Set B, which was accepted by the Moses decoder as input. Finally, the decoder generated its Chinese and Spanish translation of Set B for evaluation. Figure 1 illustrates this approach within the structure of Moses.

MEMT3

MEMT3 was identical to MEMT2 except that in-domain data were added to the LMs. While many parallel corpora were available for MT research and training, such as the Holy Bible and the Europarl corpus (Koehn, 2005), we hoped that an LM consisting of additional data of the same type, that is, metadata records, would lead to better translation results.

For the Spanish LM, we acquired metadata records from the Redalyc database of UAEM (Universidad Autónoma del Estado de México, 2015) in Mexico. The data consisted of approximately 228,240 metadata records, most of which reflected scholarly papers. Some of these records contained non-Spanish string sequences. We therefore developed an in-house language identifier to exclude the non-Spanish records. This identifier used the Europarl corpus (Koehn, 2005) to build a character-level bigram language model, applied it to determine whether the language of each record was Spanish, and removed all non-Spanish records. Roughly 25% of the records were thus excluded by the identifier, leaving around 171,000 Spanish records. Figure 2 presents one of the Spanish records in its original Dublin Core format.

For the Chinese LM, we used 13,088 metadata records acquired from Shenzhen Library located in Guangdong Province in China. These were part of the bibliographic records for Chinese books purchased by the library in 2012. The original metadata records were in MARC Format (MARC Formats, n.d.). We processed them and used their most informative elements, such as title, creator, publisher, description, and subject in MEMT3. Table 2 presents one sample record in its original MARC format and the elements we extracted for use for MEMT3. We removed the mark-up tags from both Spanish and Chinese metadata records before using them for training the language models in Moses.

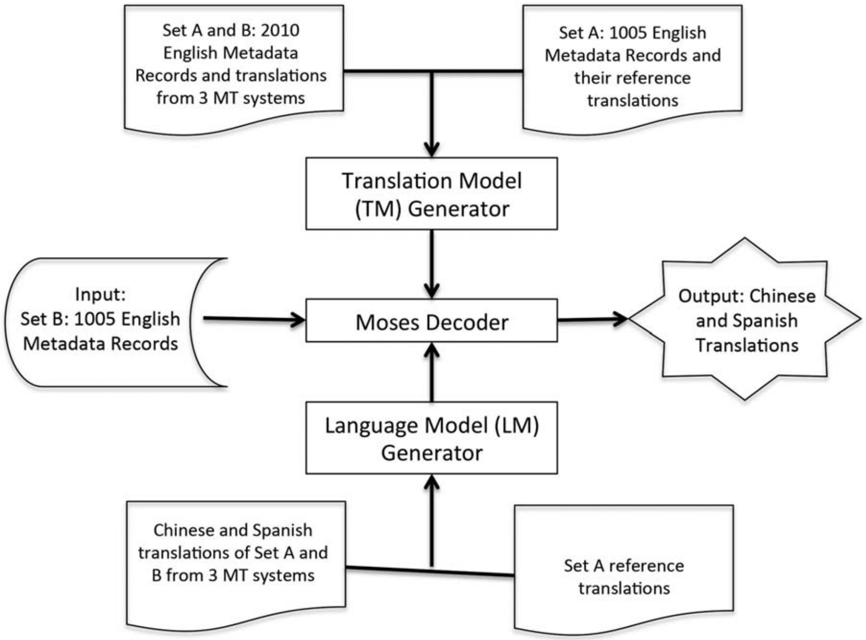
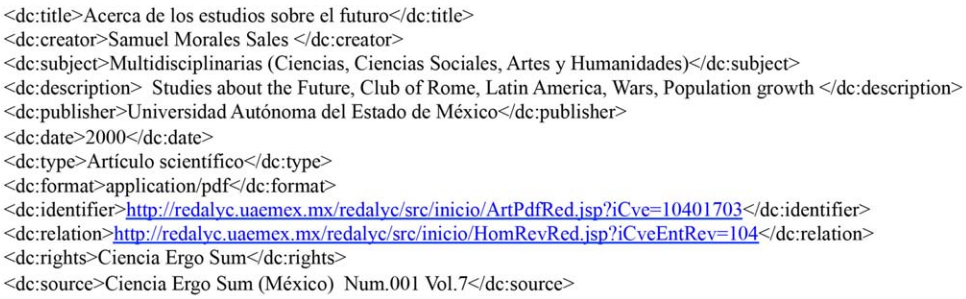

Set A and B: 2010 English Metadata Records and translations from 3 MT systems
Set A: 1005 English Metadata Records and their reference translations
Translation Model (TM) Generator
Input: Set B: 1005 English Metadata Records
Moses Decoder
Output: Chinese and Spanish Translations
Language Model (LM) Generator
Chinese and Spanish translations of Set A and B from 3 MT systems
Set A reference translations

FIG. 1. The structure and data of MEMT2.



```
<dc:title>Acerca de los estudios sobre el futuro</dc:title>
<dc:creator>Samuel Morales Sales </dc:creator>
<dc:subject>Multidisciplinarias (Ciencias, Ciencias Sociales, Artes y Humanidades)</dc:subject>
<dc:description>  Studies about the Future, Club of Rome, Latin America, Wars, Population growth </dc:description>
<dc:publisher>Universidad Autónoma del Estado de México</dc:publisher>
<dc:date>2000</dc:date>
<dc:type>Artículo scientífico</dc:type>
<dc:format>application/pdf</dc:format>
<dc:identifier>http://redalyc.uaemex.mx/redalyc/src/inicio/ArtPdfRed.jsp?iCve=10401703</dc:identifier>
<dc:relation>http://redalyc.uaemex.mx/redalyc/src/inicio/HomRevRed.jsp?iCveEntRev=104</dc:relation>
<dc:rights>Ciencia Ergo Sum</dc:rights>
<dc:source>Ciencia Ergo Sum (México)  Num.001 Vol.7</dc:source>
```

FIG. 2. A sample Spanish metadata record. [Color figure can be viewed at wileyonlinelibrary.com]

The corpora used to generate the translation models for MEMT3 were identical to those for MEMT2 for the two languages, while the Chinese and Spanish LMs for this approach were appended by 13,088 and 171,000 records, respectively.

