A Data-driven Approach for Cross Transformation Between Mongolian texts

Dawa Yidemucao[a] , Muheyat Niyazbek[b]※, Ayjarken Amantay[c]※

School of Information Science and Engineering Xinjiang University, Urumqi, China

[a]idawa@sina.com, [b]muheyatn@xju.edu.cn, [c]jiarektek@163.com

Keywords: Mongolian texts; cross language transformation; DP; data driven approach

Abstract. This paper discusses a data-driven approach to transforming different graphic texts of Mongolian. Using the proposed approach, it is possible to transcribe or translate texts between similar languages such as Mongolian graphic texts used in different regions and countries, as well as the Altaic family languages like Uygur Turkic and Kazakh. The approach has been implemented based on DP (dynamic programming) matching supported by the knowledge-based sequence matching, referred to a multilingual dictionary and a data-driven approach of the target language corpus. Experimental results demonstrate that the proposed method achieves 86.4% transformation accuracy (in F-measure) for the NM (Cyrillic) to the TM (Traditional Mongolian) mainly used in the inner Mongolia, and 91.1% NM to Todo, which is mainly used in Xinjiang areas in China.

Introduction

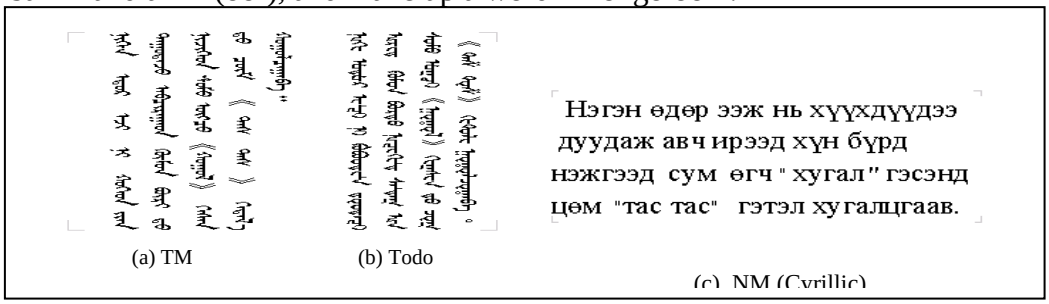
Mongolian people may have difficulties to communicate with each other due to the various texts and dialects used in different areas and countries nowadays [1].

Mongolian belongs to Altaic language family, and it is an agglutinative language. Some examples of the texts printing by different versions are shown in Fig. 1. In Fig. 1, (a) TM (written by the traditional Mongolian scripts, found at the 13th century) is now used mainly in the area of the inner Mongolia; (b) Todo (written by Todo scripts, found at the 17th century) is now used mainly in the Xinjiang area in China; (c) Cyrillic (writing by Cyrillic alphabet, found at the beginning of 20th century), is used in Mongolia and other areas such as Kalmyk and Buryat in Russia today. Figure 2 shows the distribution and population of spoken Mongolian.

For the modern communication via the internet, a transformation system among different Mongolian variations is an urgent need. An example of the sentence alignment by words of TM,Todo and NM is illustrated in Fig. 3. The structure is similar to English with a different SOV grammar, the order of subject, predicate, object and

A phrase transformation pair of TM and Cyrillic is shown in Fig. 4. Similarly, we observe that a word, in either TODO or Cyrillic, corresponds to two or more words of TM, and there is a clear difference in the word formation. This implies that it is rather difficult to transcribe multiple texts by a script unit or word unit or a Unicode showed in Fig. 5.

In Todo and Cyrilli, an entry generally consists of two or three parts: named root (Mongol), inflection suffix and affix (oor), and make up a word “Mongoloor”.



