Taste of Two Different Flavours: Which Manipuri Script Works Better for English-Manipuri Language Pair SMT Systems?

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

Abstract

The statistical machine translation (SMT) system heavily depends on the sentence aligned parallel corpus and the target language model. This paper points out some of the core issues on switching a language script and its repercussion in the phrase based statistical machine translation system development. The present task reports on the outcome of English-Manipuri language pair phrase based SMT task on two aspects – a) Manipuri using Bengali script, b) Manipuri using transliterated Meetei Mayek script. Two independent views on Bengali script based SMT and transliterated Meitei Mayek based SMT systems of the training data and language models are presented and compared. The impact of various language models is commendable in such scenario. The BLEU and NIST score shows that Bengali script based phrase based SMT (PBSMT) outperforms over the Meetei Mayek based English to Manipuri SMT system. However, subjective evaluation shows slight variation against the automatic scores.

1 Introduction

The present finding is due to some issue of sociolinguistics phenomenon called digraphia - a case of Manipuri language (a resource constrained Indian languages spoken mainly in the state of Manipur) using two different scripts namely Bengali script[1] and Meetei Mayek[2]. Meetei Mayek (MM) is the original script which was used until the 18th century to represent Manipuri text. Its earliest use is dated between the 11th and 12th centuries CE[3]. Manipuri language is recognized by the Indian Union and has been included in the list of 8th scheduled languages by the 71st amendment of the constitution in 1992. In the recent times, the Bengali script is getting replaced by Meetei Mayek at schools, government departments and other administrative activities. It may be noted that Manipuri is the only Tibeto-Burman language which has its own script. Digraphia has implications in language technology as well despite the issues of language planning, language policy and language ideology. There are several examples of languages written in one script that was replaced later by another script. Some of the examples are Romanian which originally used Cyrillic then changed to Latin; Turkish and Swahili began with the Arabic then Latin, and many languages of former Soviet Central Asia, which abandoned the Cyrillic script after the dissolution of the USSR. The present study is a typical case where the natural language processing of an Indian language is affected in case of switching script.

Manipuri is a monosyllabic, morphologically rich and highly agglutinative in nature. Tone is very prominent. So, a special treatment of these tonal words is absolutely necessary. Manipuri language has 6 vowels and their tone counterparts and 6 diphthongs and their tone counterparts. Thus, a

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

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

[3] http://en.wikipedia.org/wiki/Meitei_language

Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 11–18,
Atlanta, Georgia, 13 June 2013.

Manipuri learner should know its tone system and the corresponding word meaning.

Natural language processing tasks for Manipuri language is at the initial phase. We use a small parallel corpus and a sizable monolingual corpus collected from Manipuri news to develop English-Manipuri statistical machine translation system. The Manipuri news texts are in Bengali script. So, we carry out transliteration from Bengali script to Meetei Mayek as discussed in section 3. Typically, transliteration is carried out between two different languages –one as a source and the other as a target. But, in our case, in order to kick start the MT system development, Bengali script (in which most of the digital Manipuri text are available) to Meetei Mayek transliteration is carried out using different models. The performance of the rule based transliteration is improved by integrating the conjunct and syllable handling module in the present rule based task along with transliteration unit (TU). However, due to the tonal characteristic of this language, there is loss of accents for the tonal words when getting translated from Bengali script. In other words, there is essence of intonation in Manipuri text; the 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.

2 Related Work

Several SMT systems between English and morphologically rich languages are reported. (Toutonova et al., 2007) reported the improvement of an SMT by applying word form prediction models from a stem using extensive morphological and syntactic information from source and target languages. Contributions using factored phrase based model and a probabilistic tree transfer model at deep syntactic layer are made by (Bojar and Hajič, 2008) of English-to-Czech SMT system. (Yeniterzi and Oflazer, 2010) reported syntax-to-morphology mapping in factored phrase-based Statistical Machine Translation (Koehn and Hoang, 2007) from English to Turkish relying on syntactic analysis on the source side (English) and then encodes a wide variety of local and non-local syntactic structures as complex structural tags which appear as additional factors in the training data. On the target side (Turkish), they only perform morphological analysis and disambiguation but treat the complete complex morphological tag as a factor, instead of separating morphemes. (Bojar et al., 2012) pointed out several pitfalls when designing factored model translation setup. All the above systems have been developed using one script for each language at the source as well as target.