As mentioned earlier, we doubled the MT results from Microsoft Bing Translator for all MEMT experiments. This was also considered necessary due to the limited training materials we had, and the different performance of the three online MT systems. Bing translations received higher BLEU scores during our experiments. Table 3 summarizes the data that were fed into LM and TM for each MEMT approach and for each language.

We experimented with the aforementioned three MEMT approaches and tested the 1,005 records in Set B. Our experiments produced three sets of translation results for each language. These results were manually evaluated, as described below.

Human Evaluation of MT Results

We used fluency and adequacy as evaluation criteria of MT performance. Fluency referred to the degree to which the target language output was well-formed according to the rules of a particular language, in this case Spanish and Chinese. Adequacy referred to the degree to which information present in the source language text was represented in the target language output (Linguistic Data Consortium, 2005; Chen, 2016, p. 105). To facilitate the measure of these criteria, we provided participants with definitions of five-point scales for each of them. Table 4 presents the definitions of the fluency and adequacy scales.

The evaluation method and processes have been described in Chen (2016). In general, we recruited graduate students and professionals who were fluent in Chinese or Spanish to perform two tasks using a web evaluation system we built specifically for this project. The first task was to assign adequacy and fluency scores to each element of a record, and then to each record as a whole. Two reference translations were presented alongside the MT results. The

DOI: 10.1002/asi

TABLE 2. A sample Chinese metadata record.

Format	The Sample Record
The original record in MARC format	00100130000000500170001300900150003001000280004503500250007310000410009810100080013910200150014710 50018001621060006001802000029001862100031002152150016002463300181002626060029004436860011 004726900014004837010034004978010025005319050046005556–012003060542-20051214170349.0-CRLN 2004013961-a7–218-04182–5dCNY30.00-a(012001)012004005117-a20031231d2003 m y0chiy0110 ea-0 achi-aCNb44 0000-ay z 000yy-ar-1 a中国近代思潮论f丁和著-a广州c广东人民出版社d2003-a647页d21cm-a本书收录了20 世纪 80 和 90 年代的有关近代思想文化的文章22篇。即《近代中国探索发展道路的历史考察》、《中国近代思潮的思考》、《民主科学在中国的命运》、《中国传统文化与现代化问题》等。-0 a思想史x研究y中国z近代-aB25v2-aB250.5v4-0a丁守和f(1927)4著3A9512864- 0aCNb012001c20040317-aSTfB25/146s3b1527279b1527280b1527281-
After processing, the extracted elements of the same record	Title: 中国近代思潮论 Creator: 丁守和 Publisher: 广州##广东人民出版社 Description: 本书收录了20世纪80和90年代的有关近代思想文化的文章22篇。 即《近代中国探索发展道路的历史考察》、《中国近代思潮的思考》、《民主科学在中国的命运》、 《中国传统文化与现代化问题》等。 Subject: 思想史##研究##中国##近代

TABLE 3. MEMT approaches and data for training LM and TM.

Approach	Language	LM	TM
MEMT1	Chinese	**8,040 Chinese Records**: Chinese translation from Google, Bing (Doubled), & Yahoo! for the 2010 English records	**8,040 Parallel Records**: The 2010 English records and their Chinese translations from 3 Systems
	Spanish	**8,040 Spanish Records**: Spanish translation from Google, Bing (Doubled), & Yahoo! for the 2010 English records	**8,040 Parallel Records**: The 2010 English Records and their Spanish translations from 3 Systems
MEMT2	Chinese	**10,050 Chinese Records**: All records from MEMT1 Chinese LM plus Set A's Reference Chinese Translations *(*2,010 total records*)*	**10,050 Parallel Records**: Same as MEMT1 Chinese TM plus Set A (1,005 English metadata records and their Reference Chinese Translations)
	Spanish	**10,050 Spanish Records**: All records from MEMT1 Spanish LM plus Set A's Reference Spanish Translations (2,010 records)	**10,050 Parallel Records**: Same as MEMT1 Spanish TM plus Set A (1,005 English metadata records and their Reference Spanish Translations)
MEMT3	Chinese	**10,050+13K Chinese Records**: All records from MEMT2 Chinese LM plus 13K Chinese Metadata Records	Same as MEMT2 Chinese TM
	Spanish	**10,050+177K Spanish Records**: All records from MEMT2 Spanish LM plus 177K Spanish Metadata Records	Same as MEMT2 Spanish TM

Note. Set A served for training; Set B served for testing and evaluation.

evaluators consulted the reference translations when judging adequacy and fluency. They were not given the original English records. The second task was to identify the best and worst translations for each element and for the record as a whole. For this task, the web evaluation system displayed the MT results of all three systems. The evaluators selected "Other" if two or more systems provided identical translations.