(a) TM (b) Todo (c) NM (Cyrillic)

Figure 1. Mongolian texts writing by different versions

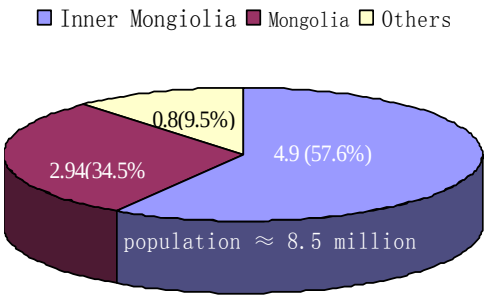


Figure 2. Population ratio of speaking Mongolian

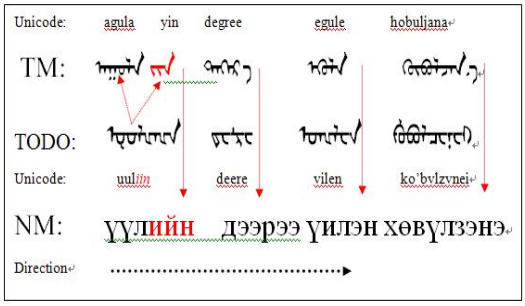


Figure 3. a sentence and word alignment by TM,Todo and NM

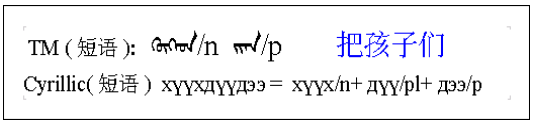


Figure 4. the phrase (NP) by TM and Cyrillic

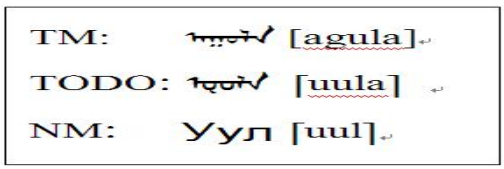


Figure 5. Unicode layout by Mongolian system

The case particle and suffix, also called function words, are usually attached with the root word (item) in Cyrillic and Todo, but they are separated from the root words in TM. This structure is just like Japanese and English. However, it is free in a case of Todo (as shown in Fig. 4).

To create word to word alignments and transformation by mutual ways, some rule based and statistical processing, such as segmentation of the suffixes and syntactic analysis of root word in the case of NM and Todo, are used for Mongolian and other language processing [2]. Currently, researches related to the Mongolian languages and the technical processing among their transformations or translations are rare.

A recent study proposed to using fundamental linguistic rules and a character processing unit for converting Cyrillic to TM texts and vice versa [3]. Although satisfactory conversion results have been reported, the authors also pointed that it was rather difficult to use their approach when the source languages were different and when out-of-rule words occurred frequently.

Another recent study reports a transformation method between two scripts based on the linguistic rules [4]. However, it has been reported that the method has limited capability to transform other texts, such as TM. Additionally, the method cannot be used in the case of unlisted words in a limited corpus.

In this paper, we report a novel data-driven based approach using a target language corpus with DP, which achieves efficient and effective transformation performance.

Approach

A. *System Overview*

The block diagram in Fig. 6 shows the main components of our system, which converts the NM text (source language) to others (target language) in a word by word manner. The main process of this system uses the following three steps:

- step-1: The entries in NM (e.g, Mongol, ajilaasu and bolj in Fig 6) are searched with Dict.MED. If they are found in Dict.MED, a pair is then formed. For example, an entry, “Mongol” was transcribed to “mongol” in the target language.
- step-2: If an entry cannot be found in Dict.MED, the entry is checked whether it is an item appended with a suffix according to the common suffix list (as shown in Table 1). Here, we set 90 commonly used suffixes for NM). If so, the entry is segmented into two parts of root and suffix. Finally, DP matching is performed (described in 2.2) between the root and the entries in Dict.MED.

If a better match is found, the corresponding pairs of both root and suffix from Dict.MED is produced. An example “*mongoloor*” is turned into “*mongol*” and “*oor*” and to “*mongol yer*” shown in Fig. 6.

