Master Thesis

English to Bangla Phrase-Based Statistical Machine Translation

Md. Zahurul Islam
2519786

Supervisors:

Dr. Andreas Eisele
PD. Dr. Valia Kordoni
Prof. Hans Uszkoreit
Department of Computational linguistics
Saarland University
Germany

&

Dr. Jörg Tiedemann
Department of Humanities Computing
University of Groningen
The Netherlands

Submitted to

Department of Computational linguistics
at
Saarland University

Erasmus Mundus European Masters Program in
Language and Communication Technologies (LCT)

August 2009

university of groningen

faculty of arts

Master Thesis

English to Bangla Phrase-Based Statistical Machine Translation

Md. Zahurul Islam
1788213

Supervisors:

Dr. Jörg Tiedemann
Department of Humanities Computing
University of Groningen
The Netherlands

&

Dr. Andreas Eisele
DFKI GmbH, Language Technology Lab and
Department of Computational linguistics
Saarland University
Germany

Submitted to

Faculty of Arts
at
University of Groningen

Research Master Linguistics
Erasmus Mundus European Masters Program in
Language and Communication Technologies (LCT)

August 2009

Acknowledgements

I would like to express my gratitude to all the people who supported and accompanied me during the progress of this thesis. Special thanks to my family, specially my mom and dad, my sisters, my brother and brother in laws, for being part of every bit of my life. Their endless motivation, constant mental support and unconditional love have been influential in whatever I have achieved so far.

Secondly, I would like to thank my supervisors Dr. Jörg Tiedemann and Dr. Andreas Eisele. I would always be grateful for their insightful and guidance they provide me in order to cross the final hurdle. Specially, Dr. Tiedemann who listened to all my problems I faced during this thesis and showed me the way to overcome them.

I would also like to thanks Prof. Gisela Redeker and PD. Dr. Valia Kordoni for their guidance in the LCT program. They helped me a lot during my stay in Saarbrücken and Groningen. Whenever I faced any problem in Groningen, Prof. Redeker helped me to solve the problems.

I would also like to thanks all my friends in LCT program in Groningen and Saarbrücken specially, Pranesh and Mehwish for their mental support. And also Stefan and Leila from Martini House, Max and Alejandra from Saarbrücken for their support during my stay in Groningen and Saarbrücken.

I would also like to thank Prof. Mumit khan and all my colleagues from the Center for Research on Bangla Language Processing (CRBLP), BRAC University, Bangladesh for their encouragement to join LCT program and their help with linguistic resources.

Finally, all my friends in Bangladesh and here in the Europe deserve special thanks. They are the ones who would always there to spend time with, and share my joys and sorrows. In the end, I would like to thank almighty God for giving me the strength to achieve whatever I have achieved so far.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Zahurul Islam)

Abstract

Machine Translation (MT) is the task of automatically translating a text from one language to into another. In this work, we describe the phrase-based Statistical Machine Translation (SMT) system that translates English sentences to Bangla sentences. Though SMT systems are trained using large parallel corpora, they can never hope to have a complete coverage of unrestricted text. Particular problems arise for highly productive word classes like proper nouns. We have added a transliteration module to the translation system to transliterate out-of-vocabulary words.

Prepositional systems across languages vary to a considerable degree, and this cross-linguistic diversity increases as we move from core, physical senses of preposition into the metaphoric extensions of prepositional meaning [Naskar and Bandyopadhyay, 2006a]. Where English uses prepositions, Bangla typically uses post positions and in some cases attaches inflections to the head nouns. A preposition-handling module is added to the translation system to handle English prepositions during translation.

We have shown the improvement of our approach through the effective impact on the BLEU, NIST and TER scores. The BLEU score of our system is 23.3 for short sentences and 11.7 overall.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine Translation (MT) is the task of automatically translating text in a source language to text in a target language. MT was envisioned as a computer application back in the 1950's. After more than 50 years of research, MT is still an open problem.

Nowadays, the demand for MT is quickly growing. Multilinguality is considered to be a part of democracy. In the European Union (EU), documents are being translated to 23 official languages. There is an ongoing project called *EuroMatrixPlus* [1] to a build machine translation system for all European language pairs. The United Nations (UN) is also translating a large number of documents into several languages. UN corpora for some language pairs like Chinese – English, Arabic – English are among the largest bilingual corpora distributed via the Linguistic Data Consortium (LDC). On the World Wide Web (www), many web pages are available in their national languages. MT systems can be used to make the content of those websites accessible for understanding the content of those website for people who do not understand those languages. MT can help to reduce the language barrier and make communication easier.

There are two different approaches to address MT problems. One is the rule-based approach and another is the data-driven approach. In the rule-based approach, the source language text is analyzed using various tools like: a morphological analyzer and a parser, and transformed to an intermediate representation. Some rules are used to generate target language text from this intermediate representation. A larger number of rules is required to capture the natural language phenomena. These rules transfer the grammatical structure of the source language into the target language. As the rules grow, the system becomes very complicated. Formulating a large number of rules is a time consuming process.

In the data-driven approach, large parallel and large monolingual corpora are used as the source of knowledge. This approach can be further divided into statistical approaches and example-based approach. In statistical approaches, target text is generated on the basis of a statistical model and parameters are derived from corpora. Here, MT is also treated as a decision problem, a best target language sentence is decided from a given source language sentence. Bayes rule and statistical decision theory are used to solve this decision problem. Sta-

[1] Euromatrix Plus website: http://www.euromatrixplus.net/

tistical decision theory and Bayesian decision rules are used to minimize decision errors. Statistical Machine Translation (SMT) gives better results as more and more training data is available. In the example-based approach, the basic idea is translation by analogy. An example-based Machine Translation (EBMT) is given a set of source language sentences and their corresponding translations in the target language, and uses those examples as source of knowledge to translate others, similar source language sentences into the target language. The basic hypothesis is that, if a previously translated sentence occurs again, then the translation is likely to be correct. EBMT uses the case-based[2] reasoning approach of machine learning.

SMT requires enormous amounts of parallel text in the source and target language to achieve high quality translation. However, many languages are considered to be low-density languages, either because the population speaking the language is not very large, or because insufficient digitized text material is available in a language even though it spoken by millions of people. Bangla/Bengali is one such language. Bangla, an Indo-Aryan language, is a language of Southeast-Asia, which comprises present day Bangladesh and the Indian state of West Bengal. With nearly 230 million speakers, Bangla is one of the most spoken languages in the world.

In this thesis, our aim is to present a phrase-based SMT system for translating English to Bangla. The current state-of-the-art phrase-based SMT system available for this task is based on a log-linear translation model, which is used as our baseline system. We have incorporated a transliteration module as a component with our baseline to handle proper names/ (out of vocabulary words (OOV)). The transliteration module is same as phrase-based SMT, but it works on character level instead of phrase level. Instead of prepositions in Egnish, Bangla uses postpositions and in some cases attaches inflections to the head nouns. A preposition-handling module is added to the translation system to handle English prepositions during translation. We evaluate the components of our SMT system through their effective impact on BLEU score, NIST score and Translation Error Rate (TER).

[2] Case-based reasoning is the process of solving a new problem based on the solutions of similar past problems. More can be found here: http://en.wikipedia.org/wiki/Case-based_reasoning

Chapter 2

Related Work

Although being among the top ten most widely spoken languages in the world, the Bangla language still lacks significant research in the area of natural language processing, specifically MT. Some research work has been done in Bangla MT in the context of rule-based MT. [Dasgupta et al., 2004] proposed a way of Machine Translation from English to Bengali. Their proposed architecture uses the syntactic transfer of English to Bangla with optimal time complexity. This architecture follows five steps: a) Tagging, b) Parsing, 3) Change CNF[1] parse tree to normal Parse tree, 4) Transfer of English parse tree to Bangla parse tree, 5) Generate output translation with morphological analysis. The Cocke-Younger-Kasami (CYK) parsing algorithm is used to parse English sentences. The CYK parsing algorithm outputs parse tree based on CNF form of English grammar. These CNF parse trees are transferred to normal parse trees using some transformation rules. They came up with two types of nodes in a CNF parse tree. Node-type1, whose childern will remain as their childern after transforming to normal parse tree, rest of the nodes are Node-type2. A transformation rule $S- > NP+VP$ will be applied where NP is Node-type1 and VP is Node-type2. 2.1 shows a tree in CNF from and corresponding transformation which is taken from [Dasgupta et al., 2004]. The transformed normal parse trees are converted to Bangla parse trees using a bilingual dictionary. These syntactic transfers depend on mapping between surface structures of sentences. In the generation stage [Dasgupta et al., 2004] used a dictionary to identify subject, object and also other information like person, number and generate target sentences.