Results

We enlisted the help of humans to evaluate the MT output of six systems: Google (Translate), Bing (Translation), Yahoo! (Translation), MEMT1, MEMT2, and MEMT3. Two rounds of evaluation were conducted with the sample evaluation approach: In the first round we had the evaluators assess the three online systems, and in the second round they judged the three MEMT results. A crowdsourcing type of approach was applied to recruit evaluators. We advertised the evaluation tasks at the University of North Texas, and through our partners in China and Mexico. In total, we recruited 16 Spanish evaluators and nine Chinese evaluators from Mexico, China, and the United States All evaluators went through an online training lesson prior to conducting the evaluation. The training lesson explained the evaluation tasks and measures, and provided tips on making appropriate judgments based on the five-point scales for adequacy and fluency as presented in Table 4. In order to improve reliability, each record was independently evaluated by three different evaluators. The translation results were presented to the

DOI: 10.1002/asi

TABLE 4. Fluency and adequacy scales (Chen, 2016, p. 105).

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

TABLE 5. Average adequacy and fluency scores for Chinese translations.

Adequacy	System	Whole	Title	Subject	Creator	Description	Publisher	Coverage
	Bing	3.16	3.24	3.64	3.46	3.08	3.47	3.31
	Google	**3.30**	**3.26**	**3.77**	**3.66**	**3.11**	**3.59**	**3.42**
	Yahoo!	2.94	2.89	3.62	2.81	2.85	3.22	3.06
	MEMT1	3.54	3.62	4.29	3.48	3.55	4.07	3.81
	MEMT2	3.62	**3.64**	4.36	3.73	**3.58**	**4.09**	3.93
	MEMT3	**3.68**	**3.64**	**4.38**	**3.85**	**3.58**	4.05	**4.10**
Fluency	System	Whole	Title	Subject	Creator	Description	Publisher	Coverage
	Bing	3.13	**3.11**	3.67	3.51	**2.73**	3.50	3.31
	Google	**3.19**	3.08	**3.77**	**3.74**	2.72	**3.64**	**3.44**
	Yahoo!	2.93	2.81	3.66	2.97	2.54	3.25	3.04
	MEMT1	3.62	3.61	4.40	3.69	3.48	4.16	3.87
	MEMT2	3.68	**3.64**	4.47	3.89	**3.49**	4.16	3.98
	MEMT3	**3.72**	3.63	**4.48**	**4.00**	**3.49**	**4.17**	**4.18**

evaluators in random order to prevent the order effects, and no evaluator knew which system had produced the output being evaluated.

Manual Evaluation Results

Once the results were collected, they were processed using the following method. Each item (an element or the metadata record as a whole) to be evaluated had three evaluation scores for adequacy and three scores for fluency. If each of the three scores was different, the median was chosen as the score for that item. For example, if the scores were 3, 4, and 5, the median, that is, 4, would be chosen. We used median because the distribution of our data was very skewed. Median was considered more appropriate than mean/average in that situation (Vaughan, 2001, p. 31). If two evaluators gave the same score, the mode was chosen as the score for that record. For example, if the scores were 4, 4, and 5, the mode, that is, 4, would be chosen. The mode represented the most popular choice. Once the final scores for each item were determined, they were averaged. Tables 5 and 6 summarize the evaluation results for Chinese and Spanish, respectively. The column "Whole" refers to the evaluation score for the metadata record as a whole.

Table 5 showed that for all elements and the record as a whole, MEMT approaches were consistently better than any single MT service in both measures for Chinese translation, with two exceptions. Especially MEMT2 and MEMT3 achieved higher average scores than even the best score of the three online MT systems on both adequacy and fluency. All MEMT approaches achieved a score above 3.5 on average adequacy. Also, MEMT2 and MEMT3 obtained similar average scores for almost all elements and the record as a whole. Google, and occasionally Bing, performed the best among the online MT systems. To compare these scores in a more straightforward way, we present Chinese average adequacy scores in Figure 3.

Figure 3 shows two numbers for each group, the numerical value on the left is the average adequacy score of the best-performing online MT system, while the value on the right is the average adequacy score for the best-performing MEMT system. For example, in the first group for "Whole," the value 3.30 is the average adequacy score for Google, and 3.68 is that for MEMT3. Comparing these two figures on fluency scores, we found that the MEMT approaches brought slightly larger improvements to fluency scores than the adequacy scores depicted in Figure 3. This might indicate that the MEMT strategy could identify translations closer to what a native Chinese speaker would prefer. The evaluation results for Spanish translations are presented in Table 6, and visually depicted in Figure 4 (for adequacy) using the same representation as in Figure 3.

Spanish translation results, as presented in Table 6 and depicted in Figure 4, indicate that MEMT1 did not perform

TABLE 6. Average adequacy and fluency scores for Spanish translations.

Adequacy	System	Whole	Title	Subject	Creator	Description	Publisher	Coverage
	Bing	3.95	**4.10**	**4.41**	**4.50**	4.06	4.15	4.29
	Google	**3.96**	4.06	4.34	**4.50**	**4.07**	**4.16**	**4.39**
	Yahoo!	3.75	3.77	4.06	4.12	3.76	4.02	3.95
	MEMT1	3.68	3.92	3.88	4.77	3.95	4.07	4.04
	MEMT2	**4.39**	4.56	**4.68**	**4.88**	**4.53**	**4.49**	**4.47**
	MEMT3	4.38	**4.57**	**4.68**	**4.88**	4.51	**4.49**	**4.47**
Fluency	**System**	**Whole**	**Title**	**Subject**	**Creator**	**Description**	**Publisher**	**Coverage**
	Bing	3.95	**4.05**	**4.37**	**4.51**	3.98	**4.15**	4.22
	Google	**3.96**	4.01	4.31	4.49	**3.99**	4.13	**4.33**
	Yahoo!	3.76	3.75	4.07	4.16	3.70	3.99	3.97
	MEMT1	3.59	3.86	3.81	4.76	3.90	4.04	4.00
	MEMT2	**4.47**	**4.57**	**4.71**	**4.87**	**4.53**	4.51	4.65
	MEMT3	4.46	4.56	4.69	**4.87**	4.51	**4.52**	**4.66**