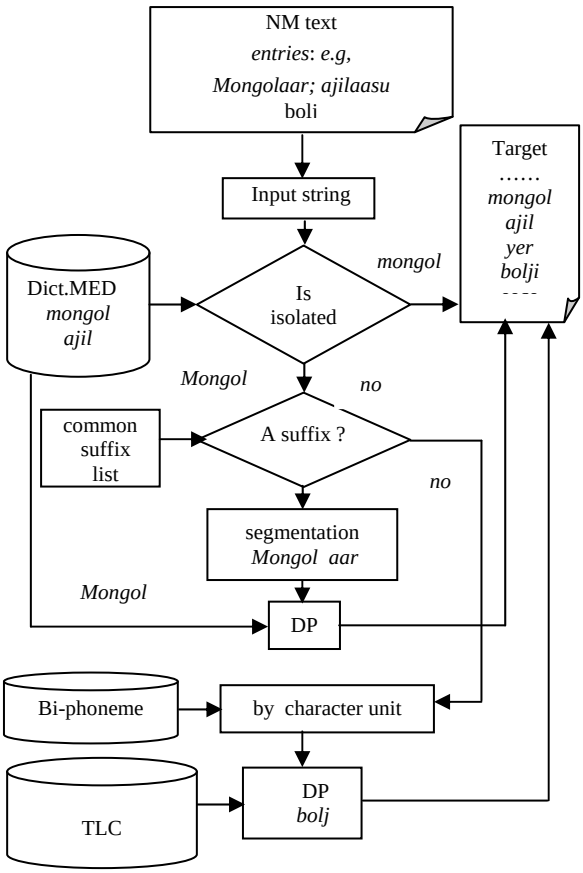


Figure 6. Block diagram of our system

TABLE I. COMMONLY USED SUFFIXES FOR NM

нийнхэн/ нийхээ/ хнээс/ хний/ лийн/ гийн/ лийг/ дохь/ нний/ ийг/ гээр/ гээс/ гүй/ гын хны/ гоор/ лээр/ бээр/ гуй/ чууд/ той/ тэй/ тай/ гоос/ хэнд/ нүүд/ пий/ гаад/ гаар/ гай/ үйс/ нууд/ ханд/ хаад/ ийн/ чүүд/ хий/ өөр/ өөс/ аас/ доо/ аар/ даа/ аан/ дээ/ төө/ нээ/ тээ/ оос/ ээр/ хоо/ хээ/ гөө/ гаа/ ын/ гоо/ ноо/ ып/ дөө/ аад/ үүд/ өөр/ хан/ ууд/ хуу/ тан/ лаа/ ер/ хөн/ хэн/ тоо/ ны/ гээ/ үүн/ хаа/ яа/ ээ/ оо/ ин/ иг/ өө/ аа/ сан/ сэн/ даг/ т/ г/ х/ ч/ д/

- step-3: If step-2 fails, the entry is first converted by a character unit by referring to a Bi-lingual phoneme set. DP matching is then conducted between the converted entries to the target language corpus (TLC). The closest possible match is produced.

B. *Dynamic Programming (DP) Matching*

DP, also known as dynamic time warping (DTW), was introduced for non-linear time alignment of two continuing patterns. DP can effectively minimize errors that occur during the time alignment of the two patterns. Compared with conventional methods of matching two sequences such as edit distance (ED) and longest common subsequence (LCS), DP is more effective because in DP, a character can correspond to more than one character during the matching, and it is more time-efficient than LCS [5,6].