[Naskar and Bandyopadhyay, 2006b] presents an example-based machine translation system. This work identifies the phrases in the input through a shallow analysis, retrieves the target phrases using a phrasal example based and finally combines the target language phrases by employing some heuristics based on the phrase reordering rules in Bangla. [Naskar and Bandyopadhyay, 2006b] have discussed some syntactic issues between English and Bangla. The NP structure differs in English and Bangla, In English, [specifier/article] [adv] [noun] [plural marker] [case marker], and in Bangla: [specifier] [adv] [adj] [noun] [plural marker] [case marker]. There are some similarities between English and Bangla pronouns as well, but these differ on gender: Bangla pronouns do not

[1] CNF: Chomsky Normal Form

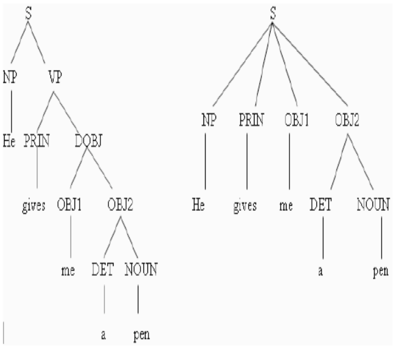

Figure 2.1: Transformation of CNF parse tree to normal parse tree

depend on gender information and for second and third persons have more than one forms. So, translating pronouns from English to Bangla involves anaphora resolution [2] In Bangla adjective forms (positive, comparative, superlative) are handled in the similar way as in English. [Naskar and Bandyopadhyay, 2006b] used a shallow parser to identify the phrases in the source language and tags with the phrase with relevant information, translated these phrases individually and arranged them using some phrase ordering heuristics rules.

[Saha and Bandyopadhyay, 2005] proposes an English to Bangla EBMT system for translating news headlines. The translation of source to target headline is done in three steps. In the first step: search in direct example base, if not found then search in generalized tagged example base. If a match is found in the second step, then extract the English equivalent of the Bangla words from the bilingual dictionary and apply some synthesis rule to generate the surface word level. If the second step fails, then the tagged input headline is analyzed to identify the constituent phrases. The target translation is generated from the bilingual example phrase dictionary and uses heuristics to reorder Bangla phrases.

There is an open source machine translation system called Anubadok[3] available for translating English sentences to Bangla sentences. It uses four steps to translate English to Bangla. First, it pre-processes the English documents. In this step Anubadok converts different kind of documents to XML documents. Second, it performs part of speech tagging of source document. In this step it also performs tokenization as a preprocess and lemmatization as a post process. Third, it performs the translation. In the beginning of this step, it determines the sentence type, subject, object, verb and tense, and then translates English words to Bangla words using a bilingual dictionary and considering linguistic properties. Finally, it joins subject, object and verbs in the SOV order.

[Naskar and Bandyopadhyay, 2006a] shown a technique of handling prepo-

[2] Bangla pronouns, unlike their English counterpart, do not differentiate for gender; the same pronoun may be used for *he* or *she*. However, Bangla pronouns encode proximity. Different pronouns are used for someone who is nearby, and for someone who is further away. In addition, each of the *second* and *third person* pronouns have different forms for the familiar and polite forms; the *second person* also has a "very familiar (dishonor)" form. For example: *you* in the sentence "Will you please give me that pen?" will be translated to তুমি (familiar form) or আপনি (polite form) or তুই (dishonor) depending on the context

[3] Anubadok can be downloaded from here: http://anubadok.sourceforge.net/

sitions in English to Bangla machine translation system. In Bangla there is no concept of preposition. English prepositions are translated to Bangla by attaching inflections to the head nouns of the prepositional phrase. The English form of preposition (preposition) (reference object) is translated to (reference object) [(inflection)] [(prepositional − word)]. The reference object plays a major role in determining the correct preposition sense. For example: *at home* should translated to the stem word বাড়ি (bari : home) and the inflection -তে(-te).

[Vilar et al., 2007] has presented an approach that treats source and target sentences as strings of letters instead a collection of words. They have treated each word as a sequence of letters, which is translated into a new sequence of letters. This approach reduces the vocabulary size significantly but it increased the average sentence length. This system could be useful for closely related languages and languages where very little parallel training data is available.

Transliteration systems are being used nowadays in MT systems. [UzZaman et al., 2006] presents a phonetics based transliteration system for English to Bangla which produces intermediate code strings that facilitate matching pronunciations of input and desired output. They have used table-driven direct mapping between English alphabet to Bangla alphabet and a phonetic lexicon – enabled mapping.

The first piece of work on Statistical Machine Transliteration was [Knight and Graehl, 1997]. They have done back transliteration of Japanese to English. The problem was decomposed to five sub-problems and recombine them using Bayes' rule. The decomposed sub-problems are:

1. An English Phrase is written

2. A translator pronounces it in English

3. The pronunciation is modified to fit the Japanese phonetic system

4. The sounds are converted into Japanese katakana, the syllabary that is used to write transliteration of foreign words.

5. The Japanese phrases are written

These subproblems rely on probabilities and Bayes' theorem. The probabilities are:

1. $Pr\,(w)$ - probability that used to generate written English word sequences

2. $Pr\,(e|w)$ - probability that used to pronounce English word sequences

3. $Pr\,(j|e)$ - probability that used to convert English sounds into Japanese sounds

4. $Pr\,(k|j)$ - probability that used to convert Japanese sounds into Japanese katakana

5. $Pr\,(o|k)$ -probability that used to introduce misspellings caused by Optical Character Recognition (OCR)

Given a katakana string o observed by OCR system, the system finds the English word sequence w that maximizes the sum over all e, j and k, of:

$$Pr\left(w\right).Pr\left(e|w\right).Pr\left(j|e\right).Pr\left(k|j\right).Pr\left(o|k\right) \tag{2.1}$$

The generation of English word sequences $Pr\left(w\right)$ is implemented as a weighted finite-state acceptor (WFSA). [Knight and Graehl, 1997] implemented the other distributions as weighted finite-state transducers (WFST). Both WFSA and WFST are trained on an appropriate corpus. To compute the most probable English transliteration they used two shortest path algorithms.

[Matthews, 2007] presents a machine transliteration system of proper names using MOSES, a phrase based Machine Translation system. [Matthews, 2007] have experimented to translate proper names between both English – Chinese and Arabic to English. [Matthews, 2007] have achieved 43.0% accuracy of forward transliteration from Arabic to English and 37.8% from English to Chinese.

[Finch and Sumita, 2008] have used have an English <-> Japanese machine transliteration system which is a part of a machine translation system. This system is a character -based machine translation system for Out of Vocabulary words in MT system. They have collected training data from the freely available Katakana to English dictionary cross lingual links for Wikipedia.

Chapter 3

Background

3.1 Statistical Machine Translation

The history of MT started in the 1950s. Initially, the MT results were seen as promising but later it turned out to be a lot more complicated than expected. In 1966, the ALPAC report found that the ten years long research had failed to fulfill the expectations, and thus the funding was dramatically reduced. In the late 1980s the IBM group started to work on Statistical MT due to increased computing power and availability of parallel corpora [Brown et al., 1993]. Currently there are many MT systems available on the web for high density language pairs like: English-French, English-German, etc..

3.1.1 Bayes Rule for Machine Translation

In Statistical MT, there is a source language sentence $e_1^I = e_1 \dots e_i \dots e_I$, which is to be translated into a sentence of target language $f_1^J = f_1 \dots f_j \dots f_J$. Using statistical decision rules we can get all the possible translations of the source language. Finally, we have to choose the sentence, with minimized errors. The Bayes decision rule is as shown below.

$$\hat{f}_1^{\hat{j}} = argmin_{J,f_1^J} \left\{ \sum_{J' f_1'^{j'}} Pr\left(f_1'^{J'} | e_1^I\right) .L\left(f_1^J, f_1'^{J'}\right) \right\} \tag{3.1}$$