Manipuri is a relatively free word order where the grammatical role of content words is largely determined by their case markers and not just by their positions in the sentence. Machine Translation systems of Manipuri and English is reported by (Singh and Bandyopadhyay, 2010b) on development of English-Manipuri SMT system using morpho-syntactic and semantic information where the target case markers are generated based on the suffixes and semantic relations of the source sentence. The above mentioned system is developed using Bengali script based Manipuri text. SMT systems between English and morphologically rich highly agglutinative language suffer badly if the adequate training and language resource is not available. Not only this, it is important to note that the linguistic representation of the text has implications on several NLP aspects not only in machine translations systems. This is our first attempt to build and compare English-Manipuri language pair SMT systems using two different scripts of Manipuri.

3 Transliterated Parallel Corpora

The English-Manipuri parallel corpora and Manipuri monolingual corpus collected from the news website www.thesangaiexpress.com are based on Bengali script. The Bengali script has 52 consonants and 12 vowels. The modern-day Meetei Mayek script is made up of a core repertoire of 27 letters, alongside letters and symbols for final consonants, dependent vowel signs, punctuation, and digits. Meetei Mayek is a Brahmic script with consonants bearing the inherent vowel and vowel matras modifying it. However, unlike most other Brahmi-derived scripts, Meetei Mayek employs explicit final consonants which contain no final vowels. The use of the killer (which refers to its function of *killing* the inherent vowel of a consonant letter) is optional in spelling; for example, while র্দ্র may be read *dara* or *dra*, র্দ্র must be read *dra*. Syllable initial combinations for vowels can

occur in modern usage to represent diphthongs. 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).

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

There is no possibility of direct one-to-one mapping for the 27 Meetei Mayek letter (Iyek Ipee) to Bengali script as given by table 1, over and above some of Bengali scripts which does not have a corresponding direct mapping to Meetei Mayek such as (ঋ, ক্ষ, ঃ, ঞ, ঁ etc.) which has resulted in the loss of target representation. The syllable based Bengali script to Meetei Mayek transliteration system outperforms the other known transliteration systems in news domain between the two scripts in terms of precision and recall (Singh, 2012). The overall conjunct representation is many-to-many characters in nature for the bilingual transliteration task of Bengali script and Meetei Mayek. Some of the example words using the conjuncts are given as:

প্রেসন ←→ ꯄ꯭ꯔꯦꯁꯅ (*press-na*)

ডিষ্ট্রিককি ←→ ꯗꯤꯁ꯭ꯇ꯭ꯔꯤꯛꯀꯤ (*district-ki*)

সেক্রেটরিয়েত্তা ←→ ꯁꯦꯀ꯭ꯔꯦꯇꯔꯤꯌꯦꯠꯇꯥ (*secretariate-ta*)

পেট্রোল ← → ꯄꯦꯇ꯭ꯔꯣꯜ (*petrol*)

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

প্রে (pre) → প + ্র + ে → ꯄ꯭ꯔꯦ

ষ্ট্রি (stri) → ষ + ট + ্র + ি → ꯁ꯭ꯇ꯭ꯔꯤ

ক্রে (cre) → ক + র + ে → ꯀ꯭ꯔꯦ

ট্রো (tro) → ট + র + ো → ꯇ꯭ꯔꯣ

A sample transliterated parallel corpus between English and Manipuri is given in the table 2. These transliterations are based on the syllable based model.

