An Approach to Ancient-to-modern and Cross-script Information Access for Traditional Mongolian Historical Collections

Batjargal, Biligsaikhan
biligsaikhan@gmail.com
Graduate School of Science and Engineering, Ritsumeikan University, Japan

Khaltarkhuu, Garmaabazar
garmaabazar@gmail.com
Mongolia-Japan Center for Human Resources Development, Mongolia

Kimura, Fuminori
fkimura@is.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

Maeda, Akira
amaeda@is.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

---

The main purpose of this research is to develop a system to keep over 800-year-old historical records written in traditional Mongolian script for future use, to digitize all existing records and to make those valuable data available for public viewing and screening. There are over 50,000 registered manuscripts and historical records written in traditional Mongolian script stored in the National Library of Mongolia. About 21,100 of them are handwritten documents. There are many more manuscripts and books stored in libraries of other countries such as China, Russia and Germany. Despite the importance of keeping 800-year-old historical materials in good condition, the Mongolian environment for material storage is not satisfactory to keep historical records for a long period of time (Tungalag, 2005). We believe that the most efficient and effective way to keep and protect old materials of historical importance while making them publicly available is to digitize historical records and create a digital library.

Mongolia introduced a new writing system (Cyrillic) in 1941. This was a radical change and alienated the traditional Mongolian script. The spelling of words and suffixes in traditional Mongolian script differs from spellings in modern Mongolian (Cyrillic). This is due to traditional Mongolian script preserving a more ancient language while modern Mongolian reflects pronunciation differences in modern dialects. Spellings used in traditional Mongolian script reflect the Mongolian language spoken in the days of Genghis Khan but also contain elements of the ancient Mongolian language spoken before that era. Thus traditional Mongolian has a different grammar and a distinct dialect from modern Mongolian. At present, people use dictionaries with transcribing suffixes to overcome differences.

Recently ancient historical documents in traditional Mongolian script are being digitized and made publicly available, thanks to advances in innovative information technologies, and the popularity of the Internet. In Windows Vista, and later versions, especially in Windows 7, text-display support for traditional Mongolian script and the input locale is enabled. The Uniscribe driver was updated to support OpenType advanced typographic functionality of complex text layouts such as traditional Mongolian script. Traditional Mongolian script is written vertically from top to bottom in columns advancing from left to right.

However, retrieval of the required information in modern Mongolian from traditional Mongolian script documents is not a simple task, due to substantial changes in Mongolian language over time. We want to offer an information retrieval method that considers language difference over time. Our goal in this research is to develop a retrieval system where a user can access cross-period and cross-script databases with a query input in a modern language.

1. Related Research

Much research has been conducted in the last decade on Cross-language information retrieval (CLIR) – a technique to retrieve documents in one language through a query in another language. On the contrary, little research has been completed regarding information retrieval techniques for historical documents. Still less, almost none of the breakthroughs in research

on information retrieval and information access have aimed at retrieving information in the native language from an ancient, cross-period and/or cross-script foreign language documents. Almost none of the CLIR systems has integrated either modern (Cyrillic) or traditional (ancient) Mongolian language due to its research backwardness.

Few approaches (Ernst-Gerlach et al., 2007; Garmaabazar et al., 2008; Kimura et al., 2009) that could be considered a cross-period information retrieval have been proposed. Little research has been completed regarding information retrieval techniques for historical documents. Ernst-Gerlach et al. (2007) developed a retrieval method that considers the spelling differences and variations over time. They focused on modern and archaic German. Kimura et al. (2009) proposed a retrieval method that considers not only language differences over time, but also cultural and age differences in the same language. They focused on retrieval techniques for modern and archaic Japanese. Khaltarkhuu et al. (2008) proposed a retrieval technique that considers cross-period differences in dialect, script and writing systems of the same language. They focused on modern and traditional Mongolian.

2. System Architecture

We propose a simple model to cope with cross-period and cross-script queries. The proposed model is shown in Figure 1.

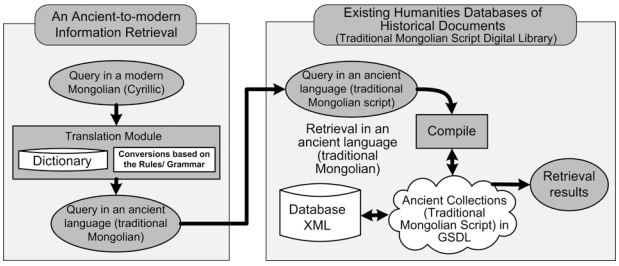

Figure 1

In the first stage, the query in a modern language will be translated into a query in an ancient form. Consequently, a query in the ancient language is submitted as a retrieval query for digital collections. The proposed model is expected to perform cross-period information retrieval and a user will be able to access ancient historical databases with a query input in a modern language. We will adopt a dictionary-based query translation approach.

