Bidirectional Bengali Script and Meetei Mayek Transliteration of Web Based Manipuri News Corpus

Thoudam Doren Singh
Centre for Development of Advanced Computing (CDAC)
Gulmohor Cross Road No 9, Juhu
Mumbai-400049, India
thoudam.doren@gmail.com

ABSTRACT

The transliteration has attracted interest of several sections of researchers. Several techniques of transliteration have been developed and used – both statistical based approaches and rule based approaches. In the present method, a simple but effective rule based technique is developed for the transliteration between Bengali script and Meetei Mayek script of written Manipuri text. Typically, transliteration is carried out between two different languages –one as a source and the other as a target. But, for the languages which use more than one script, it becomes essential to introduce transliteration between the scripts. This is the reason why the present task is carried out between Bengali script and Meetei Mayek for Manipuri language. The proposed rule based approach points out the importance of deeper linguistic rule integration in the process by making use of the monosyllabic characteristics of Manipuri language. The Bengali script to Meetei Mayek transliteration system based on the proposed model gives higher precision and recall compared to the statistical model. But, in contrast to that, the statistical based approach gives higher precision and recall compared to the rule based approach for the reverse transliteration.

KEYWORDS : Meetei Mayek, Bengali Script, Transliteration, Monosyllabic

1 Introduction

Manipuri language is in eighth schedule of the constitution of India and spoken approximately by three million people mainly in the state of Manipur in India and in the neighbouring countries namely Bangladesh and Myanmar. It is a less privileged Tibeto-Burman language and highly agglutinative in nature, influenced and enriched by the Indo-Aryan languages of Sanskrit origin and English. The Manipuri or Meeteiron is represented using two different scripts, viz. Bengali script and Meetei Mayek (also known as Meitei Mayek). So, it is essential to produce a Manipuri text both in Bengali script as well as Meetei Mayek. The Meetei Mayek is the original script used to represent Manipuri language. It may be noted that Manipuri is the only Tibeto-Burman language which has its own script. We carry out transliteration to cope up with different writing systems by converting from one writing system to another writing system between Bengali script and Meetei Mayek. Transliteration is used in several applications of natural language processing such as machine translation, named entity transliteration, out of vocabulary transliteration and cross lingual information retrieval etc. The proposed rule based approach points out the importance of deeper linguistic rule integration in the transliteration process by making use of the monosyllabic characteristics. Natural language processing tasks for Manipuri language is at the very initial stage and most of the tools available so far do not perform at the required measure and more importantly, there is not enough digitized Manipuri language resources, be it monolingual or bilingual.

2 Related Work

The transliteration models can be categorized as grapheme based, phoneme-based and hybrid based. The grapheme based model (Li et al, 2004) is the direct orthographic mapping and only uses orthography-related features while phoneme based model (Knight and Graehl, 1998) works on phonetic correspondence to generate the target text. The hybrid method refers to the combination of several different models which may use other knowledge source. The papers by (Gao, 2004), (Knight and Graehl, 1998), (Virga and Khudanpur, 2003) report the attempt to build statistical transliteration model. In this approach, the transliteration model performance is limited by what it sees during the training process from the training data. This approach faces one important challenge to disambiguate the noise introduced during the training process for reverse / backward transliteration. The reason behind is that some of the silent syllable may be lost due to the noisy model as in the case of Chinese-English transliteration (Yang et al, 2008). Some of the researchers devised methods to improve the statistical approach by making use of bilingual resources.

Manipuri is a resource constrained language and a bilingual resource is very limited (Singh and Bandyopadhyay, 2010a). The first Manipuri to English transliteration is reported for the named entities using modified joint source channel model (Ekbal et al., 2006) and is used in the parallel corpora extraction from comparable news corpora (Singh and Bandyopadhyay, 2010b) and reused in the Manipuri to English example based machine translation system (Singh and Bandyopadhyay, 2010c) and Phrase Based Statistical Machine Translation (PBSMT) system development (Singh and Bandyopadhyay, 2011). The performance of the rule based approach is improved by integrating syllable based unit as transliteration unit (TU) in the present grapheme based approach. However, Manipuri being a tonal language, there is loss of accents for the tonal words. There is essence of intonation in Manipuri text; but differentiation between Bengali characters such as ি (i) and ী (ee) or ু (u) and ূ (oo) cannot be made using Meetei Mayek. This increases the lexical ambiguity on the transliterated Manipuri words in Meetei Mayek script. To the best of our knowledge, the present task of transliterating from Meetei Mayek to Bengali is the first attempt so far. This attempt is essential for the generation who are accustomed to Bengali script only prior to full fledged reinstatement of Meetei Mayek at schools, offices and administrative levels.