English	On the part of the election department, IFCD have been intimidated for taking up necessary measures.
Manipuri in Bengali Script	ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়খত্নবা খঙহনখ্রে .
Manipuri in Meetei Mayek	ꯏꯂꯦꯛꯁꯟ ꯗꯤꯄꯥꯔꯠꯃꯦꯟꯇꯀꯤ ꯃꯥꯌꯀꯩꯗꯒꯤ ꯑꯥꯏꯑꯦꯐꯁꯤꯗꯤꯗꯥ ꯗꯔꯀꯥꯔ ꯂꯩꯕ ꯊꯕꯛ ꯄꯥꯌꯈꯠꯅꯕ ꯈꯪꯍꯟꯈ꯭ꯔꯦ ॥
Gloss	*election departmentki maykeidagee IFCDda darkar leiba thabak paykhatnaba khanghankhre .*
English	In case of rising the water level of Nambul river, the gate is shut down and the engines are operated to pump out the water.
Manipuri in Bengali Script	করিগুম্বা নম্বুল তুরেলগী ঈশিং ইমায় ৱাংখত্লক্লবদি গেট অসি থিংলগা ঈশিং অসি ইঞ্জিননা ওইনা চিংথোক্লগা লাকপনি হায়রি .
Manipuri in Meetei Mayek	ꯀꯔꯤꯒꯨꯝꯕ ꯅꯝꯕꯨꯜ ꯇꯨꯔꯦꯜꯒꯤ ꯏꯁꯤꯡ ꯏꯃꯥꯌ ꯋꯥꯡꯈꯠꯂꯛꯂꯕꯗꯤ ꯒꯦꯠ ꯑꯁꯤ ꯊꯤꯡꯂꯒꯥ ꯏꯁꯤꯡ ꯑꯁꯤ ꯏꯟꯖꯤꯅꯅꯥ ꯑꯣꯏꯅꯥ ꯆꯤꯡꯊꯣꯛꯂꯒꯥ ꯂꯥꯛꯄꯅꯤ ꯍꯥꯌꯔꯤ ॥
Gloss	*karigumba nambul turelgi eesing eemay waangkhatlaklabadi gate asi thinglaga eesing asi enginena oyna chingthoklaga laakpani hayri.*
English	The department has a gate at Samushang meant for draining out the flood water of Lamphelpat.
Manipuri in Bengali Script	শামুশঙদা ডিপার্টমেন্ট অসিগী গেট অমা লম্ফেলপাৎকি ঈশিং চিংথোক্নবা থম্মী .
Manipuri in Meetei Mayek	ꯁꯥꯃꯨꯁꯪꯗꯥ ꯗꯤꯄꯥꯔꯠꯃꯦꯟꯇ ꯑꯁꯤꯒꯤ ꯒꯦꯠ ꯑꯃꯥ ꯂꯝꯐꯦꯜꯄꯥꯠꯀꯤ ꯏꯁꯤꯡ ꯆꯤꯡꯊꯣꯛꯅꯕ ꯊꯝꯃꯤ ॥
Gloss	*samusangda department asigee gate ama lamphelpatki easing chingthoknaba thammee.*

Table 2. Transliterated texts of English – Manipuri Parallel Corpora and the corresponding Gloss

4 Building SMT for English-Manipuri

The important resources of building SMT are the training and language modeling data. We use a small amount of parallel corpora for training and a sizable amount of monolingual Manipuri and English news corpora. So, we have two aspects of developing English-Manipuri language pair SMT systems by using the two different scripts for Manipuri. The moot question is which script will perform better. At the moment, we are developing only the baseline systems. So, the downstream tools are not taken into account which would have affected by way of the performance of the script specific tools other than the transliteration system performance used in the task. In the SMT development process, apart from transliteration accuracy error, the change in script to represent Manipuri text has made the task of NLP related activities a difference in the way how it was carried out with Bengali script towards improving the factored based modes in future as well. Lexical ambiguity is very common in this language mostly due to tonal characteristics. This has resulted towards the requirement of a word sense disambiguation module more than before. This is because of a set of difference in the representation using Meitei Mayek. As part of this ongoing experiment, we augment the training data with 4600 manually prepared variants of verbs and nouns phrases for improving the overall accuracy and help solving a bit of data sparsity problem of the SMT system along with an additional lexicon of 10000 entries between English and Manipuri to handle bits of data sparsity and sense disambiguation during the training process. The English-Manipuri parallel corpus developed by (Singh and Bandyopadhyay, 2010a) is used in the experiment. Moses[4] toolkit (Koehn, 2007) is used for training with GIZA++[5] and decoding. Minimum error rate training (Och, 2003) for tuning are carried out using the development data for two scripts. Table 3 gives the corpus statistics of the English-Manipuri SMT system development.

