Reflection on Textual Transformation between the Similar Languages

Yu Qing & I. Dawa
School of Information Science and Engineering Xinjiang University, 830046, Urumqi

Wang ling &T. Aniwar
Xinjiang Laboratory of Multi-language Information Technology, 830046, Urumqi, China

ABSTRACT: In this paper, we discuss a method of a textual transformation between the similar languages taking Mongolian as an example. The textual transformation approach is performed by combining a knowledge-based rule bank with data driven method. DP algorithm (dynamic programming) is applied to matching of the source and target language words. Our experimental results demonstrate that the proposed method has achieved 83.9% transformation accuracy (in F-measure) from NM (Cyrillic) to TM (Traditional Mongolian) text, and 88.1% for NM to TODO.
KEYWORD: Mongolian language; similar language cross processing; data-driven approach; knowledge-based rule bank; DP algorithm

1 INTRODUCTION

Many writing languages are usually similar in their grammatical and word order. But they are quite different in using character types and lexicon structure such as Mongolian used nowadays in different areas and countries, or Turkic languages such as Uyghur, Kazakh and Turkish[1].

For the languages above, a textual transformation or translation system between the documents writing different scripts is very necessary for their global communication.

In a case of agglutinative languages, for example Mongolian or Turkic, due to changes of the suffixes or affixes linked a verb or word, a converting using word-by-word unit is even difficult just by using a dictionary[2].

It is true that the statistical machine translation (SMT) based on the large amounts of paralleled data is a good hand for their transformation [3]. It is still difficult to provide much paralleled and pre-processing data for the minority languages or less populated languages.

In previous studies, we discussed a method focused on similarity between words based on a dictionary [4], and a SMT method using limited parallel data [5]. In this paper we report another challenging attempt by combining a knowledge-based rule bank with data driven approach. DP algorithm is applied to matching of the source and target language words.

The paper is organized as follows: Section 2 introduces Mongolian writing system and its current situation briefly. Then system algorithm is presented in detail in section 3. Experiment results and discussion are presented in section 4. Finally, some conclusions are drawn in section 5.

2 MONGOLIAN LANGUAGE AND WRITING SYSTEMS

Mongolian language belongs to Altaic language family, and it is an agglutinative language. Because of its historical and geographical background, like some other languages, Mongolian has several dialectal variations in its linguistic phonetic and graphic expressions. Some examples of the texts printing by different scripts are shown in figure 1.

Figure 1 Mongolian writing by different graphics

In Fig. 1, (a) TM (written by the traditional Mongolian scripts, found at the 13th century) is nowadays used mainly in the area of the inner Mongolia; (b) TODO (written by called TODO scripts, found at the 17th century) is used mainly in the Xinjiang area in China and Kalmyk in Russia; and (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.

The sentence order and SOV structure (subject, predicate, object and verb) are same. But the rule of building a word and a way using the function words (suffixes or affixes) are different. An example of the sentence alignment by words of TODO and TM is illustrated in figure 2.

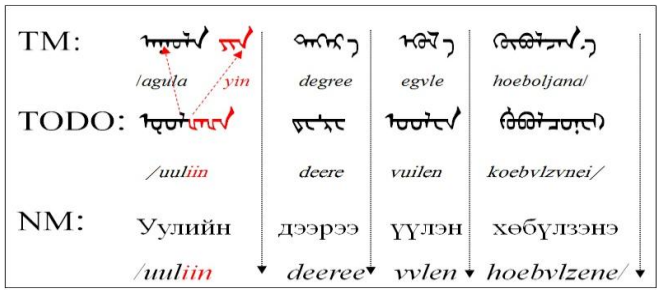

Figure 2 a sentence and word alignment by TODO and TM

Similarly, a phrase transformation pair of TM and NM is shown in figure 3. As shown in Fig.2 and Fig.3, 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 and sequence. This means that it is quite difficult to transcribe the multi-graphic documents between Mongolian by a script (Unicode) unit or word unit. And as shown in Fig. 4, some words, especially in a case of TODO and NM, are very similar when the word is converted into nominal character (Latin or Unicode) forms.

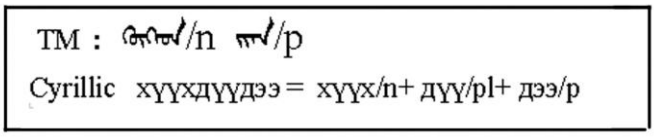

Figure 3 a phrase (NP) by TM and Cyrillic

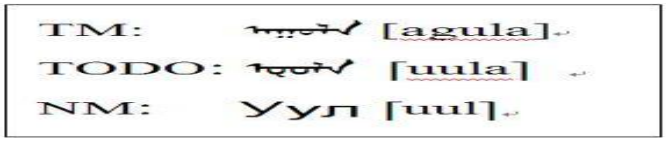

Figure 4 same words [Unicode] by different scripts

To create word-to-word alignments and transformation, 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, will be necessary for Mongolian.