3 The Manipuri News Corpus

The web walked into the ACL meetings starting in 1999 as a source of linguistic data. In recent times, there are online Manipuri news websites available such as http://www.thesangaiexpress.com/ where the news items are published in both Manipuri and English. Some other websites are http://www.ifp.co.in/ published in English; http://www.poknapham.in/ published in Manipuri and http://www.hueiyenlanpao.com/ is published in both English and Manipuri (in both Bengali script and Meetei Mayek script). However, the Manipuri news is available only in PDF format from these websites. A web based Manipuri news corpus collection is reported (Singh and Bandyopadhyay, 2010a) using the Bengali script in Unicode format. The resource constrained Manipuri language news corpus is collected from http://www.thesangaiexpress.com/. At present, there is a Manipuri news monolingual corpus of 4 million wordforms in Bengali script Unicode format. Our experiment makes use of this corpus on news domain. The news items cover national and international news, brief news, editorial, letter to editor, articles, sports etc. The local news coverage is more than the national and international news in these websites. The corpus is tokenized and cleaned to minimise spelling errors.

4 Manipuri scripts and Linguistic Features

Manipuri uses two different scripts – Bengali Script[1] and Meetei Mayek[2] (Unicode range: ABC0-ABFF of Unicode Standard Version 6.0) also known as Meitei Mayek script to some linguists. The Bengali script has 52 consonants and 12 vowels. The Meetei Mayek has 27 letters (Iyek Ipee), 8 dependent vowel signs (Cheitap Iyek), 8 final consonants (Lonsum Iyek), 10 digits (Cheising Iyek) and 3 punctuation (Cheikhei, Lum Iyek and Apun Iyek). The transliteration models convert source text to target text based on the phonetic equivalent mapping. Though there can be direct one-to-one mapping for the 27 Meetei Mayek letter (Iyek Ipee) to Bengali script, there are some Bengali scripts which does not have a one-to-one direct mapping to Meetei Mayek such as (ঋ, ক্ষ, ঃ, ঞ, ঁ etc.) which has resulted in the loss of target representation. This is the most basic problem with grapheme based transliteration system. An important question is – how do we handle these characters in Manipuri text with Bengali script for the transliteration into Meetei Mayek. An expert level agreement on mapping these characters to Meetei Mayek is the need of the hour in order to improve the rules based transliteration performance between the two scripts.

[1] http://unicode.org/charts/PDF/U0980.pdf

[2] http://unicode.org/charts/PDF/UABC0.pdf

5 Bengali script to Meetei Mayek Transliteration

Our method has two advantages. Firstly, there is a sizable monolingual Manipuri news corpus available to be used. Secondly, our method can be improved by including syllable based transliteration unit to the existing list of TU. Concretely, there are two phases involved in our approach. In the first phase, we split the individual words into syllables, and then a syllable based searching can be employed to revise the result. In the second phase, for any syllable unit or TU match, it is searched from the list of TUs and the corresponding mapping unit is picked up from the mapping table and the syllables are concatenated to form the target word. Two different models are developed – baseline and monosyllabic based model.

5.1 Baseline Model

As a baseline system, a grapheme based – that is character to character and numeral to numeral transliteration is carried out. A hand crafted one-to-one mapping rule between Bengali script and Meetei Mayek is established and transliteration is carried out. However, there is no exact one-to-one correspondence between each phoneme associated with each Bengali script grapheme and Meetei Mayek grapheme. A direct one-to-one grapheme for the conjuncts between Bengali script and Meetei Mayek is not possible based on the order of appearance of each character. This model has several drawbacks including the mishandling of conjuncts. Hence, this model does not give a reasonably good transliteration. So, the several NLP tools developed using the transliterated corpus using this model suffers very badly in terms of performance.