Consider two strings A and B with arbitrary length, say, n and m respectively in equation (1).

$$\begin{cases} A = a_1, a_2, \cdots a_n \\ B = b_1, b_2, \cdots b_m \end{cases} \quad (1)$$

Taking distance $d_n = (i, j)$ between the characters, we initialize them as follows:

$$d_n = \begin{cases} 1 & \text{if } a_i = b_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Then, the matching between strings A and B is regarded as a temporal alignment in a two-dimensional plane. Suppose that the sequence of matched pairs $c_k(i_k, j_k)$ of A and B forms a time warping function F expressed as, $F = c_1, c_2 \cdots c_k$. Let $g_k(c_k)$ denotes the minimized overall distance representing the explicitly accumulated distance from $c_1(1,1)\ to\ c_k(i, j)$. Then, $g_n(c_k) = g_n(i, j)$ can be expressed by equation (3).

$$g_n(i, j) = \min \begin{cases} g_n(i, j-1) + d_n(i, j) \\ g_n(i-1, j-1) + 2 \times d_n(i, j) \\ g_n(i-1, j) + d_n(i, j) \end{cases} \quad (3)$$

Now, if, for example, there are q candidate words to be selected, and the minimized overall distance is given by $D^q_{\min}(A,B) = 1/(n+m)g_q(n-1,m-1)$; then the word will finally be selected by equation (4).

$$D_x = \min\{D^q_{\min}(A, B)\} \quad (4)$$

Notably, the implementation of equation (3) runs in O(n,m) time. Fig. 7 shows an example of DP process for an entry "bolj" described above. In Fig. 7, Two candidates, (t1) and (t2), are given, and the best performance was (t2) for its giving lower overall distance min (n,m) = 0.111.

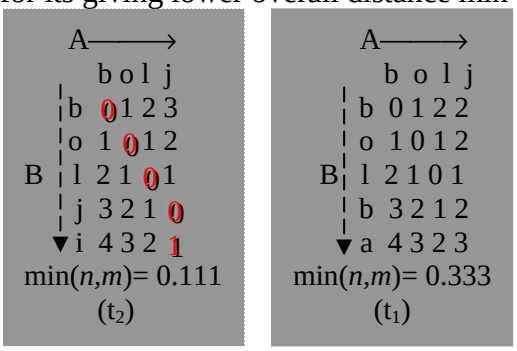


Figure 7. Performances by DP matching for entry "bolj"

Experiments and Results

(1) Data: A parallel corpus of 50,000 sentences was created by referring to a teaching book [7,8] for tests of the NM segmentation and conversion from NM to TM and to Todo, respectively.

(2) Pre-processing:

1) The NM text was first converted into Latin text using a global alphabet set.

2) In many cases, the first character of an NM is written in uppercase. Thus, the initial capital of NM was replaced by a lowercase character.

(3) Test: First, the system picks out a number of entries, which may be appended suffixes, and they are segmented based on the common suffix list (CSL). In this test, the manual comparison accuracy was 37.6%. Next, the entries are searched with Dict.MED (D) and TLC. Finally, a better DP matching result between the entries and TLC, and suffixes are produced.

F-measure expressed by equation (5) was used for test evaluation, and results were listed in Table 2.

$$P(precision) = \frac{\#\ of\ \ words\ \text{by}\ cheked\ \ manually}{\#\ of\ \ produced\ \ items\ by\ \ proposed\ \ method}$$

$$R(recall) = \frac{\#\ of\ \ \ words\ by\ cheked\ \ manually}{\text{all}\ \ entries}$$

$$F = 2 \times P \times R/(P + R) \times 100[\%] \qquad (5)$$

As can be seen from Table 2, although the performances of the proposed method slightly underperformed the previous methods (91.9% for "NM to TM" and 94.3% for "NM to Todo"), the calculation complexity is much lower than the previous method.

TABLE II. CONVERSION RESULTS (NM TO TM/TODO)

	TM			**Todo**		
NM 8,560	***CSL***	***TLC+D***	***TLC+D***	***CSL***	***TLC+D***	***TLC+D***
Acc/ F (%)	manual check (37.6)	83.4	88.9	manual Check (40.2)	86.4	91.1