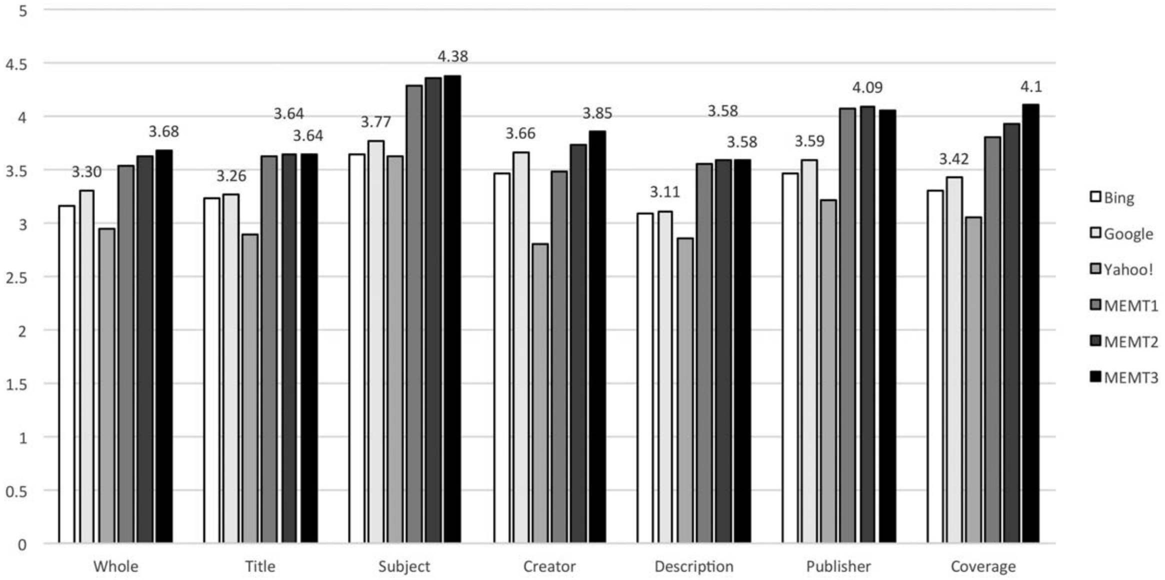


FIG. 3. Average adequacy scores for Chinese translations.

as well as Bing and Google, two of the online MT systems. The three online MT systems, especially Google and Bing, did quite well (above 3.95 on adequacy and fluency for the records as a whole) on Spanish translation. However, MEMT2 and MEMT3 achieved more than 10% higher average scores for both measures. This shows that the addition of Set A to the training data made a difference in performance. Comparing Tables 5 and 6, we found that online MT services performed much better (more than 20%) when translating English metadata records into Spanish than into Chinese.

As for comparative evaluation results, Google was considered the best among the three MT systems, while MEMT3 was the best among the three MEMT systems for Chinese. The results were consistent with the individual evaluation reported in Table 5. For Spanish, however, Google and Bing Translator were considered to have provided the same quality of translation, as they received very close scores on both adequacy and fluency (3.96 for Google and 3.95 for Bing). MEMT2 and MEMT3 were judged to have had the same performance, which was consistent with the evaluation results reported in Table 6—MEMT2 and MEMT3 received nearly identical scores on adequacy (4.39 for MEMT2 and 4.38 for MEMT 3) and adequacy (4.47 for MEMT2 and 4.46 for MEMT3).

Frequency Distribution and Significance Testing

Frequency distribution enables us to understand how the scores are distributed along the scales. Table 7 shows the frequency distribution of adequacy and fluency scores for Chinese translation. None of the translations was assigned a 1 for either adequacy or fluency. The scores in Table 7 are presented as both counts and percentages. For example, the last adequacy row of Table 7 shows that a total of 12 records

DOI: 10.1002/asi

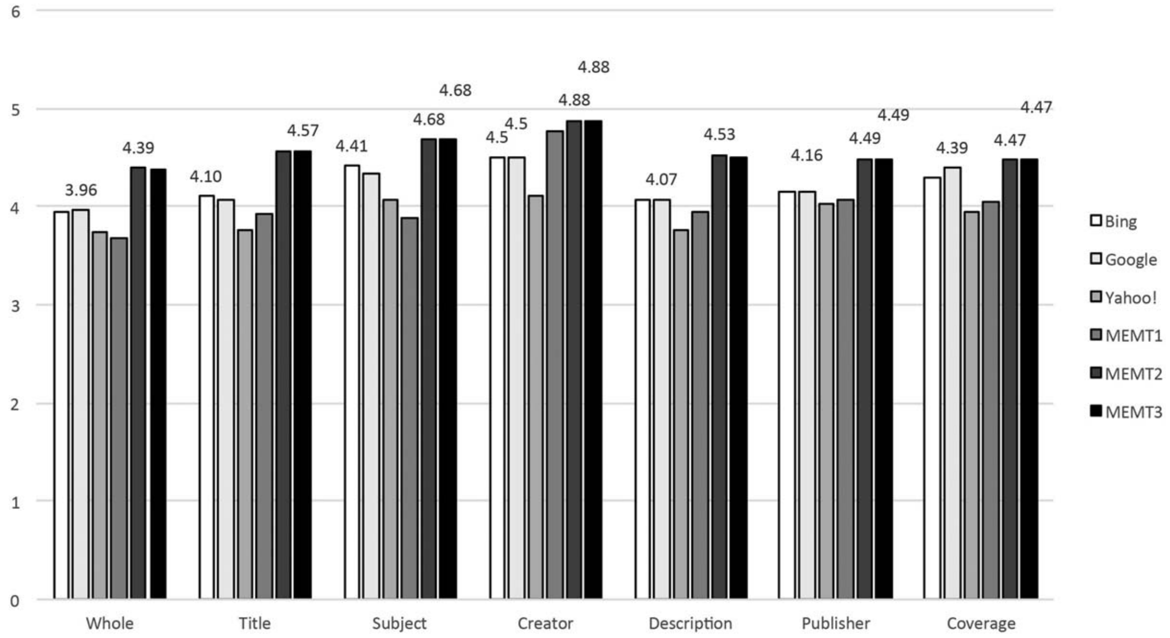


FIG. 4. Average adequacy scores for Spanish translations.