5.2 Syllable Based Model

As a baseline system, a grapheme based – that is character to character and numeral to numeral transliteration is carried out. There is no one-to-one correspondence between phoneme associated with the Bengali script grapheme and Meitei Mayek grapheme. So, a hand crafted rule between this is established and transliteration is carried out. This scheme has several drawbacks including the mishandling of conjuncts. There is no direct one-to-one grapheme for the conjuncts as well. Thus, the injective function behaviour does not suit between the two character sets for transliteration. Hence, this model does not give a reasonably good transliteration.

Bengali Script	Meetei Mayek
ষ, স, শ, ছ	ꯁ (Sam)
ন, ণ	ꯅ (Na)
ট, ত	ꯇ (Til)
থ, ঠ	ꯊ (Thou)
য়, য	ꯌ (Yang)
দ, ড	ꯗ (Dil)
ঢ, ধ	ꯙ (Dhou)
উ, ঊ	ꯎ (Un)
ই, ঈ	ꯏ (Ee)
র, ড়, ঢ়	ꯔ (Rai)
◌ি, ◌ী	ꯤ (Inap)
◌ু, ◌ূ	ꯨ(Unap)

TABLE 1 – Many-to-One mapping table.

Based on the above observation, there is no possibility of one-to-one mapping between the two scripts for all the Meitei Mayek scripts as given in table 1, a syllable based transliteration model is developed. One of the most significant shortfalls is that – the conjuncts which are appearing in the Bengali representations need to be addressed as a single transliteration unit. Use of conjuncts using Bengali script of Manipur text is very common and large in numbers. The overall conjunct representation is many-to-many characters in nature for the bilingual transliteration task of Bengali-Manipuri language pair. Some of the example words using the conjuncts are given as:

প্রেসন ←→ ꯄ꯭ꯔꯦꯁꯅ (*press-na*)
ডিষ্ট্রিককি ←→ ꯗꯤꯁ꯭ꯇ꯭ꯔꯤꯀꯀꯤ (*district-ki*)
সেক্রেটরিয়েত্তা ←→ ꯁꯦꯛ꯭ꯔꯦꯇꯔꯤꯌꯦꯠꯇꯥ (*secretariate-ta*)
পেট্রোল ← → ꯄꯦꯠ꯭ꯔꯣꯜ (*petrol*)

And the Bengali script conjuncts and its constituents along with the Meetei Mayek notation for the above examples are as given below:

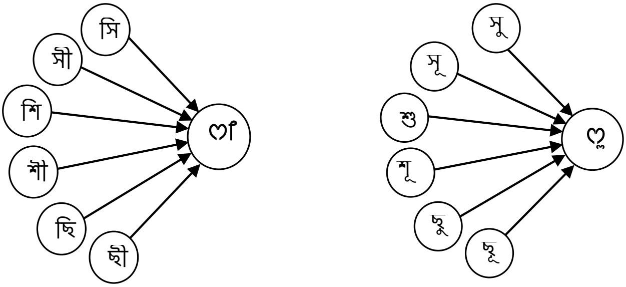
প্রে → প + ্র + ে → ꯄ꯭ꯔꯦ
ষ্ট্রি → ষ + ট + ্র + ি → ꯁ꯭ꯇ꯭ꯔꯤ
ক্রে → ক + র + ে → ꯛ꯭ꯔꯦ
ট্রো → ট + র + ো → ꯇ꯭ꯔꯣ

One of the observations is that vowels presentation of Bengali and Meetei Mayek is not in the same order. Again, the ে (ꯦ–e) is a constituent of ো (ꯣ–o) and ৌ (ꯧ -ou) . This may result in the misinterpretation in case of character by character transliteration. Over and above, the conjuncts need to be fragmented down towards its basic units. In such occurrences the one-to-one mapping fails to work as given in the order. So, a better solution to this problem is addressed in the proposed syllable based model. Since Manipuri is monosyllabic and highly agglutinative, it is essential to analyze at the syllable level for each word. Each syllable can be represented by a group of characters also termed as transliteration unit (TU). Inside this TU, there can be conjuncts such as (ঙ+ক = ঙ্ক) .In this syllable based model, the monosyllabic TU with the highest number of characters are picked from a table, i.e., a many-to-many model. This table maintains the Bengali TU and its corresponding Meetei Mayek representation as shown in table 2. The step is repeated towards the smaller conjuncts up to single character level. This technique is specifically effective for the transliteration of the same language with different scripts.