4.1 Lexical Ambiguity

Manipuri is, by large, a tonal language. The lexical ambiguity is very prominent even with Bengali script based text representation. The degree of ambiguity worsens due to the convergence as shown by the figure 1 and many to one mapping shown in the table 1. So, the Bengali script to Meetei Mayek transliteration has resulted to the lost of several words meaning at the transliterated output. Many aspects of translation can be best explained at a morphological, syntactic or semantic level. This implies that the phrase table and target language model are very much affected by using Meetei Mayek based text and hence the output of the SMT system. Thus, lexical ambiguity is one major reason on why the transliterated Meetei Mayek script based PBSMT suffers comparatively. Three examples of lexical ambiguity are given below:

(a)
মি (*mi*) → *spider* → ꯃꯤ (*mi*) meaning either *spider* or *man*

মী (*mee*) → *man* → ꯃꯤ (*mi*) meaning either *spider* or *man*

(b)
শূবা (*sooba*) → to work → ꯁꯨꯕ (*suba*) meaning either *to work* or *to hit*

শুবা (*suba*) → to hit → ꯁꯨꯕ (*suba*) meaning either *to work* or *to hit*

(c)
সিনবা (*sinba*) / শিনবা (*shinba*) → substitution → ꯁꯤꯅꯕ (*sinba*)

শীনবা (*sheenba*) → arrangement → ꯁꯤꯅꯕ (*sinba*)

শীনবা (*sheenba*) → sour taste → ꯁꯤꯅꯕ (*sinba*)

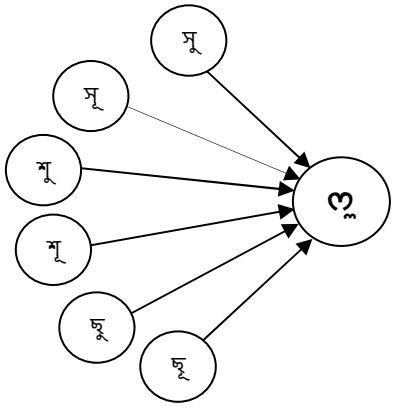

Figure 1. An example of convergence of TU (সু -su, সূ-soo etc.) from Bengali Script to Meitei Mayek

[4] http://www.statmt.org/moses/
[5] http://www.fjoch.com/GIZA++.html

The lexical ambiguity that arises are twofold, i) one after transliteration into Meetei Mayek as given by examples (a) and (b), ii) one before the transliteration as given by the example (c) for which the ambiguity is doubled after the transliteration. Thus, the scripts are functioning as a representation language for lexical ambiguity like the semantic phrase sense disambiguation model for SMT (Carpuat and Wu, 2007).

4.2 Language Modeling

The impact of the different language models is clearly seen in our experiment mostly by way of lexical variation and convergence characteristics as shown in Figure 1. Four different language models are developed: a) language model (LM1) on Bengali script based Manipuri text, b) language model (LM2) on transliterated Manipuri Meetei Mayek text, there is change in the language model parameter such as perplexity which affects the overall translation decoding process, c) language model (LM3) based on language model (LM1) with transliteration to Meitei Mayek on Manipuri text from Bengali Script texts, and d) language model (LM4) based on language model (LM2) with transliteration to Bengali script on Manipuri text from Meetei Mayek text. SRILM (Stolcke, 2002) is used to build trigram model with modified Kneser-Ney smoothing (Stanley and Joshua, 1998) and interpolated with lower order estimates which works best for Manipuri language using 2.3 million words of 180,000 Manipuri news sentences. There are variations in the language model parameters while switching the scripts.