TABLE 7. Frequency distribution of adequacy and fluency scores for Chinese translations.

Adequacy	Scores	Bing	Google	Yahoo!	MEMT1	MEMT2	MEMT3	Total
	2	40 (20.73%)	22 (11.4%)	**123 (63.73%)**	3 (1.55%)	3 (1.55%)	2 (1.04%)	193
	3	761 (22.42%)	665 (19.59%)	**815 (24.01%)**	457 (13.46%)	373 (10.99%)	323 (9.52%)	3,394
	4	203 (8.35%)	317 (13.04%)	67 (2.76%)	543 (22.34%)	628 (25.83%)	**673 (27.68%)**	2,431
	5	1 (8.33%)	1 (8.33%)	0 (0%)	2 (16.67%)	1 (8.33%)	**7 (58.33%)**	12
								6,030
Fluency	Scores	Bing	Google	Yahoo!	MEMT1	MEMT2	MEMT3	Total
	2	53 (21.99%)	43 (17.84%)	**141 (58.51%)**	2 (0.83%)	1 (0.41%)	1 (0.41%)	241
	3	773 (23.33%)	731 (22.06%)	**796 (24.02%)**	380 (11.47%)	331 (9.99%)	303 (9.14%)	3,314
	4	178 (7.33%)	229 (9.44%)	68 (2.8%)	617 (25.42%)	660 (27.19%)	**675 (27.81%)**	2,427
	5	1 (2.08%)	2 (4.17%)	0 (0%)	6 (12.5%)	13 (27.08%)	**26 (54.17%)**	48
								6,030

were assigned a score of 5. Of these 12 scores, MEMT3 was responsible for most of them, seven or 58.33%, respectively. Table 7 shows that MEMT approaches had a much smaller number of records that received a score of 2 compared to the three online MT systems. Also, most of the high fluency and adequacy scores (4 and 5) were produced by the three MEMT systems, specifically MEMT3. MEMT3 alone produced 27.68% and 58.33% of 4 and 5 adequacy scores, respectively. It also produced 27.81% and 54.17% of 4 and 5 fluency scores, respectively.

Table 8 presents the frequency distribution for adequacy and fluency scores for Spanish translations in a similar way as in Table 7. It demonstrates that MEMT1 was responsible for most of the low adequacy scores, with 68.75% of the 2 scores and 37.12% of the 3 scores. This might indicate that simply combining the MT results of multiple systems might not work well for language pairs that could be well translated by the individual MT systems.

In order to confirm that the results produced by online MT systems and those produced by the MEMT approaches were significantly different, we conducted nonparametric tests of significance. Since the distribution of scores for both Spanish and Chinese translations was not a normal distribution, we performed significance testing using the Mann–Whitney *U*-test, instead of the traditional *t*-test (Hinkle, Wiersma, & Jurs, 2003a; Hinkle et al., 2003a). Table 9 shows the results of the Mann–Whitney *U*-test for Chinese adequacy and fluency scores on the whole record. In each applicable cell, the first score is the Mann–Whitney *U statistic*, while the second score is the value of *p*. Table 9 shows that the three MEMT approaches produced results that were significantly different both from each other and from the three online MT systems. In almost all cases, $p < .00$. Similar

DOI: 10.1002/asi

TABLE 8. Frequency distribution for adequacy and fluency scores for Spanish translations.

Adequacy	Scores	Bing	Google	Yahoo!	MEMT1	MEMT2	MEMT3	Total
	2	2 (6.25%)	0 (0%)	6 (18.75%)	**22 (68.75%)**	1 (3.13%)	1 (3.13%)	32
	3	130 (15.22%)	128 (14.99%)	265 (31.03%)	**317 (37.12%)**	8 (0.94%)	6 (0.7%)	854
	4	793 (19.16%)	**794 (19.19%)**	717 (17.33%)	630 (15.22%)	594 (14.35%)	610 (14.74%)	4,138
	5	80 (7.95%)	83 (8.25%)	17 (1.69%)	36 (3.58%)	**402 (39.96%)**	388 (38.57%)	1,006
								6,030
Fluency	Scores	Bing	Google	Yahoo!	MEMT1	MEMT2	MEMT3	Total
	2	1 (1.67%)	0 (0%)	5 (8.33%)	**52 (86.67%)**	1 (1.67%)	1 (1.67%)	60
	3	122 (14.68%)	110 (13.24%)	255 (30.69%)	**328 (39.47%)**	6 (0.72%)	10 (1.2%)	831
	4	807 (20.23%)	**823 (20.63%)**	723 (18.12%)	601 (15.07%)	515 (12.91%)	520 (13.04%)	3,989
	5	75 (6.52%)	72 (6.26%)	22 (1.91%)	24 (2.09%)	**483 (42%)**	474 (41.22%)	1,150
								6,030

TABLE 9. Mann–Whitney *U*-test results for Chinese for the whole record.