Bengali Script	Meetei Mayek
ট্রো	ꯇ꯭ꯔꯣ
ক্রে	ꯛ꯭ꯔꯦ
প্রে	ꯄ꯭ꯔꯦ।
ষ্ট্রি	ꯁ꯭ꯇ꯭ꯔꯤ

TABLE 2 – Sample transliteration unit mapping between Bengali and Meitei Mayek scripts

The mapping tables have been prepared at different levels with separate tables for single characters and conjuncts with two or more than two characters. The single character mapping table contains 72 entries and the multiple characters mapping table consists of 738 entries. There are conjuncts of 2, 3 and 4 characters. Sub-tables for each of the conjuncts are prepared. The many-to-many character based syllables are collected from the news corpus based on its occurrence and iterative procedure till it covers the written text in the corpus followed by one-to-one mapping. First of all, we split the transliteration candidate TC_i in the word list into syllables, {s1, s2,…..sn}. Then this syllable sequence is used as a query for syllable-based searching and mapping using the mapping table. An input word can be represented by a vector of syllables {s1, s2,…..sn}. The syllables are the units to construct a word. The algorithm for search and map starts using the mapping table with highest degree of many-to-many mapping table. This reduces the overall complexity still keeping the performance high. Figure 1 shows two examples of convergence due to phonetic similarity of TU from Bengali script to Meetei Mayek.



(a) (b)

FIGURE 1 – Two examples of convergence of TU from Bengali Script to Meitei Mayek.

5.3 Modified Joint Source Channel

The problem of machine transliteration has been studied extensively in the paradigm of the noisy channel model. A Bengali script to Meetei Mayek script transliteration of Manipuri news text is developed based on Modified Joint Source Channel Model for transliteration (Ekbal et al, 2006). A bilingual training set of Bengali script and their respective Meetei Mayek transliterations, has been created. This bilingual training set is automatically analyzed to acquire mapping knowledge in order to transliterate new Bengali script to Meetei Mayek. Transliteration units (TUs) are extracted from the Bengali script to Meetei Mayek counterparts. Some examples are given below:

(a) মনিপুর (ꯃꯅꯤꯄꯨꯔ) [*manipur*] → ম । নি । পু । র

ꯃꯅꯤꯄꯨꯔ → ꯃ | ꯅꯤ | ꯄꯨ | ꯔ

রাজকুমার (ꯔꯥꯖꯀꯨꯃꯥꯔ) [*rajkumar*] → রা । জ । কু । মা । র

ꯔꯥꯖꯀꯨꯃꯥꯔ → ꯔꯥ | ꯖ | ꯀꯨ | ꯃꯥ | ꯔ

(b) অভিনন্দন (ꯑꯚꯤꯅꯟꯗꯟ) [*abhinandan*] → অ । ভি । ন । ন্দ । ন

ꯑꯚꯤꯅꯟꯗꯟ → ꯑ | ꯚꯤ | ꯅ |_ꯟꯗ| ꯟ

The TUs are the lexical units for machine transliteration. The Bengali script is divided into Transliteration Units (TU) with patterns C+M, where C represents a consonant or a vowel or a conjunct and M represents the vowel modifier or matra. The system learns mappings automatically from the bilingual training set of 20,000 entries. Aligned TUs along with their contexts are automatically derived from this bilingual training set to generate the collocation statistics.Transliteration units (TUs) are extracted from the Bengali script and the corresponding Meetei Mayek words, and Bengali script TUs are associated with their Meetei Mayek counterparts along with the TUs in context.

6 Meetei Mayek to Bengali Script Transliteration

The reverse/backward transliteration of Bengali to Meetei Mayek is carried out using the same mapping tables. This Meetei Mayek to Bengali transliteration faces several challenges using linguistics rules. Based on the table 2, the many-to-one mapping is one-to-many mapping for Meetei Mayek to Bengali script transliteration thus it suffers very badly due to probability distribution based on the number of target characters. A one-to-one map is not feasible option. A step to enhance the performance of the transliteration is to consider the preceding few characters and following few characters of the TU to be transliterated as a context in order to forecast the most likely target TU. This enables to choose the right mapping by a certain factor. However, in case of a single character TU, this approach does not help much.

