Building A Multilingual Test Collection for Metadata Records

Jiangping Chen[1], Min Namgoong[1], Brenda Reyes Ayala[1], Gaohui Cao[2], Xinyue Wang[3]
[1]Department of Information Science, University of North Texas
[2]School of Information Management, Central China Normal University
[3]Department of Computer Science and Engineering, University of North Texas

Abstract
This paper describes the principles and processes of building a test collection that enables multilingual information retrieval for digital metadata records. The collection includes a multilingual collection of 1,005,752 metadata records, their Chinese and Spanish machine translation results, 45 topics generated through crowd- sourcing, and their relevant judgments.
Keywords: test collection; multilingual information retrieval; metadata records; information retrieval evaluation
Citation:
Copyright: Copyright is held by the authors.
Acknowledgements: This study is funded by the National Leadership Grant for Libraries LG-06-13-0187-13 of the Institute of Museum and Library Services (IMLS)
Research Data: the final product of this study will be available by May 31, 2017
Contact: Jiangping.Chen@unt.edu

1 Introduction

Information Retrieval (IR) systems are often evaluated using test collections. A test collection usually contains a document collection, a set of test topics, and relevance judgments associated with those topics. Test collections have played a crucial role in advancing IR research and practice. The most influential test collections were those developed by the organizers of the three major IR forums: TREC, CLEF, and NTCIR. Many of these collections are about news stories and web pages. Existing test collections; however, are not always sufficient to satisfy the different needs of IR research and evaluation. In particular, very few test collections have been developed containing library metadata records.

The digital library community is striving to make broader use of the digital collections created through the hard work of many professionals. Providing multilingual information access (MLIA) to digital collections is considered crucial in order to share information resources worldwide (Perters, Braschler, & Clough, 2012, p192; Borgman, 1997). Test collections on metadata records allow libraries or digital libraries to explore and compare different MLIA models or approaches before they make decisions on how to implement MLIA for their digital collections. Our research aims to build one such test collection.

The purpose of this paper is to describe the methodologies and processes we applied to generate a test collection on metadata records. We first review the literature on test collection construction, then present our considerations on building our test collection on metadata records. Next, we describe in detail the processes and quality control involved in generating the test topics and relevance judgments. The paper concludes with a summary of the components of the test collection and the implications of our study.

2 Related Literature

The three information retrieval and access forums have provided test collections for the IR community. The TREC test collection site (Text Retrieval Conference, 2016) lists the document collections, topics, and relevance judgment used by TREC over years. However, few publications described in detail their creation process. Harman (1993) described the creation of the test collection for the first TREC. It has become a best practice for developing test collections. Creating large test collections, such as those used by TREC, NTCIR, and CLEF, is usually done collaboratively. For example, document collections are usually provided by organizations that hold copyrights, and are then processed by another organization. The topics and relevance judgment are created by the same or different organizations using human judges.

There is few work describing test collections with library records. Lykke, Larsen, Lund, and Ingwersen (2010, March) created a test collection involving physics-related library records, papers, and other objects. The test collection contained a dataset of 18,000 monographic records, 160,000 papers and 275,000 abstracts from the physics domain, 65 topics, and relevance judgment. Their relevance was

judged on 4-point scale: highly, fairly, marginally and non-relevant using a web-based relevance assessment system. This collection is freely available for educational and research purposes (Lykke, Larsen, Lund, & Ingwersen, 2010). It is the only test collection on library records we are aware of. Researchers have built some other test collections. Oard et al. (2004) built a test collection for the retrieval of spontaneous conversational speech. Soboroff, Griffitt, and Strassel (2016) created a test collection for multilingual retrieval from informal discussion forum text, which was built as part of DARPA's Board Operation Language Translation program; Almerekhi, Hasanain, and Elsayed (2016, July) built EveTAR, the first publicly-available Arabic test collection for event detection. They collected a dataset of 590M Arabic tweets, identified 66 events in this dataset, and generated a set of 134K potentially-relevant tweets using crowd-sourcing. The review of the literature helped us to design a framework for test topic generation and the approach for relevance judgment.

3 Methodology

3.1 Methodological Considerations and Workflow

We applied a TREC-like approach to create a test collection on metadata records with the following considerations:

- This test collection, once created, should be available publicly to achieve maximum usefulness. In other words, all items of the collection should have the copyright issues resolved. Because we do not hold the copyrights for any library metadata records, we needed to collaborate with libraries that generated or held the records.
- For topic generation and relevance judgment, we should apply a crowd-sourcing approach to minimize expenses. Especially, we should allow college students who are interested in information retrieval and information technology to participate in the study. Developing a web-based system to create topics for the collection and perform the relevance judgment was necessary for crowd-sourcing.
- The whole test collection should be multilingual so that cross-language information retrieval experiments can be conducted. We chose to have the collection in three languages, English, Chinese, and Spanish, as these are most widely spoken languages on the Internet (Internet World Stats, 2016). Realizing that we don't have the capabilities to manually translate the English metadata records into the two languages, we applied Google and Bing online translation services and provided their translation results as part of the test collection.
- The evaluation of the test collection and its components are important issues and somehow missing in the literature. We should consider different strategies for ensuring the quality of different components of the collection.