Here, $L\left(f_1^J, f_1'^{J'}\right)$ is the error function under consideration. It gives us the error of the candidate translation f_1^J by assuming that the correct translation is $f_1'^{J'}$. The Bayes decision rule explicitly depends on the error function. For minimizing the sentence or string error rate, the error function is:

$$L\left(f_1^J, f_1'^{J'}\right) = \begin{cases} 0 & \text{if } f_1^J = f_1'^{J'} \\ 1 & \text{otherwise} \end{cases} \tag{3.2}$$

$$= 1 - \delta\left(f_1^J, f_1'^{J'}\right) \tag{3.3}$$

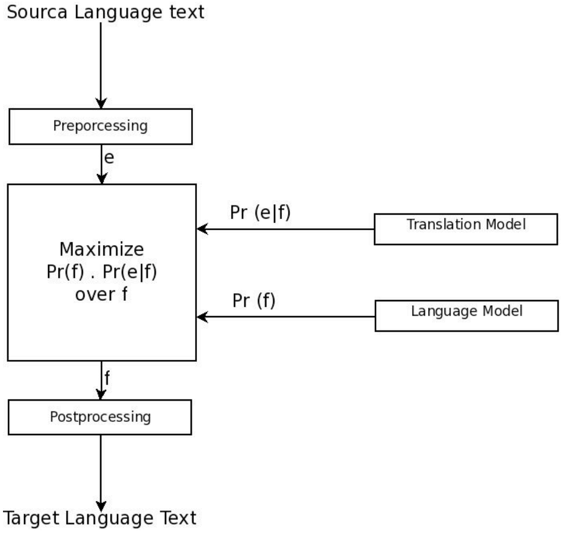

Figure 3.1: Architecture of source–channel model based translation approach

The function $\delta\left(f_1^J, f_1'^{J'}\right)$ in rule 3.3 denotes the *Kronecker delta* [1]. This function is called 0-1 error function because it assigns an error of zero to the correct solution and an error of 1 otherwise. Using the error function the rule 3.1 can be simplified to:

$$\hat{f}_1^{\hat{J}} = argmax\left\{Pr\left(f_1^J|e_1^I\right)\right\} \tag{3.4}$$

This decision rule is called maximum a-posteriori (MAP) decision rule. It gives us the posterior probability distribution over all the sentences of target language f_1^J given a source sentence e_1^I. So, we use this hypothesis, which maximizes the posterior probability $Pr\left(f_1^J|e_1^I\right)$. Figure 3.1 shows the architecture of the MT approach using Bays rule. The posterior probability can be decomposed to:

$$Pr\left(f_1^J|e_1^I\right) = \frac{Pr\left(f_1^J\right)Pr\left(e_1^I|f_1^J\right)}{Pr\left(e_1^I\right)} \tag{3.5}$$

Note that the denominator $Pr\left(e_1^I\right)$ depends only on the source sentence e_1^I. In the case of the MAP rule, we can omit the denominator during search.

$$argmax\left\{Pr\left(f_1^J|e_1^I\right)\right\} = argmax\left\{Pr\left(f_1^J\right)Pr\left(e_1^I|f_1^J\right)\right\} \tag{3.6}$$

This is the fundamental equation of statistical machine translation. It allows an independent modeling of the target language model $Pr\left(f_1^J\right)$ and the translation model $Pr\left(e_1^I|f_1^J\right)$. The translation model $Pr\left(e_1^I|f_1^J\right)$ links between source language sentence and target language sentence. The language model $Pr\left(f_1^J\right)$ describes how well formed the target language is.

[1] Kronecker delta is a function of two variables, usually integer, which is 1 if they are equal and 0 otherwise

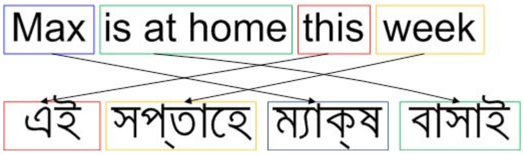

Figure 3.2: Phrasal reordering necessary when generating Bangla from English

3.1.2 Phrase Based Translation Model

A translation model generates target language sentence f_1^J from source language sentence e_1^I by assigning a probability to source language sentence and target language sentence. Statistical MT computes these probabilities by considering the behavior of the phrases. Fig 3.2 shows an example where entire phrases often need to be translated and reordered as a unit. It uses phrases as well as single words as the fundamental units of translation. A phrase-based translation has three steps [Koehn, 1993]. First, it groups the English source words into phrases $e_1, e_2 \ldots e_I$, Next, it translates each English phrase e_1^I to Bangla phrase f_1^J. Finally, it reorders each Bangla phrase using the language model. The probability model of phrase-based translation depends on translation probability and distortion probability. The factor $\phi\left(f_1^J|e_1^I\right)$ is the translation probability to generate the Bangla phrase f_1^J from the English phrase e_1^I. The distortion probability d is used to reorder the Bangla phrases. Distortion in statistical English to Bangla machine translation refers to a word having a different position in the Bangla sentence than it had in the English sentence. It is measured by the distance between the positions of a phrase in the two languages. The distortion is parameterized by $d\left(a_j - b_{j-1}\right)$, where a_j is the start position of the Bangla phrase generated by the i^{th} English phrase e_i and b_{j-1} is the end position of the Bangla phrase generated by the $i - 1^{th}$ English phrase e_{i-1}. These can be simplified to $d\left(a_j - b_{j-1}\right) = \alpha^{|a_j - b_{j-1} - 1|}$. This distortion model penalizes large distortions by giving lower probability the larger the distortion. The rule 3.4 showed the MAP decision rule for SMT. If we extend the decision rule for the translation model of phrase based MT, then it becomes:

$$Pr\left(f_1^J|e_1^I\right) = \prod_{i,j=1}^{I,J} \phi\left(f_j, e_i\right) d\left(a_j - b_{j-1}\right) \tag{3.7}$$

These parameters and the distortion constant α could be set from bilingual text[2], in which each Bangla sentence was paired with an English sentence. From the bilingual text we know exactly which Bangla sentence is the translation of which English phrase. This is called phrase alignment. We can extract these aligned phrases from another kind of alignment called word alignment. Word alignment is different from phrase alignment because it shows exactly which Bangla words exactly align with which English words inside each phrase.

[2] Text in two different languages, one text is the translation of another text in a different language

3.1.3 Word Alignment

A word alignment describes the mapping between the source words and the target words in a set of parallel sentences. Nowdays, word aligned bilingual corpora are being used as an important source of knowledge. Word alignment models were first introduced in statistical machine translation [Brown et al., 1993]. We are given source language (English) sentence $e = e_1^I$ which will be translated to a target language (Bangla) $f = f_1^J$. According to the source-channel approach, we have to choose the sentence with the highest probability among all the possible target language (Bangla) sentences

$$\hat{f} = argmax_f \left\{ Pr\left(f|e\right) \right\} \tag{3.8}$$

$$\hat{f} = argmax_f \left\{ Pr\left(e|f\right).Pr\left(f\right) \right\} \tag{3.9}$$

This decomposition allows us to use two independent knowledge sources, one is the translation model $Pr\left(e|f\right)$ and the other one is the language model $Pr\left(f\right)$. Now, the word alignment is introduced as a hidden variable in the translation model.

$$Pr\left(e|f\right) = \sum_a Pr\left(e, a|f\right) \tag{3.10}$$

There is a restriction in this model, which is that each source word is aligned with at most one target word. So, each alignment a is a mapping from source sentence positions to target sentence positions $a = a_1^I = a_1 \ldots a_i \ldots a_I, a_i$ subset symbol $\{0, \ldots J\}$. The alignment may contain an empty alignment where $a_i = 0$ which means that there is a word in the source language sentence which does not have an alignment in the target language sentence. A detailed description of the translation models IBM-1 to IBM-5 can be found in [Brown et al., 1993]. There is another word alignment technique called HMM alignment [Vogel, 2003]. It uses a first order model $p\left(a_i|a_{i-1}, J\right)$, where alignment position a_i depends on previous alignment position a_{j-1}. The distortion (distance) of the position is modeled as $p\left(\left(|a_i - a_{i-1}|\right)|J\right)$.

3.1.4 Reordering Models

Reordering models consider dependencies across phrase boundaries. These models are useful to choose a good reordering. This model can be decomposed into two parts. The distortion penalty model is based on the distance. It assigns costs by considering distance from the end position of a phrase to the start position of the next phrase. There is a very simple distance based distortion penalty model described in [Och and Ney, 2004]. The distortion penalty model assigns zero cost to monotonic translation at the phrase level. The distortion penalty goes high as more phrases are reordered.

The other is the n-gram language model. The language model is used to ensure the well-formedness of the target language sentence. An n-gram based language model considers only local context, it turns out to be quite powerful and hard to improve upon. We only need monolingual training data to train a language model. So, it can be trained on significantly larger volumes of training data. There are many freely available libraries to build a language model.

3.1.5 Decoding

Decoding is the process of finding a target translation sentence (Bangla) for a source sentence (English) using translation model and language model.

$$\hat{F} = argmax_{f \in Bangla} \left\{ Pr\left(e_1^I | f_1^J\right) .Pr\left(f_1^J\right) \right\} \tag{3.11}$$

Decoding is a search problem which maximizes the translation and language model probability. MT decoders use best-first search based on heuristics. Generally, a best-first search algorithm explores a node n based on an evaluation function $f(n)$. The variant of best-first search called A^* was first used for machine translation by IBM [Brown et al., 1995]. [Koehn, 2004] describes the phrase-based decoding for Pharaoh MT decoder. It limits the search space during decoding by only searching Bangla sentences f_1^J which are the possible translation of English sentences e_1^I.