Adequacy		Google	Bing	Yahoo!	MEMT1	MEMT2	MEMT3
	Google	—	442855.5, $p < .00$	663106.0, $p < .00$	—	—	—
	Bing	—	—	604390.0, $p < .00$	—	—	—
	Yahoo!	—	—	—	—	—	—
	MEMT1	386870.0, $p < .00$	325730.0, $p < .00$	237867.5, $p < .00$	—	463285.0, $p < .00$	435920.0, $p < .00$
	MEMT2	345785.0, $p < .00$	285344.0, $p < .00$	200857.0, $p < .00$	—	—	477412.0, $p = .01$
	MEMT3	319446.5, $p < .00$	259749.5, $p < .00$	177696.0, $p < .00$	—	—	—
Fluency		Google	Bing	Yahoo!	MEMT1	MEMT2	MEMT3
	Google	—	476067.0, $p = .00$	621409.5, $p < .00$	—	—	—
	Bing	—	—	594226.5, $p < .00$	—	—	—
	Yahoo!	—	—	—	—	—	—
	MEMT1	300523.5, $p < .00$	272380.0, $p < .00$	199927.0, $p < .00$	—	477716.0, $p = .01$	459708.5, $p < .00$
	MEMT2	275328.0, $p < .00$	247565.5, $p < .00$	177620.5, $p < .00$	—	—	486764.0, $p = .09$
	MEMT3	260386.5, $p < .00$	233088.0, $p < .00$	165082.5, $p < .00$	—	—	—

TABLE 10. Mann–Whitney *U*-test results for Spanish for the whole record.

Adequacy		Google	Bing	Yahoo!	MEMT1	MEMT2	MEMT3
	Google	—	501725.0, $p = .72$	600260.5, $p < .00$	—	—	—
	Bing	—	—	596924.5, $p < .00$	—	—	—
	Yahoo!	—	—	—	—	—	—
	MEMT1	624301.0, $p < .00$	621069.0, $p < .00$	533595.5, $p = .01$	—	223320.0, $p < .00$	227035.0, $p < .00$
	MEMT2	310336.0, $p < .00$	307629.0, $p < .00$	234398.0, $p < .00$	—	—	511382.5, $p = .56$
	MEMT3	315553.0, $p < .00$	312817.0, $p < .00$	238554.0, $p < .00$	—	—	—
Fluency		Google	Bing	Yahoo!	MEMT1	MEMT2	MEMT3
	Google	—	500235.5, $p = .71$	597637.5, $p < .00$	—	—	—
	Bing	—	—	592268.0, $p < .00$	—	—	—
	Yahoo!	—	—	—	—	—	—
	MEMT1	655307.5, $p < .00$	650016.5, $p < .00$	569057.5, $p < .00$	—	178626.5, $p < .00$	183297.0, $p < .00$
	MEMT2	273095.5, $p < .00$	271202.5, $p < .00$	209053.0, $p < .00$	—	—	510545.5, $p = .62$
	MEMT3	278989.0, $p < .00$	277029.5, $p < .00$	214361.5, $p < .00$	—	—	—

results were observed for fluency, with the exception that the difference between MEMT2 and MEMT3 did not prove to be statistically significant. This is consistent with our earlier findings that MEMT2 and MEMT3 achieved very close fluency scores on Chinese translation.

Table 10 presents the results of the Mann–Whitney *U*-test for Spanish adequacy and fluency scores. It shows that the three systems produced results that were significantly different. The only exceptions were the adequacy and fluency scores for Bing and Google, and MEMT2 and MEMT3, which were not found to have statistically significant differences. This is consistent with our earlier finding that, for Spanish translations, Google and Bing, as well as MEMT2 and MEMT3, produced similarly good results.

Correlation Between Adequacy and Fluency

Previous research has reported that the two measures of adequacy and fluency were highly associated (Callison-Burch et al., 2007). Our calculation using Pearson's *r* confirmed this conclusion for both languages. There was a very strong positive correlation between adequacy and fluency for all experiments ($r >= 0.8$). It indicates that if time and money are limited for human evaluation of MT, one can choose to use just one measure in evaluation instead of using both adequacy and fluency.

We also evaluated the results using automatic measures including BLEU and METEOR, which measure the similarity between reference translations and the MT output. They have been widely used to compare MT performance of different systems. Their limitations, however, include assigning lower scores to MT results that are very different from the reference translations, and providing little insight into the translation problems (Lommel et al., 2014). We found that both the BLEU and METEOR scores of the MEMT approaches were mostly lower than those of the best online MT systems, which were opposite to human-judged results. This might indicate that MEMT results differed more from reference translations than those of the online MT systems.

Discussion

Unique Findings and Significance

This study tried to answer the question: Which MEMT strategy can achieve better performance on metadata records? We designed and implemented three MEMT strategies using an open-source statistical MT platform, and conducted human evaluation of the translation results. Our evaluation found that combining results from multiple low-cost online MT services with a small sample of parallel data could significantly improve translation performance in terms of adequacy and fluency for both Chinese and Spanish.

MEMT has been explored by MT researchers with different strategies (Nirenburg & Frederlcing, 1994; Rosti et al., 2007); however, our study was the first to investigate this approach using library metadata records. Furthermore, our approach was different from most MEMT approaches in the literature, which treated one of the outputs as a base and aligned other outputs to the base with editing as necessary based on measures such as translation error rate (TER). In contrast, our approach fed all online MT outputs into Moses and let Moses take care of the rest, which is simpler and easier to implement for digital libraries. Libraries usually face challenges such as a lack of skilled technical staff and appropriate budgets for purchasing MT systems. Our study provided an alternative approach that libraries could utilize to have their metadata records translated into other languages effectively and efficiently.