3. Implementation of the Proposed System

Although it is not easy to develop such a retrieval system, we will utilize existing approaches (Garmaabazar et al., 2008; Kimura et al., 2009) to realize the proposed approach. A prototype called the Traditional Mongolian Script Digital Library[1] (TMSDL) (Garmaabazar et al., 2008), which could be considered a cross-period information retrieval system, has been developed. The retrieval method of the TMSDL considers cross-period differences in writing system of the ancient and modern Mongolian languages.

However, retrieval techniques of the TMSDL have not considered irregular words which have different meanings but are written and pronounced exactly the same in modern Mongolian and have different forms in traditional Mongolian. Also, word sense disambiguation has not been considered. Thus, we enhanced the TMSDL by integrating a dictionary-based cross-period information retrieval approach. We utilized the developing online version of the Tsevel's concise Mongolian dictionary[2] (Tsevel, 1966) under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported license.

The Altan Tobchi (year 1604, 164pp) – A chronological book of ancient Mongolian Kings, Genghis Khan and the Mongol Empire (the largest contiguous empire in history) is shown in Figure 2, with modern Mongolian input interface in the new version of TMSDL (TMSDLv2). A database of such historical records in the TMSDLv2 with English or modern Mongolian query input will help someone conducting research in the history of the High Middle Ages, accessing materials written in an ancient language (traditional Mongolian) in order to understand 13th-14th century history of Asia.

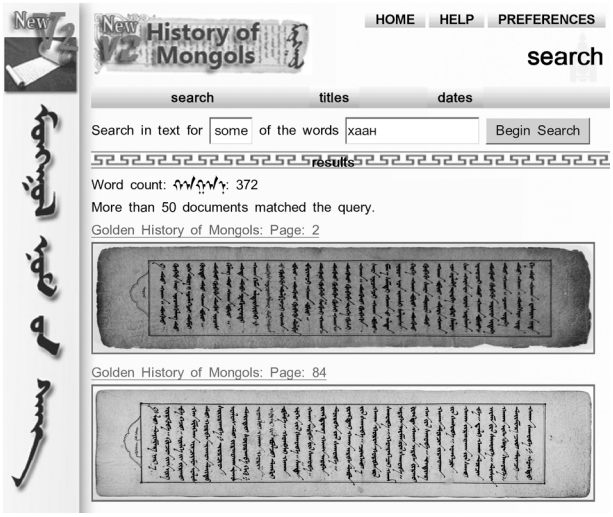

Figure 2

4. Experiments

After the enhancement, we conducted a preliminary experiment to examine the difference between the TMSDL and the TMSDLv2 as well as to check the accuracy of translations from the modern language to the ancient one. We compared our retrieval results from the two versions with the "Qad-un ündüsün quriyangγui altan tobči", (Textological Study) (Choimaa et al., 2002). This textological study contains the detailed analysis of traditional Mongolian words' frequency in the Altan Tobchi. All queries that were retrieved in the TMSDL were retrieved in the TMSDLv2. Therefore, we selected single word queries that were not retrieved in the TMSDL. In addition, selection criteria for the query input are:

- Pronounced or written differently in modern and traditional Mongolian; and
- With higher frequency in the Altan Tobchi.

In the experiment of retrieving traditional Mongolian documents via modern Mongolian (Cyrillic) utilizing a dictionary, we found that the TMSDLv2 translates and retrieves about 86% of selected input queries. However, about 64% of input queries have some variations that are less than or greater than the actual frequency because of the possible errors of translation and text digitization, or limitations of the retrieval function. We are working to distinguish the causes. Our translation module failed to translate 14% of input queries. Improvements on retrieval results are illustrated in Figure 3. Sample retrieval results are shown in Table 1. Retrieved results in the TMSDLv2 with highlights are shown in Figure 4.