Figure 1 shows the workflow of creating our test collection. We collected 1,005,752 metadata records from two digital collections, and had them translated into Chinese and Spanish by Google and Bing, as described in Section 3.2. Then we developed a test topic generation system, recruited eight college students, and selected 45 topics for the collection out of 56 generated ones, as described in Section 3.3. The relevance judgment was conducted after we implemented a web database system, which is described in Section 3.4. The evaluation and quality control for each part are also described in their respective sections.

3.2 Document Collection Extraction, Selection, and Translation

Documents were extracted from two sources: (1) a university catalog that contains more than 2.2 million metadata records for books, audios, videos, and born-digital materials; (2) the Library of Congress Catalog public portal. From each source we randomly extracted about 600,000 metadata records and kept 1,005,752 ones that contained elements such as title, description, authors, publisher, subjects, and coverage. The process is similar to the one described by Azogu and Chen (2013). The major challenges we dealt with included identifying duplicate records or multiple similar records, and removing records that contained texts in languages other than English.

Extracted and selected records were translated by Google and Bing online translation services through their translation API. The major challenges included automatically detecting records that were jumbled by the MT services and removing them from the collection. The machine translated records were included in the test collection so that they could serve as baselines for testing new translation strategies.

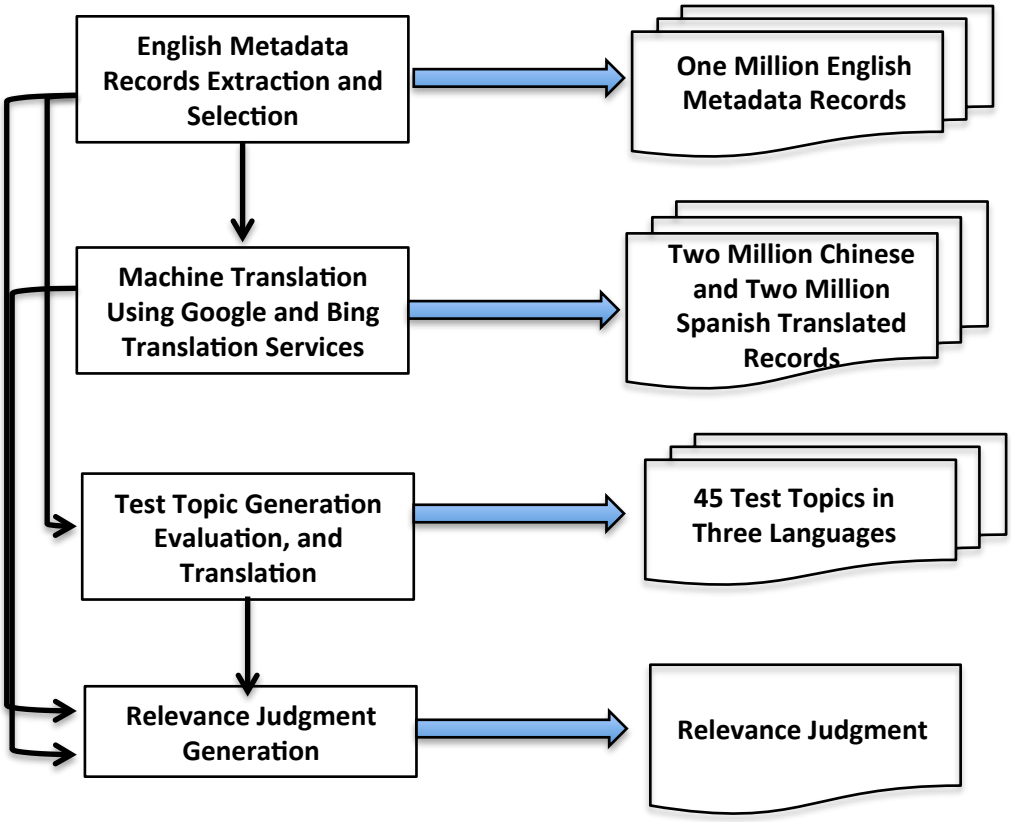

Figure 1. Workflow for Generating The Test Collection

3.3 Topic Generation

To generate topics for the test collection, we first conducted a literature review and determined the principles for this activity: topics should reflect real user needs with relevant documents in the document collection. Simultaneously, topics should be diverse enough to cover the different subject areas of the document collection. We then developed a framework and applied that framework for test topic generation. Specifically, we proposed six criteria to assess individual topics and the final topic set: Unambiguity, No Duplication, Cultural Appropriateness, Diversity, Relevancy, and Complexity (Chen, Namgoong, & Cao, 2016). A six-question survey for soliciting topics was presented to recruited participants through a web-based database system. In total we obtained 56 topics which covered areas such as health, pet care, furniture repair, music, and history. We kept 45 topics out of the 56 using the 6 evaluation criteria. These selected topics were then manually translated into both Chinese and Spanish.