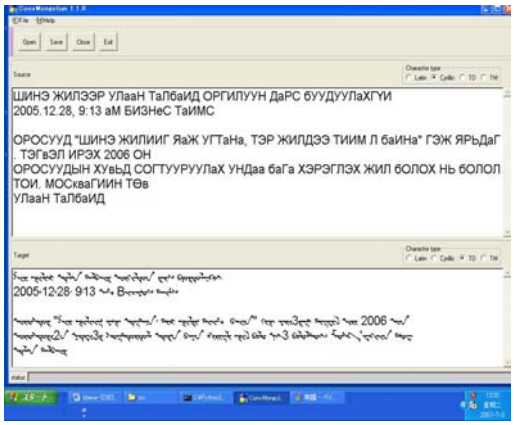


Figure 8. A demonstration of NM text to Todo by character unit

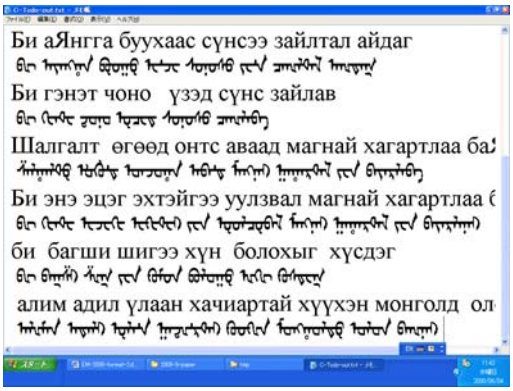


Figure 9. demonstration by proposed method

Text to Todo by character of alignment words. Meanwhile, Fig. 9 shows that the proposed method gives a better performance than that used in the Fig. 8 and several previous methods [9,10].

Conclusion

In this paper, we discussed a novel approach that converts variations of Mongolian languages. The proposed approach was implemented based on DP matching, synthesized TLC and a Dict.MED. For the segmentation and conversion of NM to TM and Todo texts, we obtained mean F-measures of 88.9% and 91.3%, respectively. Although the results show 3% and 2.2% absolute F-measure reductions comparing to the previous method, the proposed system effectively reduces the searching process on the complexity linguistic rule bank.

The Mongolian languages have complex variations in terms of scripts and dialects. There are many problems relating to standardization and natural language processing, which must be overcome. We will further explore that direction in the future.

Acknowledgement

This paper is sponsored by Natural Science Foundation project 2011211A012, Xinjiang.

※Corresponding authors: E-mail: muheyatn@xju.edu.cn;Ayjarken (graduate student) jiarektek@163.com.verb. Each sentence consists of a sequence of entries, which are separated from each other by a space.

References

[1] Ts. SHAGDARSURAN. “Mongolyn utga soyolyn tovchoo” M. Mongolia Ulaanbaatar, 1992.

[2] EHARA Terumasa, *et al*. “Mongolian to Japanese machine translation system C. Proceedings of second international symposium on information and language processing, 2007, pp.27-33.

[3] T.ISHIKAWA, *et,al.* “A Bidirectional Translation Method for the Traditional and Modern Mongolian Scripts” C. Proceeding of the Eleventh Annual Meeting of The Association for Natural Language Processing. 2005,pp.360-363.

[4] Y.NAMSURAI, *et,al*. “The database Structure for BI-Directional Textual Transformation Between Two Mongolian Scripts” C. ICEIC 2006, pp.265-270.

[5] John Coleman. “Introducing Speech and Language Processing” M.COMBRIDGE: Cambridge University Press 2005.

[6] Francois Nicolas, Eric Rivals, “Longest common subsequence problem for unoriented and cyclic strings” J.Theoretical Computer Science 370(2007), pp.1-18.

[7] D.Tserenpil, R.Kullmann, “Mongolian Grammar” M. Mongolia Ulaanbaatar, 2005.

[8] I.Dawa, *et al*, “Multilingual Text –Speech corpus of Mongolian”, CSLP 2006, pp759-770.

[9] Idomucogiin Dawa, Satoshi Nakamura, “A Study on Cross Transformation of Mongolian Family Language”, Journal of Natural Language Processing, J-STAGE, Vol.15 No.5,2008, pp3-21.