In the Meetei Mayek to Bengali transliteration, there are certain limitations which are caused by many-to-one character mapping and identification and addressing of the right representation of Bengali. However, this issue can be further addressed using possible language model and its probability parameters. One typical example of many-to-one transliteration such as ষ, স, শ and ছ to ꯁ (as given in the Table 2) which results in the lost of tone that could affect other NLP activities such as various lexical ambiguity of word sense disambiguation, POS tagging etc using the transliterated corpus. The convergence shown in figure 1 has to be resolved for the backward transliteration through a statistical approach since the exploration and establishment of rules is quite difficult for such cases for the Meetei Mayek to Bengali script transliteration process. Similarly, there are several other cases of many-to-one and one-to-many sets in between Bengali script and Meetei Mayek. For such instances, the transliteration deems to face the difficulty to make the right choice of the target TU using the rule based approach.

7 Evaluation

Once the experimental corpus is decided, a metric to measure the system's precision and recall is required. The appropriate metric depends on the scenario in which the transliteration system is to be used. The human evaluation takes enormous time compared to automatic evaluation. In our present task, we used 50,000 wordforms as gold standard reference test set. Table 3 gives the precision and recall of the different transliteration systems.

Transliteration model	Baseline Model		Syllable Based Model		Modified Joint Source Channel Model	
	Precision	Recall	Precision	Recall	Precision	Recall
Bengali to Meetei Mayek	68.31	65.23	96.23	94.57	93.23	92.45
Meetei Mayek to Bengali	59.56	58.87	89.78	88.56	93.43	91.97

TABLE 3 – Precision and Recall of the Transliteration Models

Conclusion and Discussion

A bidirectional transliteration system between Bengali script and Meetei Mayek of Manipuri text is developed exploiting the monosyllabic characteristics of Manipuri language. The system abruptly outperforms compared to the baseline transliteration systems of this language using the given two scripts. Since, the Meetei Mayek script is in the process of induction at different administrative and academic levels, there is a need still for Meetei Mayek to be transliterated back to Bengali script to reach other section of Manipuri speakers who know only the Bengali script. The grapheme based approach works well for the transliteration of the same language using the two different scripts. In future, statistical model for transliteration between Bengali script and Meetei Mayek can be experimented using a decent size of training data collected from different sources. Over and above, the Meetei Mayek to Bengali script transliteration has the shortfall using the same TU list using the linguistics rules and a statistical approach is the alternative to yield a better result.

Acknowledgments

I, sincerely, thank Dr. Zia Saquib, Executive Director, CDAC (Mumbai), Prof. Sivaji Bandyopadhyay, Jadavpur University, Kolkata and the anonymous reviewers for their support and valuable comments.

References

Ekbal, A., Sudip K. N., Bandyopadhyay, S. (2006). A Modified Joint Source-Channel Model for Transliteration, *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Pages 191–198, Sydney.

Gao, W. (2004). Phoneme-based Statistical Transliteration of Foreign Name for OOV Problem. *A thesis of Master*. The Chinese University of Hong Kong.

Knight, K. and Graehl, J. (1998). Machine Transliteration. *Computational Linguistics* 24(4).

Li, H., Zhang, M., and Su, J. (2004). A joint source-channel model for machine transliteration. *In Proc. 42nd ACL Annual Meeting*, pages 159–166, Barcelona, Spain.

Singh, T. D., Bandyopadhyay, S. (2010a). Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM, *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, the 23rd International Conference on Computational Linguistics (COLING), Pages 35–42, Beijing.

Singh, T. D., Bandyopadhyay, S. (2010b) Semi Automatic Parallel Corpora Extraction from Comparable News Corpora, *In the International Journal of POLIBITS*, Issue 41 (January – June 2010), ISSN 1870-9044, Pages 11-17.

Singh, T. D., Bandyopadhyay, S. (2010c). Manipuri-English Example Based Machine Translation System, *International Journal of Computational Linguistics and Applications* (IJCLA), ISSN 0976-0962, Pages 147-158.

Singh, T. D., Bandyopadhyay, S. (2011). Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System, *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Pages 1304–1312, Chiang Mai, Thailand, November 8 – 13.

Virga, P., Khudanpur, S. (2003). Transliteration of proper names in cross-lingual information retrieval. *In Proc. of the ACL workshop on Multilingual Named Entity Recognition*.

Yang, F., Zhao, J., Zou, B., Liu, K., Liu, F. (2008). Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages, *Proceedings of ACL-08: HLT*, pages 541–549, Columbus, Ohio, USA, June.