3.4 Relevance Judgment Generation

The next step will be to generate relevance judgments for the topics. Both Indri (The Lemur Project, 2016) and Terrier (University of Glascow, 2012) are used to generate information retrieval results using their multiple IR models and configurations. We will choose about 30 configurations with different models/pseudo relevance feedback to retrieve metadata records from the English document collection. We have created a web application that allows users to judge the relevance of those retrieved records. Participants will be recruited to classify a retrieved metadata record into one of the following four categories: (1) relevant; (2) maybe relevant; (3) irrelevant; (4) not sure. To ensure the quality of the judgment, each retrieved metadata record for each topic will be judged by at least two judges. Inconsistent judgments will be reviewed and determined by a third judge. Any metadata records that are judged “not sure” will be re-judged.

4 Summary of the Test Collection

The test collection, once completed, will contain the following items: (1) An English document collection of 1 million metadata records; (2) Two Chinese translations for each English metadata record obtained from machine translation using Google Translate and Bing Translator; (3) Two Spanish translations of each record obtained from machine translation using Google Translate and Bing Translator; (4) 45 topics in English, Chinese, and Spanish; and (5) A relevance judgment file containing the identification numbers of judged metadata records for each topic. With this test collection, digital library developers can test different cross-language information retrieval strategies with different translation tools and resources.

This paper describes the building of a test collection of library metadata records. As very few such collection is available, this study contributes a new resource to information retrieval and digital library communities for exploring multilingual information access strategies.

5 References

Almerekhi, H., Hasanain, M., & Elsayed, T. (2016, July). EveTAR: A New Test Collection for Event Detection in Arabic Tweets. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 689-692). ACM.

Almerekhi, H., Hasanain, M., & Elsayed, T. (2016). EveTar. [Dataset]. Retrieved from https://www.dropbox.com/sh/3797o85z8g4drwg/AACx5OUKz9HCIf76cIORTwvRa?dl=0.

Azogu, O. & Chen, J. (2013). Randomized sampling: an approach to extraction of metadata records. Proceedings of Iconference 2013, Fort Worth, TX, Feb. 12-15.

Bawden, D., & Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science,* 35(2), 180–191. doi:10.1177/0165551508095781

Case, D. O. (2012). *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Pub Limited.

Charmaz, K. (1998). Grounded Theory: Objectivist and Constructivist Methods. In: N. K. Denzin & Y. S. Lincoln (Eds.), *Strategies of Qualitative Inquiry* (pp. 509-535). Thousand Oaks, London, New Delhi: SAGE Publications.

Chen, J., Namgoong, M., Cao, G. (2016). A framework for test topic generation. Proceedings of Iconference 2015, Philadelphia, PA, March 20-23.

Harman, D. (1993). Overview of the First TREC conference. In ACM SIGIR'93.

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

Linhares, A., & Brum, P. (2007). Understanding our understanding of strategic scenarios: What role do chunks play? *Cognitive Science*, 31(6), 989-1007.http://dx.doi.org/doi:10.1080/03640210701703725

Lykke, M., Larsen, B., Lund, H., & Ingwersen, P. (2010, March). Developing a test collection for the evaluation of integrated search. In European Conference on Information Retrieval (pp. 627-630). Springer Berlin Heidelberg.

Lykke, M., Larsen, B., Lund, H., & Ingwersen, P. (2010). iSearch collection. [Dataset]. Retrieved from http://itlab.dbit.dk/~isearch/.

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., ... & Kharevych, L. (2004, July). Building an information retrieval test collection for spontaneous conversational speech. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 41-48). ACM.

Piper, A. M., & Hollan, J. (2009). Tabletop displays for small group study: Affordances of paper anddigital materials. *Proceedings of ACM CHI '09*, 1227-1236.

Smith, H. (2012). Information as exclusion: Towards a critical understanding of everyday life. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1–11. doi:10.1002/meet.14504901106

Soboroff, I., Griffitt, K., & Strassel, S. (2016, July). The BOLT IR Test Collections of Multilingual Passage Retrieval from Discussion Forums. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 713-716). ACM.

Text Retrieval Conference (TREC). (2016). Test collection. Retrieved from http://trec.nist.gov/data/test_coll.html

The Lemur Project. (2016). Indri (Version 5.11) [Computer software]. Retrieved from http://www.lemurproject.org/indri.php
University of Glascow (2012). Terrier (Version 4.1) [Computer software]. Retrieved from http://www.terrier.org/