The log probability and perplexity of the sentence (considering the first sentence from Table 2) using Bengali script, “ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়খত্নবা থঙহনখ্রে ।” are given as:

logprob= -22.873 ppl= 193.774 ppl1= 347.888

while the parameters for the same sentence using Meetei Mayek, i.e.,

“ꯏꯂꯦꯛꯁꯟ ꯗꯤꯄꯥꯔꯇꯃꯦꯟꯇꯀꯤ ꯃꯥꯌꯀꯩꯗꯒꯤ ꯑꯥꯏꯑꯦꯐꯁꯤꯗꯤꯗꯥ ꯗꯔꯀꯥꯔ ꯂꯩꯕꯥ ꯊꯕꯛ ꯄꯥꯌꯈꯠꯅꯕꯥ ꯊꯡꯍꯟꯈ꯭ꯔꯦ ꯫” are given as:

logprob= -26.7555 ppl= 473.752 ppl1= 939.364

It is also observed that some of the n-grams entries on one language model are not available in the other language model. For example,

-2.708879 মদুদা চেন্নবনি -0.3211589

is a bigram found in Bengali script based language model but not found in the Meetei Mayek based language model. Similarly,

-6.077539 [illegible] -0.06379553

is a bigram found in the Meetei Mayek based language model but not available in Bengali script based language model. Above all, for the same n-gram in the language models, the log(P(W)) and log(backoff-weight) are found to be different. Two bigram examples are given below:

-1.972813 মদুদা থোক্ককপা -0.09325081
-6.077539 ꯃꯗꯨꯗꯥ ꯊꯣꯛꯀꯄꯥ -0.06379553

and,

-1.759148 মদুদা থোরকপা -0.3929711
-6.077539 ꯃꯗꯨꯗꯥ ꯊꯣꯔꯀꯄꯥ -0.06379552

4.3 Evaluation

The systems are developed using the following corpus statistics.

	# of Sentences	# of Words
Training	10000	231254
Development	5000	121201
Testing	500	12204

Table 3. Corpus Statistics

The evaluations of SMT systems are done using automatic scoring and subjective evaluation.

4.4 Automatic Scoring

We carry out the comparisons of automatic evaluation metrics scores for the SMT systems. The various models developed are evaluated using BLEU (Papineni et al, 2002) and NIST (Doddington, 2002) automatic scoring techniques. A high NIST score means a better translation by measuring the precision of n-gram.

	BLEU Score	NIST Score
Meetei Mayek based Baseline using LM2 language model	11.05	3.57
Meetei Mayek based Baseline with LM3 language model	11.81	3.33
Bengali Script based Baseline using LM1 language model	15.02	4.01
Bengali Script based Baseline using LM4 language model	14.51	3.82

Table 4 . Automatics Scores of English to Manipuri SMT system

BLEU metric gives the precision of n-gram with respect to the reference translation but with a brevity penalty.

	BLEU Score	NIST Score
Bengali Script based Baseline	12.12	4.27
Meetei Mayek based Baseline using	13.74	4.31

Table 5. Automatics Scores of Manipuri to English SMT system

4.5 Subjective Evaluation

The subjective evaluation is carried out by two bilingual judges. The inter-annotator agreement is 0.3 of scale 1. The adequacy and fluency used in the subjective evaluation scales are given by the Table 6 and Table 7.