The search process starts with the null hypothesis as initial search position/state. The hypothesis is expanded by choosing each possible English word or phrase that could generate a Bangla sentence initial phrase. Each position/state is associated with costs: current cost and future cost. The current cost is the total probability of the phrases that have been translated already in the hypothesis. That means, the current cost is the product of translation model probability, distortion and language model probability. For the set of partially translated phrases $S = \left(e_1^I, f_1^J\right)$, the current cost is:

$$Cost\left(e_1^I, f_1^J\right) = \prod_s \phi\left(e_i, f_j\right) d\left(a_j - b_{j-1}\right) Pr\left(f_1^J\right) \tag{3.12}$$

The future cost is the estimation of the cost of translating the remaining words in the English sentences. By combining these costs each position/state gives the estimation of the probability of search path for the complete translation sentence f_1^J. Nowdays decoders use beam search rather than best-first search or A^* search. At each ply of the search there is a stack. The stack fits with only n entries. At each ply of the search all the states are expanded and pushed onto the stack. They are ordered by cost, the best n entries are kept and the rest are deleted.

3.1.6 Minimum Error Rate Training

Modern SMT systems use some variants of the exponential log linear model. This model combines diverse knowledge sources to score the output translation. This model chooses the translation using the *argmax* decision rule with highest probability. Choosing the parameters of this exponential model or scaling factors for each knowledge source in the noisy channel model significantly affects translation quality, since they are used to drive the search space of possible target language translations. So, a method will be useful which explores the parameter space of scaling factors and picks a value that maximizes translation quality according to an automatic evaluation metric. Minimum Error Rate training was introduced in [Och, 2003] using the n-best list as an approximation to the translation search space, and considers rescoring the translations with different choices of scaling parameters. This explicitly generates an error surface whose minimum can be inspected, and the corresponding parameter choices reported. [Och, 2003] shows that the space of parameter configurations can be limited to those that actually cause the error surface to change, making this search strategy quite feasible when used in a greedy search through the parameter space.

3.2 Machine Transliteration

Transliteration is a process of transforming phonemes or graphemes of the source language to phonemes or graphemes of the target language. Transliteration can be defined as the task of transcribing the words in the source script to words in the target script. Transcribing without a bilingual lexicon is a challenging task as the output words produced in the target script should be such that it is acceptable to the readers. Transliteration systems find wide applications in MT systems and Cross Lingual Information Retrieval Systems (CLIR). For MT or CLIR difficulties arise due to huge numbers of OOV[3] words which are continuously added into the languages. These OOV words are proper names, technical words and foreign words. There are two types of transliteration: forward transliteration and backward transliteration. The forward transliteration is the process of transforming words in source language to words in target language. The reverse process is backward transliteration, transforming target language approximations back into original source language.

Transliteration between English and languages that can be viewed as having a superset of English alphabet, e.g. most European languages, rarely, if ever transliterate names, for example Guillaume (French) and Jorge (Spanish) are usually left unchanged rather than translated into their Anglicized equivalents William and George[Matthews, 2007].

Transliteration becomes essential when two different languages use different writing scripts, for example: English to Chinese,Japanese, Bangla, Hindi and Arabic. The transliteration process could be a lossy process for languages like English to Chinese, English to Arabic or English to Japanese, due to the lack of direct correspondence between the languages' phonetic systems [Matthews, 2007]. This is not the case for English ↔ Bangla transliteration. Most of the English nouns are pronounced the same in Bangla, but are written using Bangla script. So, a transliteration module could be used for the English to Bangla SMT system. There are some exceptions, where Bangla translations are available for English nouns. Table 3.1 shows the example of English–Bangla Transliteration. Table 3.2 shows some example of nouns which cannot be transliterated. The following sections describe the basics about machine transliteration.

[3] OOV Words: words for which a system has no translation

Table 3.1: Example of English - Bangla Transliteration

English	Bangla
Socrates	সক্রেটিস (soc-ra-tis)
Burkina Faso	বুর্কিনা ফাসো (bur-ki-na fa-so)
Benjamin Harrison	বেঞ্জামিন হ্যারিসন (ben-ja-min ha-rri-son)
Austria	অস্ট্রিয়া (au-s-tri-aa)

Table 3.2: Non transliterated nouns

English	Bangla
India	ভারত (bha-rat)
Mississippi River	মিসিসিপি নদী (mi-ssi-ssi-ppi no-di)
Pacific Ocean	প্রশান্ত মহাসাগর (pro-shan-to moha-shagor)
China	চীন (chin)
Republic of Texas	প্রজাতন্ত্রী টেক্সাস (proja-ton-tri texas)

3.2.1 Statistical Transliteration Model

Assume that given a source language word, represented as a sequence of letters $s = s_1^I = s_1 \ldots s_i \ldots s_I$, needs to be transliterated as a sequence of letters in the target language, represent as $t = t_1^J = t_1 \ldots t_j \ldots t_J$. The job of the transliteration module is to find the best target language letter sequence among the candidate letter sequences, which can be represented as:

$$\hat{t} = argmax_{t_1^J} \left\{ Pr\left(t_1^J | s_1^I\right) \right\} \tag{3.13}$$

This transliteration model is based on noisy channel model. We can reformulate using Bayes' rule:

$$\hat{t} = argmax_{t_1^J} \left\{ Pr\left(s_1^I | t_1^J\right) \right\} .Pr\left(t_1^J\right) \tag{3.14}$$

Formula 3.14 allows a $N - gram$ model $Pr\left(t_1^J\right)$of target language letters and a transcription (letter translation) model $Pr\left(s_1^I | t_1^J\right)$. The best target sequence is obtained based on the product of the probabilities of the transcription model and the probabilities of the language model and their respective weights. Determining the best weights is necessary for obtaining the right target language's letter sequences. Minimum error rate training described in section 3.1.6 is used for determining the best weights.

3.3 Preposition

Prepositional systems across languages vary to a considerable degree, and this cross-linguistic diversity increases as we move from core, physical senses of preposition into the metaphoric extensions of prepositional meaning [Naskar and Bandyopadhyay, 2006a]. The lexical meaning of preposition is important,

Table 3.3: Bangla Postposition Examples

Postposition	**Example**
আগে (before)	সকালের আগে (before the morning)
ঐ-পারে (across)	নদীর ঐ-পারে (across the river)
নিচে (under)	বই এর নিচে (under the book)
পরে (after)	সন্ধ্যার পরে (after the evening)

because it is intended for use in an MT system, where the meaning of a sentence, a phrase or lexical entry of the source language must be preserved in the target language, even though it may take different syntactic form in the source and target language.

There is no concept of preposition in Bangla. Instead of prepositions Bangla typically uses postpositions and some cases attaches inflections to the head noun. The postpositions follow the nouns. The noun is usually in the genitive/accusitive case unless the two words are placed under the rules of সন্ধি (Sandhi)[4] or সমাস (Samas) [5] in which case, the noun is not inflected. Table 3.3 shows Bangla postposition examples. In many cases, English prepositions are translated to Bangla by attaching appropriate inflections to the head noun. For example: inflection -তে (-te) attaches to the noun বাড়ি (home) and become বাড়িতে(at home), inflection -য় (-y) attaches to the noun সন্ধ্যা (the evening) and becomes সন্ধ্যায় (in the evening). The English form of preposition (preposition) (reference object) is translated to Bangla (reference object) [(inflection)] [(postpositional-word)]. Handling Bangla preposition would improve the MT performance.

[4] Sandhi is the euphonic change when words are conjoined. For example: ঢাকা (Dhaka) + ঈশ্বরী(Godess) = ঢাকেশ্বরী (Queen of Dhaka), বধু (bride) + উৎসব (festival) = বধুৎসব (festival to welcome a bride). shandi takes place on the simple joining of words in a sentence, on the formation of compound words and on the adding of affixes to noun or verbs.

[5] Samas is the rules of compounding words. For example: তুমি (you) এবং (and) আমি (I) = আমরা (we).

Chapter 4

Experimental Framework

4.1 Data

This section describes the corpora used in this thesis, format and preprocessing steps. We have used a parallel corpora of south Asian languages called Enabling Minority Language Engineering (EMILLE) corpus developed by Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. This corpus is distributed by the European Language Resources Association. This corpora contains 200, 000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. Bangla translation contains 189,495 words. Table 4.1 shows the EMILLE corpus statistics.