Currently, researches related to the Mongolian natural language processing, especially a textual transformation among their texts are rare.

T.ISHIKAWA research group introduced a performance based on the fundamental linguistic rules and a character unit for converting Cyrillic to TM texts and vice versa [6]. 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 (OOV) words occurred frequently. The report [7] challenged a transformation method between TM and NM two scripts based on the linguistic rules. However, it has been reported that the method has limited capability to transform others, such as TM. Additionally, the method cannot be used in the case of unlisted words in a limited corpus.

3 APPROACH

3.1 *system overview*

The block diagram in Fig. 5 shows the main algorithm of our system, which presents a transformation process using NM as a source language in a word by word manner. And the target language is set in TM. The system process is going on the following three steps:

- step-1: The entries in NM (*e.g*, *Mongol*, *ajilaasu* and *bolj* in Fig.5) are searched with a Dict.MED. If they are found in Dict.MED, a pair is then formed. For example, an entry, "*Mongol*" was transcribed to "*mongol*" of TM.
- 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 and these linguistic rule for NM as a knowledge-based rule bank). If so, the entry is segmented into two parts of root and suffix. Finally, DP matching is performed (described in 3.2) 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 figure 5.
- 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 knowledge-based corpus (TLC). Then, the closest possible match is produced.

3.2 *Dynamic Programming (DP)*

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 [8-9]. Consider two strings A and B with arbitrary length, say, *n* and *m* respectively in equation (1).

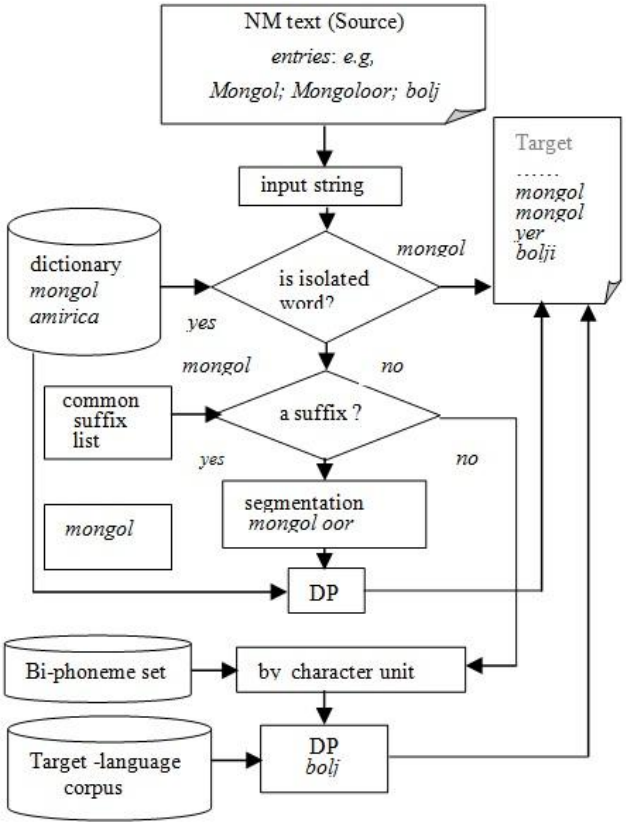

figure 5 system algorithm

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

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 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).

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

Notably, the implementation of equation (3) runs in O(*n*,*m*) time. Fig. 6 shows an example of DP process for an entry "*bolj*" described above. In figure 6, two candidates, (t_1) and (t_2), are given, and the best performance was (t_2) for its giving lower overall distance min (*n*,*m*) = 0.111.

