Language Model for Cyrillic Mongolian to Traditional Mongolian Conversion

Feilong Bao, Guanglai Gao, Xueliang Yan, and Hongwei Wang

College of Computer Science, Inner Mongolia University,
Hohhot 010021, China
{csfeilong,csggl,csyxl}@imu.edu.cn, wanghongwei6136@163.com

Abstract. Traditional Mongolian and Cyrillic Mongolian are both Mongolian languages that are respectively used in china and Mongolia. With similar oral pronunciation, their writing forms are totally different. A large part of Cyrillic Mongolian words have more than one corresponds in Traditional Mongolian. This makes the conversion from Cyrillic Mongolian to Traditional Mongolian a hard problem. To overcome this difficulty, this paper proposed a Language model based approach, which takes the advantage of context information. Experimental results show that, for Cyrillic Mongolian words that have multiple correspondence in Traditional Mongolian, the correct rate of this approach reaches 87.66%, thereby greatly improve the overall system performance.

Keywords: Cyrillic Mongolian, Traditional Mongolian, Language Model.

1 Introduction

Mongolia, as a widely used language over different countries and multiple regions, has a significant impact on the world. Its main users are distributed over China, Mongolia and Russia. A major difference between the Mongolian used in China (called Traditional Mongolian) and that used in Mongolia (called Cyrillic Mongolian or Modern Mongolian) is that they have same pronunciation but different written forms.

As a derivative language, Cyrillic Mongolian has both similar grammar and vocabulary to Traditional Mongolian. This means that the conversion of the two languages does not need to follow the traditional machine translation framework. We can just convert the two languages word by word according to their correspondence relationship. A serial of research that focus on the conversion from Cyrillic Mongolian to Traditional Mongolian has been carried out by Bao Sarina, Wuriliga and Hao Li [1-4] et al with either dictionary based approaches or rule based ones and achieved acceptable results. However, none of them have considered the multiple correspondence problems.

Observed that the correct converted word has a strong relationship to its context, we proposed a language model based approach to overcome the multiple correspondence problem. The rest of the paper is organized as follows: section 2 introduces the characteristic of Traditional Mongolian and Cyrillic Mongolian; section 3 depicts in

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 13–18, 2013.

detail the language model based conversion approach; in section 3, experiments and the corresponding results are discussed; at last, we conclude the paper in section 4.

2 Comparison between Traditional Mongolian and Cyrillic Mongolian

Although having a strong relationship to each other, the Traditional Mongolian and Cyrillic Mongolian, as two different languages, still have some significant difference as follows:

1. Tradition Mongolian is composed of 35 characters, in which 8 are vowels and 27 are consonants[5]; Cyrillic Mongolian, on the other hand, has also 35 characters. But 13 of them are vowels and 20 are consonants. Besides, it also includes a harden-character and soften-character[6]. The complete alphabets for the two languages are listed in Tab. 1 for comparison.
2. Cyrillic Mongolian is a case-sensitive language while Traditional Mongolian is not. In Cyrillic Mongolian, the usage of case is similar to English. For the Traditional Mongolian, although it's not sensitive to the case, its form will be different according to the position (top, middle or bottom) in a word [7].

Table 1. Comparison of the characters of Cyrillic Mongolian and Traditional Mongolian

Cyril	Traditional	Cyril	Traditional	Cyril	Traditional	Cyril	Traditional
Аа		Ии		Рр		Шш	
Бб		Йй		Сс		Щщ	
Вв		Кк		Тт		Ъъ	
Гг		Лл		Уу		Ыы	
Дд		Мм		Үү		Ьь	
Ее		Нн		Фф		Ээ	
Ёё		Оо		Хх		Юю	
Жж		Өө		Цц		Яя	
Зз		Пп		Чч			

3. The written direction is different for Cyrillic Mongolian and Traditional Mongolian. For Cyrillic Mongolian, the words are written from left to right and the lines are changed top-down; for Traditional Mongolian, the words are written top-down and the lines are changed from left to right.

4. The degrees of unification between the written form and oral pronunciation are different for Cyrillic Mongolian and Traditional Mongolian. Cyrillic Mongolian is a well-unified language. It has a consistent correspondence between the written form and the pronunciation; on the other hand, however, that for the Traditional Mongolian is not 1-to-1 mapping. Sometimes the vowel or consonant will be dropped, added or transformed when converting the written form to the pronunciation.

In some cases, a Cyrillic Mongolian word would have more than one Traditional Mongolian word corresponded, as shown in Fig. 1, where the three Traditional Mongolian words are different but all correspond to the Cyril word "*асар*".