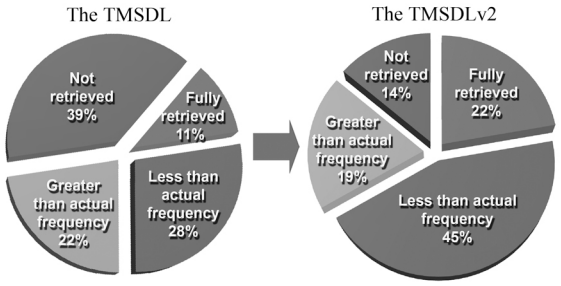

Figure 3

A word in Modern Mongolian (Cyrillic)	Pronunciation		A word in Traditional Mongolian	Meaning, English translation	Retrieved results		Actual frequency, the word count	Retrieval state
	Modern	Ancient			The TMSDL	The TMSDLv2		
хүн	hün	kümün		man, person, human	0	135	135	Fully retrieved
хаан	qaan	qayan		king	0	372	372	
даян	dayan	dayan		all, whole	0	0	32	Not retrieved
сайн	sain	sayin		good, well, fine, nice, pretty	0	0	75	
гэр	ger	ger		ger, yurt, home, residence,family	0	51	47	Greater than actual frequency
төр	tör	törü		law, kingdom, rule	0	50	47	
эзэн	ejen	ejen		lord	0	144	146	Less than actual frequency
богд	bogd	boγda		holy, sacred, divine	0	39	40	

Table 1

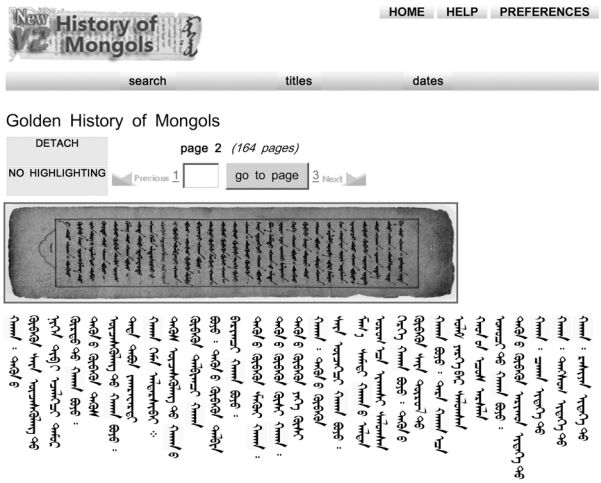

Figure 4

5. Conclusion

In this paper, we proposed a model that utilizes cross-period and cross-script digital collections, which can be used to access old documents written in an ancient language using a query in a modern language. The proposed system is suitable for full text searches on databases containing cross-period and cross-script documents. Such research would involve extensive research in an ancient language

that users and humanities researchers may or may not understand. It could apply to humanities researchers who are conducting research on ancient culture and looking for relevant historical materials written in that ancient language. The proposed model will enable users and humanities researchers to search for such materials easily in a modern language.

However, in our experiment, the TMSDLv2 translates and retrieves about 86% of input queries; only 22% is retrieved without any error. The other 64% has some differences on the number of retrieved query terms. For future development, improvements on translation and retrieval techniques need to be considered to increase the retrieval precision.

For future research, enhancements such as the retrieval of information from two distinct languages and retrieval via single query input from multiple humanities databases in multiple languages need to be developed. Our future work includes evaluating retrieval effectiveness by conducting extensive experiments that consider language differences over time. Performance of this approach will be compared with other approaches.

References

Choimaa, Sh. and Shagdarsuren, Ts. (2002). *Qad-un ündüsün quriyangyui altan tobči, (Textological Study)*. Ulaanbaatar: Centre for Mongol Studies, NUM.

Ernst-Gerlach, A. and Fuhr, N. (2007). 'Retrieval in text collections with historic spelling using linguistic and spelling variants'. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2007)*. Vancouver, Canada, June 2007, pp. 333-341.

Garmaabazar, Kh. and Maeda, A. (2008). 'Developing a Traditional Mongolian Script Digital Library'. *Proceedings of the 11th International Conference on Asia-Pacific Digital Libraries (ICADL2008)*. Bali, Indonesia, December 2008, pp. 41-50.

Kimura, F. and Maeda, A. (2009). 'An Approach to Information Access and Knowledge Discovery from Historical Documents'. *Conference Abstracts of the Digital Humanities 2009 (DH09)*. College Park, MD, June 2009, pp. 359-361.

Tsevel, Y. (1966). *Mongol helnii tovch tailbar toli*. Ulaanbaatar (Mongolian).

Tungalag, D. (2005). *Mongol ulsiin undesnii nomiin san dahi Mongoliin tuuhiin gar bichmeliin nomzuin sudalgaa*. Ulaanbaatar (Mongolian). V. 1.

Notes

1. http://www.dl.is.ritsumei.ac.jp/tmsdl
2. http://toli.query.mn/