To prepare the data, we need to do some preprocessing. Preprocessing was as follows:

1. Convert encoding from UTF-16 to UTF-8
2. Extract sentences from XML mark up text
3. Align sentences
4. Tokenize English and Bangla corpus and lower case of English

We also used KDE4 system messages as a corpus, English and Bangla translation of BN_BD (Bangla in Bangladesh) and BN_IN (Bangla in West-Bengal/India) domains. This KDE4 system message corpus contains 221,409 words and 33,365 sentence pairs with UTF-8 encoding. We used monolingual corpus from EMILLE project and the Prothom-Alo corpus developed by BRAC University, Bangladesh. The EMMILE monolingual corpus contains 1,867,452

Table 4.1: EMILLE English - Bangla Corpus Statistics

Encoding:	UTF-16
Total number of files :	72 (English) and 70 (Bangla)
Total English sentences :	12,654
Total Bangla sentences :	12,633

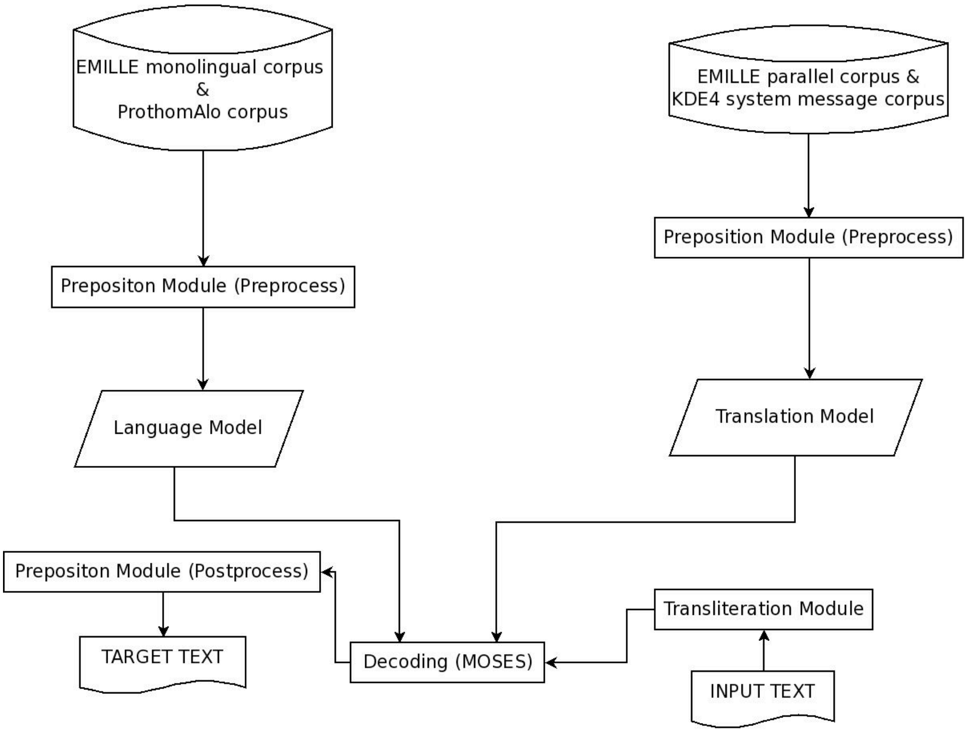
EMILLE monolingual corpus
&
ProthomAlo corpus
EMILLE parallel corpus &
KDE4 system message corpus
Preposition Module (Preprocess)
Prepositon Module (Preprocess)
Language Model
Translation Model
Prepositon Module (Postprocess)
TARGET TEXT
Decoding (MOSES)
Transliteration Module
INPUT TEXT

Figure 4.1: Combined System Architecture

words and the Prothom-Alo corpus contains 19,496,884 words. The Prothom-Alo corpus contains some lines with English words. We have deleted those from lines in the final version text for the language model.

4.2 System Architecture

We integrated two external modules with the baseline system, as we believe these modules will improve the translation quality and accuracy of our MT system. We already described the importance of handling preposition during translation. We added a module that puts the post-positional words before the nouns and separates inflectional suffixes before training the system and later put the postpositional words after the noun adds suffixes with nouns as post–processing task. Our second module is the transliteration module which is responsible for identifying out of Vocabulary words (OOV) and transliterating those words to avoid the presence of English words in the target Bangla translation. Figure 4.1 shows the combined system architecture and Figure 4.2 shows the architecture of the transliteration module.

4.3 Software

The following sections briefly describe the software that was used during the project.

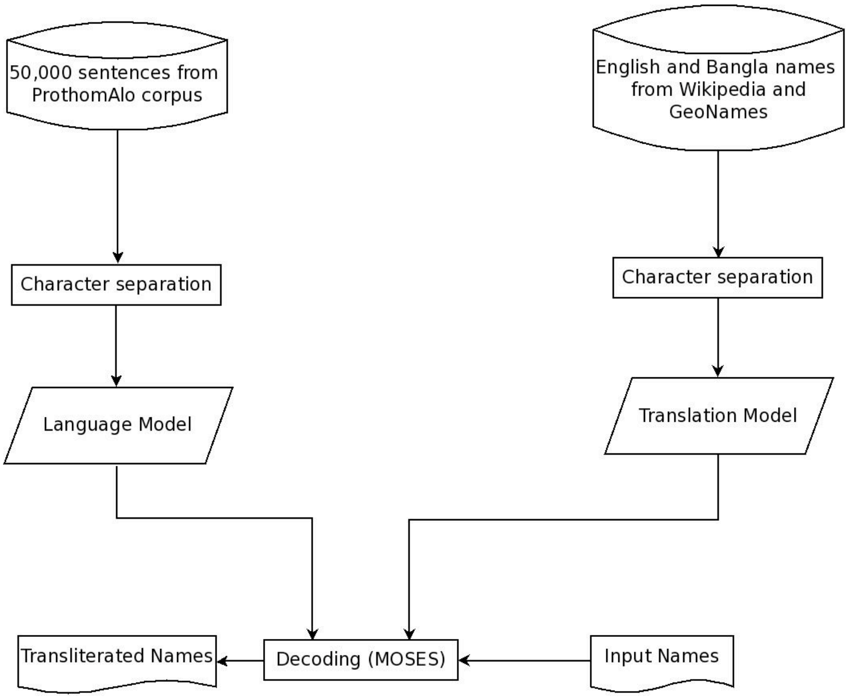

Figure 4.2: Transliteration Module Architecture

4.3.1 MOSES

MOSES is an open-source toolkit for statistical machine translation. MOSES is an extended phrase-based MT system with factors and confusion network decoding. We can integrate morphological, syntactic and semantic information as factors during training. In the factored translation model the surface form may be augmented with different factors such as POS tags or lemma. Figure 4.3 shows the factored translation, which is taken from [Koehn et al., 2007]. The confusion network allows the translation of ambiguous sentences. This enables, for instance, the tighter integration of speech recognition and machine translation Instead of passing along the one best output of the recognizer, a network of different word choices may be examined by the machine translation

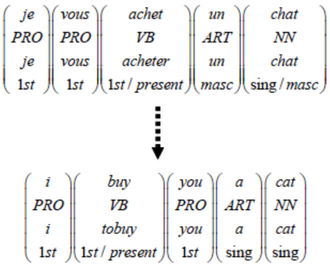

Figure 4.3: Factored translation

system [Koehn et al., 2007]. MOSES has an efficient data structure that allows memory-intensive translation model and language model by exploiting larger data resources with limited hardware. It implements an efficient representation of phrase translation table using the *prefix tree* structure, which allows to load only the fraction of phrase table into memory that is needed to translate the test sentences. MOSES uses the *beam-search* algorithm that quickly finds the highest probability translation among the exponential number of choices.

4.3.2 GIZA++

GIZA++ is a statistical machine translation toolkit that is used to train IBM Models 1-5 and an HMM word alignment model. It is an extension of GIZA (part of the SMT toolkit EGYPT). It includes IBM models 3 and 4. It uses the *mkcls* [Och, 1999] tool for unsupervised classification to help the model. It also implement the HMM alignment model and various smoothing techniques for fertility, distortion/alignment parameters. We can build a bilingual dictionary from a parallel corpus using this tool. More details about GIZA++ can be found in [Och and Ney, 2003]

4.3.3 SRILM

SRILM is a toolkit for language modeling that can be used in speech recognition, statistical tagging and segmentation, and statistical machine translation. It is a freely available collection of C++ libraries, executable programs, and helper scripts. It can build and manage language models. SRILM implements various smoothing algorithm such as Good-Turing, Absolute discounting, Written-Bell and modified Kneser-Ney. Besides the standard word-based N-gram backoff models, SRILM implements several other LM types [Stolcke, 2002], such as Word-Class based N-gram models, Cache-based models, Disfluency and hidden event language models, HMM of N-gram models and more.

4.3.4 Mert

MERT is a tool for minimum error rate training, which is included in MOSES. This tool is the implementation of minimum error rate training in [Och, 2003] and [Venugopal and Vogel, 2005]. This tool has been extended to randomized initial conditions, permuted the model order to deal with the greedy nature of the algorithm, and tune the dynamic parameter range to increase their potential relative impact. This tool is used to optimize decoding performance.