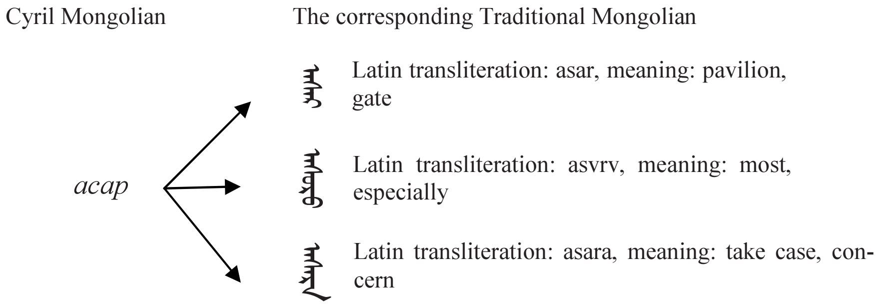

Fig. 1. An example of multiple correspondence for Cyrillic Mongolian to Traditional Mongolian

3 Language Model Based Conversion Approach

Generally speaking, Cyrillic Mongolian and Traditional Mongolian words, when converting, are one-to-one correspondence. However, a large part of Cyrillic Mongolian words have more than one corresponds in Traditional Mongolian. Take the Cyrillic Mongolian sentence "*Танай амар төвшинийг хамгаалхаар явсан юм.*" for example. The words "*амар*" and "*юм*" have more than one correspondences in Traditional Mongolian as shown in Fig. 2, where the corresponding Traditional Mongolian is represented in Latin-transliteration form. More specifically, the Cyril word "*амар*" has four correspondences in Mongolian: "*amara*", "*amar*", "*amar_a*" and "*amvr*"; the Cyril word "*юм*" has two correspondences in Traditional Mongolian: "*yagam_a*" and "*yvm*". The correct conversion for the whole sentence is denoted by the path with the line in bolder, i.e., "*tan-v amvr tobsin-I hamagalahv-bar yabvgsan yvm*" (" ").

If we consider the conversion as a stochastic process and make the final decision according to the probability of the Traditional Mongolian word sequence T conditioned on the Cyrillic Mongolian word sequence *C*, then the conversion problem can be represented as finding the words sequence that satisfies (1):

$$T' = \arg\max_{T \in Q} P(T \mid C) \tag{1}$$

where $T=\{t_1 t_2 \ldots t_m\}$ denotes the possible path and C denotes the Cyrillic Mongolian sentence to be converted.

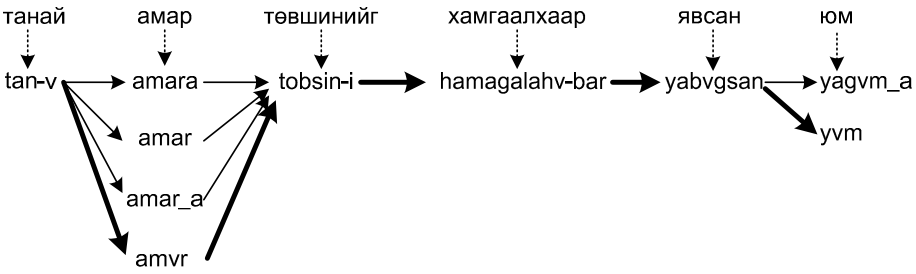

Fig. 2. A conversion example for Cyrillic Mongolian to Traditional Mongolian

As we all know, the conditional probability for $T=\{t_1t_2...t_m\}$ can be decomposed as:

$$P(T \mid C) = P(t_1 \mid C)P(t_2 \mid t_1, C)P(t_3 \mid t_1t_2, C)...P(t_m \mid t_1t_2...t_{m-1}, C) = \prod_{j=1}^{m} P(t_j \mid t_1^{j-1}, C) \quad (2)$$

then formula (1) can be represented as:

$$P(T \mid C) = \arg\max_{T=t_1t_2...t_m \in Q} \prod_{j=1}^{m} P(t_j \mid t_1^{j-1}, C) \quad (3)$$

If we further assume the N-gram language model assumption[8], formulate (3) can then be further simplified as:

$$P(T \mid C) = \arg\max_{T=t_1t_2...t_m \in Q} \prod_{j=1}^{m} P(t_j \mid t_{j-N+1}^{j-1}, C) \quad (4)$$

We use the Maximum Likelihood Estimation to estimate the parameters in (4) and adopt Kneser-ney technique[8] to overcome the sample sparseness problem.

4 Experiment

We take the Conversion Accurate Rate (CAR) as the evaluation metric, which is defined as:

$$CAR = \frac{N_{correct}}{N_{total}} \quad (5)$$

Where $N_{correct}$ denotes the total number of words that are correctly converted and N_{total} denotes the number of all the words need to be converted.