This study was one of the very few studies that conducted human evaluation of MT on metadata records. It confirmed the conclusion from the previous study (Chen et al., 2012) that online MT systems could produce non-native yet sufficiently good translations that might help information users in many ways overcome language barriers. However, it significantly extended the MT evaluation in our previous study (Chen et al., 2012) to more languages and a larger sample size. Especially, we conducted evaluation on different elements and found that some online MT systems, such as Google Translate and Bing Translator, produced translations with high levels of fluency and adequacy for certain metadata elements such as subject and creator. For example, most subject terms could be correctly translated by Google Translate (adequacy score 3.77 on Chinese, and 4.31 on Spanish). Developers of digital libraries might consider providing MLIA for their collections by integrating online MT services to translate subject terms.

Our results indicated that Spanish translations scored higher than Chinese translations. Two of the online MT systems produced quite good Spanish translations. This may due to the fact that English and Spanish are more similar to each other than Chinese is to English, as described earlier. Specifically, named entities kept their original forms and did not need to be translated from Spanish to English. Providing Spanish information access for English digital collections is likely to be achieved using current online MT services.

The Three MEMT Approaches

Among the three MEMT approaches, MEMT1 did not prove as effective as MEMT2 and MEMT3. Adding manually generated parallel data in MEMT2 proved effective, which is consistent with the MT literature that in-domain bilingual corpora are often considered requisite for effective MT (Ananthakrishnan, Prasad, Stallard, & Natarajan, 2013; Chelba & Acero, 2006). Also, adding a monolingual corpus in the target language had a significant, positive effect on the quality of Chinese translations, but not for Spanish. This indicates that MT effectiveness needs to be tested on an individual basis for language pairs.

It is expected that the MEMT system could achieve even better performance with a larger and more diverse in-domain bilingual data set. The data set could be created by manual human translation, which would be time-consuming and costly, or a postediting solution. The postediting translation method, which generates human translation by allowing translators to edit MT results (Allen, 2003), has become popular with the advancement of MT technologies. Our future work will include the creation of a larger multilingual parallel data set using a postediting approach.

Limitations

This study had several limitations. We applied the measures of adequacy and fluency to assess the MT quality of metadata records as a whole, as well as each of the six elements. However, for some elements, such as creator and subject, it would be more appropriate to use adequacy only, as these elements are mostly short text segments or phrases. Another limitation was that we found MT systems returned the same results for many of the elements of some metadata records. We should perform comparative evaluation only on whole records and compare two systems at a time. Lastly, we found that some quality issues in reference translations were due to the fact that we did not establish a clear set of translation rules for translators beforehand. For example, some acronyms specific to metadata records were not consistently translated.

Conclusion and Future Research

This study experimented with three MEMT strategies and evaluated Chinese and Spanish translations generated by six MT systems for 1,005 sample English metadata records extracted from two digital collections. It provided evidence and useful information about the performance of current MT technologies on metadata records, which is much needed by the digital library community in order to design and implement value-added services, such as MLIA, for their digital collections. Especially, we found that MT strategies combining translation outputs from multiple low-cost MT systems with a small, linguistically-appropriate corpus could significantly improve translation performance.

The ultimate goal of our research will be to investigate effective and efficient MLIA for digital collections. The study reported in this paper served as the first step toward that goal. Future studies will include integrating additional language resources to improve named entity translation, and investigating cross-language information retrieval effectiveness based on metadata records translation using different approaches, including online MT services and two of the above effective MEMT strategies. We will also develop more parallel metadata records in English, Chinese, and Spanish and use them for improving MT performance.

Acknowledgments

This study was supported by the Institute of Museum and Library Services (IMLS) under the National Leadership Grants LG-06-10-0162-10 and LG-06-13-0187-13. We thank Dr. Ying Zhang for his advice and assistance on MT and Moses. We are grateful for other research assistants Cheng Chieh Lien, Olajumoke Agozu, and Wenqian Zhao on assisting with system implementation, metadata records extraction, and data analysis. We also thank our translators, evaluators, and other researchers who participated in the study, and the anonymous reviewers for their valuable edits, comments, and suggestions.

References

Allen, J. (2003). Chapter 16: Post-editing. In H. Somers (Ed.), Computers and translation: A translator's guide (pp. 297–316). Amsterdam: John Benjamins Publishing Company.

Ananthakrishnan, S., Prasad, R., Stallard, D., & Natarajan, P. (2013). Batch-mode semi-supervised active learning for statistical machine translation. Computer Speech and Language, 27, 397–406.

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., & Mercer, R.L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19, 263–313.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136–158). Prague, Czechoslovakia: Association for Computational Linguistics.

Chelba, C., & Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. Computer Speech and Language, 20, 382–399.

Chen, J. (2016). Multilingual access and services for digital collections. Santa Barbara, CA: Libraries Unlimited.

Chen, J., & Bao, Y. (2009). Information access across languages on the web: From search engines to digital libraries. Proceedings of the American Society for Information Science and Technology, 46, 1–14. doi: 10.1002/meet.2009.1450460278

Chen, J., Ding, R., Jiang, S., & Knudson, R. (2012). A preliminary evaluation of metadata records machine translation. The Electronic Library, 30, 264–277.

Chowdhury, S., Landoni, M., & Gibb, F. (2006). Usability and impact of digital libraries: A review. Online Information Review, 30, 656–680. doi: 10.1108/14684520610716153.

Diekema, A.R. (2012). Multilinguality in the digital library: A review. The Electronic Library, 30, 165–181.

Gaspari, F. (2004). Online MT services and real users' needs: An empirical usability evaluation. In R.E. Frederking & K.B. Taylor (Eds.), Machine translation: From real users to research lecture notes in computer science (vol. 3265, pp. 74–85). Berlin: Springer.

Google. (2016). Google Translate [Computer software]. Retrieved from https://translate.google.com/.