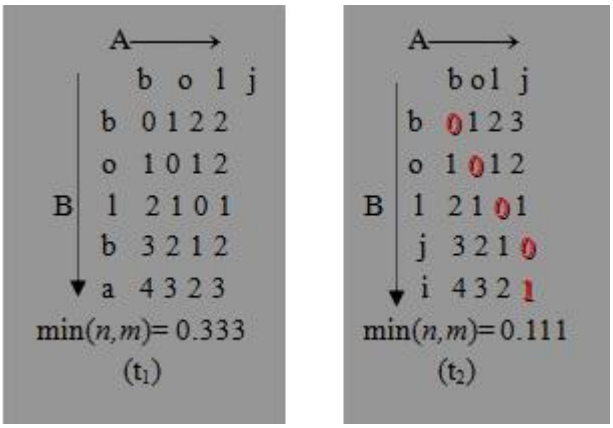

Figure 6 Performances by DP matching for entry "*bolj"*

4 EXPERIMENTS AND RESULTS

(1) Data: A parallel corpus of 50,000 sentences was created by referring to a teaching book [10,11] for tests of the NM segmentation and conversion from NM to TM and to TODO, respectively.

(2) Pre-processing:

①The NM text was first converted into Latin text using a Unicode nominal character Latin alphabet set. ②In many cases, the first character of NM is usually written in uppercase. Thus, the initial capital of NM was replaced by a lowercase character.

(3) Test_1:

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 check(MC) accuracy(ACC) was 37.6%. Next, the entry is searched with Dict.MED (D) and TLC. Finally, a better DP matching between the entry and TLC, and suffixes is produced. F-measure, expressed by equation (5), was used for the evaluation, and results were listed in Table 2.

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

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

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

As can be seen from Table 2, the best performances of the proposed method are 83.9 of NM to TM, and 88.1 in a case of NM to TODO. And it is clear that results from proposed method is close to the test results t(similarities of 86.4 and 87.2 respectively) by SMT, which has processed only 50,000 phrases.

Table 2 conversion results (from NM to TM/TODO)

	TM			TODO			SMT
NM 8,560	CSL	TLC+D	TLC+D	CSL	TLC+D	TLC+D	NM→TM (→TODO)
Acc/F (%)	MC 37.6	80.4	83.9	MC 40.2	85.4	88.1	86.4/(87.2)

The figure 7 and figure 8 showed the real test demonstrations transformation, from TODO to TM and from NM to TODO, by using the proposed method respectively. We can confirm here that the figure 8 gives a better performance than figure 7.

Figure 7 demonstration from TODO to TM

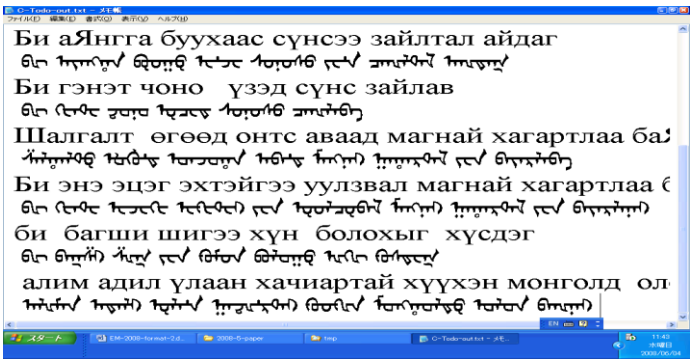

Figure 8 demonstration of NM to TODO

5 CONCLUSION

This paper has discussed a textual transformation method between Mongolian languages. Our system is tested by combining a knowledge-based rule bank with data driven methods. In this test, from TODO to TM and NM to TODO conversion and vice versa respectively, we have obtained mean F-measures of 83.9% and 88.1% respectively. The result is improved by 2.5% in a case of NM to TM compared with the prior-result of 86.4%, and 1.9 % in a case of NM to TODO by SMT approach.

We will further discuss more approaches and challenge cross language transformation between more similar languages combining the linguistic rule with data-driven approach such as SMT.

6 ACKNOWLEDGEMENT

This paper supported by NFSC 61163030 and NSFXG 201291116.

REFERENCES

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

[2] I.Dawa, N. Muheyate, Cross Information Processing Between the Similar Language Texts, Information (an international interdisciplinary Journal), 2014.

[3] 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.

[4] Idomucogiin Dawa, Satoshi Nakamura, A Study on Cross Transformation of Mongolian Family Language, Journal of Natural Language Processing Japan, J-STAGE, 2008, Vol.15 (5), 3-21.

[5] I.Dawa,Wang Xianhui, Mier Adilijiang Maimaiti, An approach of transformation between the traditional and TODO Mongolian texts based on statistical machine translation technology, Journal of The Western Mongolian Studies, 2014, No. 263-71.

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

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

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

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

[10] I.Dawa, *et al.*, Multilingual Text–Speech corpus of Mongolian, CSLP 2006, 759-770.

[11] I.Dawa, *et al*. Processing of Mongolian by Computer, Ojrnal of Chinese Information Processing, Vol(20), 2006, 56-62.