Level	Interpretation
4	Full meaning is conveyed
3	Most of the meaning is conveyed
2	Poor meaning is conveyed
1	No meaning is conveyed

Table 6. Adequacy Scale

Level	Interpretation
4	Flawless with no grammatical error
3	Good output with minor errors
2	Disfluent ungrammatical with correct phrase
1	Incomprehensible

Table 7. Fluency Scale

The scores of adequacy and fluency on 100 test sentences based on the length are given at Table 8 and Table 9 based on the adequacy and fluency scales give by Table 6 and Table 7.

	Sentence length	Fluency	Adequacy
Baseline using Bengali Script	<=15 words	3.13	3.16
	>15 words	2.21	2.47
Baseline using Meetei Mayek	<=15 words	3.58	3.47
	>15 words	2.47	2.63

Table 8. Scores of Adequacy and Fluency of English to Manipuri SMT system

	Sentence length	Fluency	Adequacy
Baseline using Bengali Script	<=15 words	2.39	2.42
	>15 words	2.01	2.14
Baseline using Meetei Mayek	<=15 words	2.61	2.65
	>15 words	2.10	1.94

Table 9. Scores of Adequacy and Fluency of Manipuri to English SMT system

5 Sample Translation Outputs

The following tables show the various translation outputs of English-Manipuri as well as Manipuri-English PBSMT systems using Bengali script and Meetei Mayek scripts.

English	On the part of the election department, IFCD have been intimidated for taking up necessary measures.
Manipuri Reference Translation (Bengali Script)	ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা খবক পায়খল্লবা খঙহনখ্রে .
Gloss	*election departmentki maykeidagee IFCDda darkar leiba thabak paykhatnaba khanghankhre* .
Baseline Translation output (Bengali Script)	ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা খবক পায়খল্লবা খঙহনখ্রে.

Table 10. English to Manipuri SMT system output using Bengali Script

English	On the part of the election department, IFCD have been intimidated for taking up necessary measures.
Manipuri reference Translation (Meetei Mayek)	ꯏꯂꯦꯛꯁꯟ ꯗꯤꯄꯥꯔꯇꯃꯦꯟꯇꯀꯤ ꯃꯥꯏꯀꯩꯗꯒꯤ ꯑꯥꯏꯑꯦꯐꯁꯤꯗꯤꯗꯥ ꯗꯔꯀꯥꯔ ꯂꯩꯕ ꯊꯕꯛ ꯄꯥꯏꯈꯠꯅꯕ ꯈꯪꯍꯟꯈ꯭ꯔꯦ ॥
Gloss	*election departmentki maykeidagee IFCDda darkar leiba thabak paykhatnaba khanghankhre .*
Baseline Translation output	ꯏꯂꯦꯛꯁꯟ ꯗꯤꯄꯥꯔꯇꯃꯦꯟꯇꯅ ꯑꯥꯏꯑꯦꯐꯁꯤꯗꯤꯗꯥ ꯗꯔꯀꯥꯔ ꯂꯩꯕ ꯊꯕꯛ ꯄꯥꯏꯈꯠꯅꯕ ꯍꯥꯏꯈ꯭ꯔꯦ ॥

Table 11: English to Manipuri SMT system output using Meetei Mayek

Input Manipuri sentence	ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়থত্নবা খঙহনখ্রে .
Gloss	*election departmentki maykeidagee IFCDda darkar leiba thabak paykhatnaba khanghankhre .*
Reference English translation	On the part of the election department, IFCD have been intimidated for taking up necessary measures.
Baseline Translation output	the election department notified IFCD to take up necessary steps

Table 12: Manipuri to English translation output using Bengali script:

Input Manipuri sentence	ꯏꯂꯦꯛꯁꯟ ꯗꯤꯄꯥꯔꯇꯃꯦꯟꯇꯀꯤ ꯃꯥꯏꯀꯩꯗꯒꯤ ꯑꯥꯏꯑꯦꯐꯁꯤꯗꯤꯗꯥ ꯗꯔꯀꯥꯔ ꯂꯩꯕ ꯊꯕꯛ ꯄꯥꯏꯈꯠꯅꯕ ꯈꯪꯍꯟꯈ꯭ꯔꯦ ॥
Gloss	*election departmentki maykeidagee IFCDda darkar leiba thabak paykhatnaba khanghankhre .*
Reference English translation	On the part of the election department, IFCD have been intimidated for taking up necessary measures.
Baseline Translation output	the election department intimidated IFCD to take up necessary steps