Guzmán, G., Joty, S., Márquez, L., & Nakov, P. (2015). Pairwise neural machine translation evaluation. In *Proceedings of the 53rd ACL Annual Conference* (pp. 805–814). Beijing: Association for Computational Linguistics.

Heafield, K., & Lavie, A. (2010). Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. The Prague Bulletin of Mathematical Linguistics, 93, 27–36.

Hillman, D. (2008). Metadata quality: From evaluation to augmentation. Cataloging & Classification Quarterly, 46, 65–80.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (2003a). Hypothesis testing: One-sample case for the mean. In Applied statistics for the behavioral sciences (5th ed., pp. 174–200). Boston: Houghton Mifflin.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (2003b). Other nonparametric tests. In Applied statistics for the behavioral sciences (5th ed., pp. 572–586). Boston: Houghton Mifflin.

Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of context-based machine translation evaluation. Machine Translation, 17, 43–75.

International Children's Digital Library Foundation (n.d.). International Children's Library. Retrieved from http://en.childrenslibrary.org/.

Internet World Stats. (2016). Internet world users by language. Retrieved from http://www.internetworldstats.com/stats7.htm.

Jones, G.J.F., Fantino, F., Newman, E., & Zhang, Y. (2008, January). Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. *Second international workshop on cross-lingual information access: Addressing the information need of multilingual societies*. Workshop conducted at the meeting of International Joint Conference on Natural Language Processing, Hyderabad, India.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: The Tenth Machine Translation Summit* (pp. 79–86). Phuket, Thailand: Asia-Pacific Association for Machine Translation.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 177–180). Stroudsburg, PA: Association for Computational Linguistics.

Library of Congress (n.d.). World Digital Library. Retrieved from https://www.wdl.org/en/.

Linguistic Data Consortium. (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations revision 1.5. Retrieved from http://wayback.archive.org/web/20100622130328/http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf.

Lommel, A., Burchardt, A., Popovic, M., Harris, K., Avramidis, E., & Uszkoreit, H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)* (pp. 165–172). Dubrovnik, Croatia: European Association for Machine Translation.

MARC Formats (n.d.). MARC standards. Retrieved from https://www.loc.gov/marc/

Matusiak, K.K., Meng, L., Barczyk, E., & Shih, C. (2015). Multilingual metadata for cultural heritage materials. The case of the Tse-Tsung Chow collection of Chinese scrolls and fan paintings. The Electronic Library, 33, 136–151.

Ménard, E. (2012). TIIARA: The “making of” a bilingual taxonomy for retrieval of digital images. Library Hi Tech, 30, 643–654. https://doi.org/10.1108/07378831211285103

Microsoft. (2016). Bing Translate [Computer Software]. Retrieved from http://www.bing.com/translator.

Moses Statistical Machine Translation System: Version 3.0. [Computer software]. Retrieved from http://www.statmt.org/moses/

Nirenburg, S., & Frederlcing, R. (1994). Toward multi-engine machine translation. In Proceedings of the workshop on human language technology (pp. 147–151). Stroudsburg, PA: Association for Computational Linguistics.

Oard, D.W., He, D., & Wang, J. (2008). User-assisted query translation for interactive cross-language information retrieval. Information Processing and Management, 44, 181–211.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 311–318). Stroudsburg, PA: Association for Computational Linguistics.

Peters, C., Braschler, M., & Clough, P. (2012). Multilingual information retrieval: From research to practice. New York: Springer.

Purday, J. (2012). Europeana: Digital access to Europe’s cultural heritage. Alexandria, 23, 1–13.

Ren, F., & Shi, H. (2002). A multi-engine translation approach to machine translation. International Journal of Information Technology & Decision Making, 1, 349.

Robertson, R.J. (2005). Metadata quality: Implications for library and information science professionals. Library Review, 54, 295–300.

Rosti, A.-V.I., Ayan, N.F., Xiang, B., Matsoukas, S., Schwartz, R., & Dorr, B.J. (2007). Combining outputs from multiple machine translation systems. In Proceedings of NAACL HLT 2007 (pp. 228–235). Stroudsburg, PA: Association for Computational Linguistics.

Roxas, R.E., Oñate Borra, A., Ko Cheng, C., Lim, N.R., Ong, E.C., & Tan, M.W. (2008). Building language resources for a multi-engine English-Filipino machine translation system. Language Resources and Evaluation, 42, 183–195. doi: 10.1007/s10579-007-9037-5.

Sakai, T., Kando, N., Lin, C.J., Mitamura, T., Shima, H., Ji, D., … Nyberg, E. (2008). Overview of the NTCIR-7 ACLIA IR4QA task. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan. Retrieved from: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/01-NTCIR7-OV-IR4QA-SakaiT.pdf

Tripathi, S., & Sarkhel, J.K. (2010). Approaches to machine translation. Annals of Library and Information Studies, 57, 388–393.

Universidad Autónoma del Estado de México. (2015). Red de Revistas Científicas de América Latina y el Caribe, España y Portugal (Redalyc) [Database]. Retrieved from: http://www.redalyc.org/

University of North Texas Libraries. (2016). The portal to Texas history — About the portal. Retrieved from https://texashistory.unt.edu/about/portal/

University of North Texas Libraries. (n.d.). University of North Texas library catalog. Retrieved from http://library.unt.edu

Vaughan, L. (2001). Statistical methods for the information professionals: A practical, painless approach to understanding, using, and interpreting statistics. ASIS&T monograph series (Vol. 367). Medford, NJ: American Society for Information Science and Technology.

Zhang, J., & Zhong, C. (2016). Machine translation. In Chinese information processing trend report. Retrieved from Chinese Information Processing Society of China: http://cips-upload.bj.bcebos.com/cips2016.pdf