4.3.5 BLEU

Bilingual Evaluation Understudy (BLEU) is a machine translation evaluation technique that is quick, inexpensive, and language independent. [Papineni et al., 2001] claimed that BLEU correlates well with human judgment on both fluency and adequacy. There are however a number of criticisms that have been voiced. [Callison-Burch et al., 2006] shows that improved BLEU score is neither necessary nor sufficient for achieving an actual improvement in translation quality. But still BLEU remains one of the most popular metrics for MT evaluation.

BLEU uses a modified form of *N-gram* precision to compare a candidate translation with multiple reference translations. It applies a length penalty (brevity penalty) if the generated sentence is shorter than the best matching (in length) reference translation. This evaluation technique does not work well for languages without word boundaries like Chinese.

We have used two different versions of BLEU to evaluate our system. One is NIST BLEU (version 11b) and the other is the BLEU tool that comes with MOSES. There is a difference between these two. In case of multiple references, BLEU (MOSES) uses the closest reference translation length whereas BLEU (NIST) uses the shortest reference translation length to calculate the brevity penalty.

4.3.6 NIST

The NIST metric is derived from the BLEU evaluation criterion but differs in one fundamental aspect: instead of n-gram precision the information gain from each n-gram is taken into account. The idea behind this is to give more credit if a system gets an n-gram match that is difficult, but to give less credit for an n-gram match which is easy [Doddington, 2002]. The brevity penalty BP again penalizes shorter system outputs compared to reference translations.

4.3.7 TER

Translation Error Rate (TER) is an error metric for machine translation that measures the number of edits required to change a system output into one of the references. This technique is a more intuitive measure of “goodness” of machine translation output – specifically, the number of edits needed to fix the output so that it semantically matches a correct translation [Snover et al., 2006]. Human-targeted TER yields higher correlations with human judgment than BLEU.

4.3.8 HunAlign

HunAlign is a tool for aligning bilingual text at the sentence level [Varga et al., 2005]. It takes tokenized sentence segments as input and outputs a sequence of bilingual sentences. HunAlign is a dictionary-based sentence alignment method. It uses dictionary and sentence length during alignment. The user can provide a dictionary otherwise it will create a dictionary using sentence length information then it will perform alignment using both dictionary and sentence length information. HunAlign does not support cross alignment.

4.3.9 GMA Sentence Aligner

Geometric Mapping and Alignment (GMA) uses the Smooth Injective Map Recognizer (SIMR) algorithm for mapping bilingual text. The SIMR is a greedy algorithm which relies on the high correlation between the lengths of mutual translations [Melamed, 1996]. SIMR infers bilingual text maps from likely two-dimensional point of correspondence. After finding these points, SIMR selects points whose geometric arrangement most resembles the typical arrangement of point of correspondence.

4.3.10 LingPipe Name Entity Recognizer

LingPipe is Java based natural language processing toolkit distributed by Alias-i[1] . It contains a Name Entity Recognition (NER) tool. The NER tool has an HMM interface with several decoders: first-best (Viterbi), n-best (Viterbi forward) and confidence-based (forward-backward) [Carpenter, 2006].

4.3.11 OpenNLP

Open NLP is an open source NLP toolkit that hosts a variety of java-based NLP tools which perform sentence detection, tokenization, POS-tagging, chunking and parsing, named-entity detection, and coreference using the OpenNLP Maxent machine learning package[2]. This tool is used for Parts of Speech (POS) tagging.

4.4 Experiments

4.4.1 Baseline Translation System

The EMILLE English corpus has 72 text files, where as Bangla EMILLE corpus has 70 text files. We deleted two files from the English side and finally got 10,850 sentences on both sides using the GMA sentence aligner. The English corpus contains 199,973 words and The Bangla corpus contains 189,495 words. Each side of the KDE corpus contains 35,366 sentence pairs. The KDE Bangla corpus contains 221,409 words and English corpus contains 157,392 words. We separated 500 sentence pairs from EMILLE corpus and 1,000 sentence pairs from the KDE corpus for development sets. We also separated same number of sentences from both corpus as a test set. The *5-gram* language model was built from the EMILLE monolingual corpus, Prothom-Alo corpus and training data, which together contain more than 21 million words. Table 4.2 shows sample output of our Baseline System.

4.4.2 Transliteration Module

Generally, SMT systems are trained using large parallel corpora. These corpora consist of several million words, still they can never be expected to have a complete coverage especially over highly productive word classes like proper nouns. When translating a new sentence, SMT systems use the knowledge acquired from training corpora. If they come across a word not seen during training, then they will at best either drop the unknown word or copy into the translation. The Table 4.2 shows that there are some words in the output (Bangla) text which are not translated by the baseline system, These are the OOV words or English words in the Bangla training corpus. So, a transliteration system is an emerging system, which can be incorporated with the baseline system to handle proper nouns or OOV words. But, there are risks of using a transliteration module for OOV words or names. When we transliterate names, the output translation contains some English words (unknown words) and when we

[1] LingPipe is available at:http://alias-i.com/lingpipe/index.html

[2] taken from OpenNLP website: http://opennlp.sourceforge.net/index.html

Table 4.2: Sample output of Baseline System

English	**Bangla**
a shopper‘s guide	একটি shopper এগুল
your legal rights	আপনার আইনগত অধিকার
office of fair trading	office of ন্যায্য ব্যবসা
dti publications orderline	dti publications orderline
the office of fair trading also has a new director general mr john vickers october 2000	office - ন্যায্য ব্যবসা বাধ্যতামূলকভাবে একটি নতুন শতাংশই সাধারণ মিঃ john vickers অক্টোবর উনিশশ
this is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	এই নয় হিসাবে কঠিন এইটি হিসেবে ধ্বনি এবং করতে threat সেটা যায়নি যথেষ্ট পারেন।
mahmoud ahmadinejad has denied the holocaust, describing it a ”myth”.	mahmoud ahmadinejad হয়েছে প্রত্যাখ্যাত যে holocaust, এটা ”myth”
ban ki-moon	ban ki-moon

Table 4.3: Sample Output of Transliteration Module

Input (English)	Output (Bangla)	Reference
Kunchinjunga	কাঞ্চনজঙ্ঘা	কাঞ্চনজঙ্ঘা
Saudi Arabian Kingdom	সৌদি আরব কিংডম	সৌদি আরব
Mauritius	মৃতুস	মরিশাস
Costa Rica	কোস্টারিকা	কোস্টারিকা
Greece	গ্রিকা	গ্রীস

transliterate OOV words, there are some words which should not be transliterated.

We collected 2,200 unique names from Wikipedia and geonames. To build this system, we tried to go a step further than the translation system and treat the words (names) as sequences of letters, which have to be translated into a new sequence of letters. We used the same tools as the translation system. For the language model we extracted 50,000 lines of text from Prothom-Alo corpus. We just put one space between each character in corpora. We used the same tools as the translation system (e.g. MOSES, GIZA++, MERT) and followed the same steps (e.g. training, tuning and testing) as well. Table 4.3 shows some sample output of the transliteration system.

4.4.3 Combined System

We combined the transliteration system with the translation system. The Transliteration system is only responsible for transliterating names or OOV words. As a preprocess, we identified names or OOV words. To identify names, we used the LingPipe [3] Name Entity Recognizer (NER). All names identified by LingPipe were sent to transliteration. Then, we replaced names (English) in

[3] LingPipe is available at : http://alias-i.com/lingpipe/index.html

Table 4.4: Sample English names and its transliteration with XML markup

<np translation=”অফিস ”> office </np>)
<np translation = ”জন ভিকারস”> john vickers </np>
<np translation = ”বান- কি-মুন”> ban-ki-mon</np>

Table 4.5: Sample Output of Combined System

English	**Bangla**
A shopper‘s guide	একটি শপর এগুলো
Your legal rights	আপনার আইনগত অধিকার
Office of fair trading	office of ন্যায্য ব্যবসা
DTI publications orderline	dti publications orderline
The office of fair trading also has a new director general Mr John Vickers October 2000	office - ন্যায্য ব্যবসা বাধ্যতামূলকভাবে একটি নতুন শতাংশই সাধারণ মিঃ জন ভিকারস অক্টোবর উনিশশ
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	এই নয় হিসাবে কঠিন এইটি হিসেবে ধ্বনি এবং করতে threat সেটা যায়নি যথেষ্ট পারেন ।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	মামড হমদনজদ হয়েছে প্রত্যাখ্যাত যে োলকস্ , এটা " িথ " ।
Ban Ki-Moon	বান- কি-মুন

the test data with the transliterated names (Bangla). Each transliterated name was XML marked up. Table 4.4 shows example of XML markup. MOSES has an advanced feature by which we can provide external knowledge to the decoder during decoding. The *-xml-input* flag was raised with *exclusive* value so that the XML-specified translation (transliterated name) is used for the input phrase and any phrase from phrase tables that overlaps with that span is ignored. The same procedure was followed for OOV words as well. We used *english.vcb* as a vocabulary list which is generated by GIZA++ during sentence alignment. For both the cases the BLEU score was very low as compared to the baseline system described above. The Presence of many English words (or even complete sentences) in the Bangla training, development and test corpus is the reason behind the significant drop of the BLEU score. We will show the result in a later chapter. Transliterating OOV words outperforms transliterating names (identified by Lingpipe) in terms of BLEU score. Table 4.5 shows some sample output of our combined system where all OOV words were transliterated.