Table 13: Manipuri to English translation output using Meetei Mayek:

The English to Manipuri SMT system output using Bengali Script suffers fluency and adequacy scores as given by table 8 compared to English to Manipuri SMT system output using Meetei Mayek script. In the case of Manipuri to English SMT system, the Meetei Mayek based SMT system outperforms the Bengali script based SMT in terms of both fluency and adequacy as given by table 9 as well as automatic scores as given by table 5.

6 Conclusion and Discussion

The detailed study of grapheme-to-phoneme indicates missing tone for several words using present Meetei Mayek script. The training process based on the Bengali script training data is found to have higher vocabulary coverage all across since the representation is done with a finer glyph as compared to Meetei Mayek so is the higher automatic scores in case of English-to-Manipuri PBSMT system. But, considering the subjective evaluation scores, the Meetei Mayek based SMT systems outperforms the Bengali script based English-to-Manipuri SMT systems as against the automatic scores. In the case of Manipuri-to-English PBSMT systems, both the automatic score and subjective evaluation scores using Meetei Mayek outperforms the Bengali script based systems. Statistical significant test is performed to judge if a change in score that comes from a change in the system with script switching reflects a change in overall translation quality. The systems show statistically significant result as measured by bootstrap re-sampling method (Koehn, 2004) on BLEU score. In future, the experiments can be repeated with special treatment of individual morphemes in bits and pieces on a decent size of parallel corpora. More notably, SMT decoders may have the feature of handling two or more scripts of the same language in the future SMT systems for languages like Manipuri.

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

Andreas Stolcke. 2002. *SRILM – an extensible language modeling toolkit*. In Proceedings of the International Conference on Spoken Language Processing.

Franz J. Och. 2003. *Minimum error rate training in Statistical Machine Translation*, Proceedings of ACL.

George Doddington. 2002. *Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics*. In Proceedings of HLT 2002, San Diego, CA.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. In Proceedings of 40th ACL, Philadelphia, PA.

Kristina Toutanova, Hisami Suzuki and Achim Ruopp. 2008. *Applying Morphology Generation Models to Machine Translation*, In Proc. 46th Annual Meeting of the Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. *How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine*, 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007). pages 43-52, Skövde, Sweden, September 2007.

Ondřej Bojar and Jan Hajič. 2008. *Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation*, Proceedings of the Third Workshop on Statistical Machine Translation, pages 143–146, Columbus, Ohio, USA.

Ondřej Bojar, Bushra Jawaid and Amir Kamran. 2012. *Probes in a Taxonomy of Factored Phrase-Based Models*, Proceedings of the 7th Workshop on Statistical Machine Translation of Association for Computational Linguistics, pages 253–260, Montréal, Canada.

Philipp Koehn. 2004. *Statistical significance tests for machine translation evaluation*. In EMNLP-2004: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 25-26 July 2004, pages 388-395, Barcelona, Spain.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Reyyan Yeniterzi and Kemal Oflazer. 2010. *Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish*, In proceeding of the 48th Annual Meeting of the Association of Computational Linguistics, Pages 454-464, Uppsala, Sweden.

Stanley F. Chen and Joshua Goodman. 1998. *An empirical study of smoothing techniques for language modeling*. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010a. *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.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. *Manipuri-English Bidirectional Statistical MachineTranslation Systems using Morphology and Dependency Relations*, Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 83–91, COLING 2010, Beijing, August 2010.

Thoudam Doren Singh. 2012. *Bidirectional Bengali Script and Meetei Mayek Transliteration of Web Based Manipuri News Corpus*, In the Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP) of COLING 2012, IIT Bombay, Mumbai, India, pages 181-189, 8th December, 2012.