The SRILM is adopted for training the language model[9]. A dictionary that contains the Cyrillic Mongolian word to its multiple correspondences in Traditional Mongolian words is constructed for our experiment. This dictionary has 4679 Cyrillic Mongolian words in total. A Traditional Mongolian text corpus, which contains 154MB text in international standard coding, is adopted for n-gram language model training. We use a Cyrillic Mongolian corpus which contains 10000 sentences to test our approach. This corpus is composed of 87941 words, among which 14663 have

more than one Traditional Mongolian words corresponded. Our conversion progress can be divided into two steps: in the first step, we convert all the Cyrillic Mongolian words to their corresponding Traditional Mongolian words according to the rule-based approach; and then, for each word, we check whether there is only one Traditional Mongolian word generated. If not, we further determine the best one according to the Language Model based approach proposed in section 3. The data set for the rule-based approach is composed of three parts: a mapping dictionary for Cyrillic Mongolian stem to Traditional Mongolian stem, which contains 52830 entries; a dictionary for Cyrillic Mongolian static inflectional suffix to Traditional Mongolian static inflectional suffix, which contains 336 suffixes; and a dictionary for Cyrillic Mongolian verb suffix to Traditional Mongolian verb suffix, which contains 498 inflectional suffixes. Based on the word formation rule of Traditional Mongolian and Cyrillic Mongolian, together with the above mentioned stem mapping dictionary and suffix mapping dictionary, we constructed a rule-based conversion system.

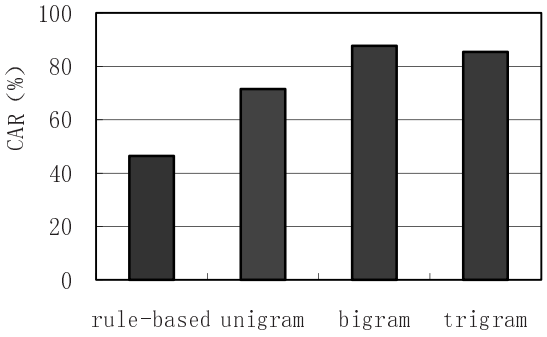

Fig. 3. Performance comparison between the LM based approaches

For the words that have more than one Traditional Mongolian correspondence, we compare the Language Model based approach with different grams (unigram, bigram and trigram) to the rule-based approach. The experiment results are illustrated in Fig 3, from where we can see that all the Language Model based approaches significantly outperform the rule-based approach, among which the bigram achieved the best performance (CAR: 87.66%). Affected by the sample sparseness problem, the trigram approach is slightly worse than the bigram approach, but still much better than the unigram one which has considered only the occurrence frequency, but no context information. This again reconfirm the fact that if the context information is not considered, the performance would be badly decreased.

We also test the overall system performance of rule-based approach and the improved one on all the Mongolian words (both 1-to-1 and 1-to-N). The experimental results are illustrated in Fig 4. In Fig 4, we can see that the conversion correctness for the rule-based approach is 81.66%. When it's integrated with the LM-based approach, the overall system correctness is greatly improved, which reaches 88.14%.

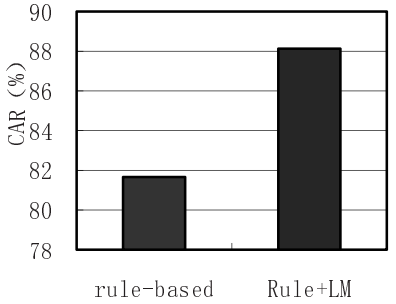

Fig. 4. Overall system performance comparision

5 Conclusions

When converting the Cyrillic Mongolian to the Traditional Mongolian, a lot of problem emerged. In this paper, we focus our attention on the multiple correspondences problem and proposed a language model based conversion approach which takes the context information into consideration. The proposed approach effectively settled this problem and thereby greatly improved the overall conversion system performance. However, there is still some issues to be considered, like the conversion problem for newly-added words and that for the words borrowed from other languages. We will take all these problems as our future work.

Acknowledgements. This work is supported by the Natural Science Foundation of China (NSFC) (NO. 61263037, NO. 71163029) and the Natural Science Foundation of Inner Mongolia of China (NO. 2011ZD11).

References

1. Sarina, B.: The Research on Conversion of Noun and Its Case from Classic Mongolian into Cyrillic Mongolian. Inner Mongolia University, Hohhot (2009)
2. Wuriliga: The Electronic Dictionary Construction of the Traditional Mongolian-Chinese and Cyrillic Mongolian-Chinese. Inner Mongolia University, Hohhot (2009)
3. Li, H., Sarina, B.: The Study of Comparison and Conversion about Traditional Mongolian and Cyrillic Mongolian. In: 2011 4th International Conference on Intelligent Networks and Intelligent Systems, pp. 199–202 (2011)
4. Gao, H., Ma, X.: Research on text-transform of Cyrillic Mongolian to Traditional Mongolian conversion system. Journal of Inner Mongolia University for Nationalities 18(5), 17–18 (2012)
5. Quejingzhabu: Mongolian code. Inner Mongolia University press, Hohhot (2000)
6. Galsenpengseg. Study Reader of Cyrillic Mongolian. Inner Mongolia education press, Hohhot (2006)
7. Qinggeertai. Mongolian Grammar. Inner Mongolia People's Publishing Press, Hohhot (1992)
8. Zong, C.: Statistical Natural Language Processing. Tsinghua University Press, Beijing (2008)
9. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado (2002)