4.4.4 Corpus Cleaning

Table 4.5 shows that the output contains many English words. There are many English words available in the Bangla side of training, development and test corpus, even some cases of entire English sentences. This happened for both the EMILLE and the KDE corpus. So, the corpora had to be cleaned to get better accuracy. To clean the EMILLE corpus, we experimented with the Interactive Sentence Aligner (ISA) [Tiedemann, 2006] tool. ISA is an interactive tool

Table 4.6: English Names, Bangla Transliteration and Bangla Translation in Training Corpus

English Corpus	**Bangla Corpus**
Third Report (Volumes 1 & 2) of the Low Pay Commission	পরিবর্তনের সূচনা লো পে কমিশন (Low Pay Commission)- এর তৃতীয় প্রতিবেদন (খন্ড ১ ও ২)
We analysed relevant data and worked with the Office for National Statistics (ONS) in order to establish better estimates of the incidence of low pay	আমরা সমস্ত প্রাসঙ্গিক তথ্য বিশ্লেষণ করেছি এবং কম মজুরির প্রভাবের হিসাব আরো ভালোভাবে করতে পারার জন্য অফিস ফর ন্যাশনাল স্ট্যাটিস্টিকস্ (Office for National Statistics - ONS) বা জাতীয় পরিসংখ্যান দফতরের সঙ্গে একত্রে কাজ করেছি ।
7 The definition of the National Minimum Wage adopted in the Regulations has worked well and the great majority of employers has found the operation of the National Minimum Wage Regulations unproblematic.	৭ রেগুলেশনস্ (Regulations) বা আইনকানুনে গৃহিত জাতীয় ন্যূনতম মজুরীর সংজ্ঞা ভালোভাবে কার্যকর হয়েছে এবং নিয়োগকারীদের গরিষ্ঠাশ ন্যাশনাল মিনিমাম ওয়েজ রেগুলেশনস্ (National Minimum Wage Regulations) বা জাতীয় ন্যূনতম মজুরী আইনকানুনের কার্যকারিতার ব্যাপারে

with web interface for sentence alignment of parallel XML documents. It uses sentence length to align sentences. The hard boundaries can be added manually to improve the quality of the automatic sentence alignment and correct the existing alignment by adding/removing the segment boundaries.

ISA corpus alignment tool was not enough to clean/align the EMILLE corpus. Most of the files in English corpus vary on line numbers of the same translation file in Bangla side. For example: the English text about child education has a total of 662 sentences, whereas the Bangla translation of this file has a total of 425 sentences. In some cases, translations on either side were missing. Another noticeable observation was that for any organization or group name in the English text, there were Bangla translation, Bangla transliteration and English name available in the Bangla corpus. Table 4.6 shows some examples of this irregularity. We found two among seventy files where we could not align the sentences at all. Either of the sides was not the translation of the other side. These two file deleted from both sides.

The KDE corpus also contains many English words in Bangla side. We extracted the same line on both the sides where there was no English character on the Bangla side. Finally, we have got total 9,111 sentence pairs from the EMILLE corpus and 16,389 sentence pairs from the KDE corpus. We also cleaned the names for the transliteration system. There were some English names with Bangla translation, which should not have been in the training corpus of the transliteration system. This is because there are Bangla translations available for those words. We identified those names manually and deleted them from the corpus.

Table 4.7: Sample Output of New Translation System

English	Bangla
A shopper‘s guide	একটি shopper এমন নির্দেশিকা
Your legal rights	আপনার আইনগত অধিকার
Office of fair trading	অফিসে ন্যায্য লেনদেনের
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	হিসেবে এই কঠিন নয় এটি ধ্বনি এবং শুধু যে ভীতি প্রদর্শন করা সেটা যথেষ্ট হতে পারে , যাতে অভিযোগটির মীমাংসা করা যায় ।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	mahmoud ahmadinejad আছে যে খান্ডবদহণ , বর্ণনা করা হচ্ছে এটা করা হয়েছে একটি ”myth” ।
Ban Ki-Moon	ki-moon নিষেধ

4.4.5 New Translation Systems

From the manually cleaned and aligned corpora. We selected 500 sentences from EMILLE corpus and 1,000 sentences from KDE corpus as development set and the same number of sentences for test set. The new language model has been cleaned as well, we deleted all the sentences which contain any English characters. This time, the *8-gram* language model is used. In chapter five, we will compare different *N-gram* language models for Bangla. Finally, we found better output and a significant improvement on the old baseline system. Table 4.7 shows some sample outputs of the new baseline translation system. One noticeable observation is that the word ”Ban” is wrongly translated. Here ”Ban” is the part of the name ”Ban Ki-Moon”, but in English ”Ban” is also a verb and the Bangla translation is নিষেধ which we can see in the output. This type of ambiguity will remain in our system.

We also cleaned the names for the transliteration system. Finally, we came up with 3,735 non-unique names. We followed the same procedure as our old baseline transliteration system. The *8-gram* character level (spaced) language model was built from the same text used for the language model of the translation system. We separated 135 unique names for testing the transliteration system.

4.4.6 New Combined System

Again we combined transliteration module with the translation system as we described before. We identified OOV words the same way as the old combined system. The output of the combined system looks better than the new translation system's output. Table 4.8 shows the sample output. Here, the names are correctly transliterated except the word ”Ban”.

4.4.7 Handling Preposition

We have already mentioned in chapter three, the importance of handling preposition during English to Bangla machine translation. English prepositions are

Table 4.8: Sample Output of New Combined System

English	**Bangla**
A shopper‘s guide	একটি শোপপার এমন নির্দেশিকা
Your legal rights	আপনার আইনগত অধিকার
Office of fair trading	অফিসে ন্যায্য লেনদেনের
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	হিসেবে এই কঠিন নয় এটি ধ্বনি এবং শুধু যে ভীতি প্রদর্শন করা সেটা যথেষ্ট হতে পারে , যাতে অভিযোগটির মীমাংসা করা যায় ।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	মাহমুদ আহমেদিনাজাদ আছে যে , বর্ণনা করা হচ্ছে এটা করা হলোকাস্ট একটি "মিথ" ।
Ban Ki-Moon	কি-মোন নিষেধ

translated to Bangla by attaching inflections to the head noun of the prepositional phrase or as a post position word after the head noun. We implemented this idea. We took the intersection of word alignment using the “intersection” option during the training of MOSES. Then, we extracted the intersection word list from our training corpus. As there is no freely available Parts of Speech (POS) tagger for Bangla, we used the OpenNLP [4] tool to POS tag English words and transfer the tags to the aligned Bangla words. For many English words there were more than one candidate tags. In this case we considered only the top 1 candidate. Finally, we separated words, which are tagged as noun.

We preprocessed corpora in two steps. Firstly, we come up with 19 postpositional words. We identified those postpositional words in the corpora and moved them before the reference object (head noun). Secondly, we came up with a group of 9 suffixes which can be attached to nouns. We just stripped those suffixes from the nouns and put them in front of the noun with a suffix mark (#X#, where X is a suffix). We did these for training, development and monolingual corpus for language model. Table 4.9 shows the sample output of combining the preposition handling module with the previous combined system.

[4] OpenNLP is available at: http://opennlp.sourceforge.net/index.html.

Table 4.9: Sample Output of Final Combined System

English	**Bangla**
A shopper‘s guide	একটি স্পার এর নির্দেশিকা
Your legal rights	আপনার আইনগত অধিকার
Office of fair trading	)অফিসে ন্যায্য বাণিজ্য
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	এটি কঠিন নয় এবং এটা কেবল এই গণ্ডগোলের হুমকি এটা হতে পারে যথেষ্ট অভিযোগটির মীমাংসা করার জন্য ।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	মাহমুদ আহমেদিনাজাদ আছে যে , বর্ণনা করা হচ্ছে এটা করা হলোকাস্ট একটি " মিথ " ।
Ban Ki-Moon	কি-মোন নিষেধ
In some cases it may be necessary to go to court to get the matter settled.	কোন কোন ক্ষেত্রে প্রয়োজনীয় হতে পারে কোর্টে যে সেটেল্ড পেতে পারেন ।
This offers the additional benefit to consumers of a 14 day cooling-off period on most goods sold by members of the direct selling association .	আপনার ক্রেতাদের যে অতিরিক্ত বেনিফিট কোন সময়ের ১৪ দিন কলিঙ্গ-অফফ ওপর বেশীর ভাগ মালবাহী sold সদস্যদের বিক্রি করে সরাসরি অ্যাসোসিয়েশন ।

Chapter 5

Results

In this chapter we present the results of the experiments described in the previous chapter. For the translation system, we used 500 sentence pairs from EMILLE corpus and 1,000 sentence pairs from KDE corpus as test sets; 135 English names are used to evaluate the transliteration system.

5.1 Evaluation Criteria

The evaluation of machine translation output is the measurement of the quality of the output. Assessing the quality of a translation is inherently subjective, there is no objective or quantifiable quality of machine translation. A metric will be understood as a measurement of the machine translation quality. The task of any metric is to assign scores of quality in such a way that they correlate with human judgement of quality. The measure of evaluation for metrics is the correlation with human judgement. Recently, many researches have focused on MT evaluation metrics which resulted in a variety of different metrics. A single metric criterion for the evaluation of machine translation does not exist. Therefore, we have used three different metrics for evaluating translation system and another three different metrics for transliteration evaluation.

5.1.1 Translation System's evaluation Metrics

- BLEU [Papineni et al., 2001]
  The metric calculates scores for individual segments, generally sentences, and then averages these scores over the whole corpus in order to reach a final score. It measures the precision of unigrams, bigrams, trigrams, and fourgrams with respect to a reference translation with a penalty for too short sentences.

- NIST [Doddington, 2002]
  The NIST score is based on the BLEU metric, but with some alternations. It is a weighted n-gram precision in combination with a penalty for too short sentences.

- TER [Snover et al., 2006]

The TER is an extension of WER [1]. In addition to the standard edit operations insertion, substitution and deletion a new operation is introduced: shift of whole phrases are permitted.

5.1.2 Transliteration System's Evaluation Metrics

- Word Accuracy in Top-1 (ACC)
  This metric is also known as Word Error Rate, it measures the correctness of the top transliteration candidate in the n-best candidate list produced by transliteration system. The ACC value 1 signifies that all top candidates are correct transliteration i.e. they match with the reference, and ACC = 0 signifies that none of the top transliterations are correct.

$$ACC = \frac{1}{N}\sum_{i=1}^{N}\{1 \quad \text{if} \quad r_i = c_i, \ 0 \quad \text{otherwise}\}$$

- Top 5, Top 20
  The percentage of correct transliteration in the top 5 and top 20 candidates. These are slightly different from ACC.

$$Top\ J = \frac{1}{N}\sum_{i=1}^{N}\{1 \quad \text{if} \ \exists\ c_{i,j} \quad c_{i,j} = ri, 0 \quad \text{otherwise}\}$$

- Mean F-score
  The mean F-score measures how different, on average the candidate transliteration is from its reference. For each source word, the F-score is a function of Precision and Recall. F-score 1 means candidate translation matches the reference, and 0 means there are no common characters between the candidate and reference. Recall and Precision are calculated based on the Longest Common Subsequence (LCS) between the candidate and the reference:

$$LCS(c, r) = 1/2\,(length\,(c) + length\,(r) - editDistance\,(c, r))$$

editDistance(c,r) is the edit-distance between candidate (c) and reference (r). Recall, Precision and F-score for i^{th} word are then calculated as:

$$R_i = \frac{LCS\,(c_i, r_i)}{length\,(r_i)} \quad P_i = \frac{LCS\,(c_i, r_i)}{length\,(c_i)} \quad F_i = 2\frac{R_i \times P_i}{R_i + P_i}$$

[1] The WER is an MT evaluation metric, computed as the minimum number of substitution, insertion and deletion operation that have to be performed to convert the output sentence into the reference sentence.

Table 5.1: Evaluation of Baseline System

Test corpus	BLEU (MOSES)	BLEU (NIST)	NIST score	TER
EMILLE	1.19	1.40	1.62	0.89
KDE	13.71	15.20	4.22	0.70
Combined	5.16	5.20	2.66	0.84

Table 5.2: Evaluation of Transliteration Module

Test corpus	ACC(Top1)	Top 5	Top 20	Mean F-score
English Names	0.122	15.57	18.85	0.686

5.2 Evaluation

5.2.1 Baseline System

Table 5.1 shows the evaluation result of the baseline translation described in section 4.4.1. We are using a training corpus with a small number of tokens for SMT, the result was unexpected. The average sentence length of EMILLE corpus is 91.22 and the average sentence length of KDE corpus is 24.99. The BLEU score of KDE test set is 11 times higher than the EMILLE test set. Sentence length is paying as a factor for this irregularity. The result shows that it is very easy to beat the baseline system.

5.2.2 Transliteration Module

We have built the transliteration module to avoid English words in translated text. Table 5.2 shows the result of baseline transliteration module. It can produce only 12% accurate words. The top 5 and top 20 candidate translations don't match with reference transliteration as we expected. But the mean F-score shows that the generated transliteration candidates are reasonably close to reference transliteration.

5.2.3 Combined System

Table 5.3 shows the output of the combined system described in section 4.4.3. The BLEU score goes down when we combine the transliteration module with the baseline translation system. Our reference translation contains many English words and in many cases entire sentences. The combined system is transliterating all the OOV words. So, some words in the test set are transliterated while these are English words in the reference translation.

5.2.4 New Translation System

After cleaning the corpus there are no English words in the training, development and test sets except some abbreviations. Table 5.4 shows the output of new translation system. There is a significant amount of improvement after cleaning the corpora. The BLEU score is double than the baseline system. The TER error rate also goes down. Still, the BLEU score for longer sentences (EMILLE

Table 5.3: Evaluation of Combined System

Test corpus	BLEU (MOSES)	BLEU (NIST)	NIST score	TER
EMILLE	1.18	1.20	1.65	0.90
KDE	13.21	14.00	4.19	1.02
Combined	4.93	5.10	2.70	0.89

Table 5.4: Evaluation of New Translation System

Test corpus	BLEU (MOSES)	BLEU (NIST)	NIST score	TER
EMILLE	4.84	5.10	3.1	0.84
KDE	21.63	22.50	5.18	0.65
Combined	10.73	11.10	4.24	0.78

test set) is very low. But, the BLEU score of KDE test set looks reasonable for a low density language pair.

5.2.5 New Transliteration Module

Table 5.5 shows the result of new transliteration module. The accuracy is far better than for the old transliteration module, especially the top 20 transliteration candidate list contains 80% correct transliteration. Only 18% times the top transliteration candidates are accurate. In most cases, the transliteration module is making mistakes with one or few characters and the whole transliterated text is counted as wrong. Also, the mean F-score shows that the candidate transliterations are very close to the reference transliterations. The F-mean of our transliteration system is better than the transliteration system described in [Jiang et al., 2009]. They have used more than five hundred thousand names for training.

5.2.6 New Combined System

Table 5.6 shows the evaluation of combining the transliteration module with the translation system. We have transliterated all the OOV words. There is not much improvement due to low accuracy of the transliteration module. But, the BLEU score of combined test set goes up 0.05 than new translation baseline and the TER error rate goes down by 0.01. The transliteration module has a positive effect in the combined system though it is not visible enough. The output of the combined system looks better than the output of the translation system. The result will improve if we train the transliteration module with more names.

Table 5.5: Evaluation of New Transliteration Module

Test corpus	ACC(Top1)	Top 5	Top 20	Mean F-score
English Names	0.187	29.68	79.68	0.797

Table 5.6: Evaluation of Combing New Transliteration Module with Translation System

Test corpus	BLEU (MOSES)	BLEU (NIST)	NIST score	TER
EMILLE	4.85	5.40	3.13	0.83
KDE	21.67	23.20	5.16	0.63
Combined	10.78	11.40	4.25	0.77

Table 5.7: Evaluation of Combing New Transliteration Module and Preposition Module with Translation

Test corpus	BLEU (MOSES)	BLEU (NIST)	NIST score	TER
EMILLE	4.88	5.70	3.16	0.83
KDE	21.67	23.30	5.18	0.63
Combined	10.80	11.70	4.27	0.76

5.2.7 Final Combined System

We have added the preposition handing module with the previous combined system. Table 5.7 shows the evaluation result. Again, we got some improvement over the previous combined system of the transliteration module with the translation system. This result shows that preposition should be handled during English to Bangla machine translation. One important observation is that the BLEU score of KDE test set is same as the previous combined sysetm. We have checked the KDE test set, most of the sentences are instruction messages and prepositions is not used very often. This final combined system gained the BLEU score of 5.64 than the baseline translation system.

5.2.8 Comparison with Anubadok

We have compared our system with an available open source MT system for English to Bangla called Anubadok [2]. We have used the same test set as we used to evaluate our system. Our system clearly out performs Anubadok. Table 5.8 shows the comparison result. The BLEU score of the Anubadok system is 0.84 while our system's BLEU score is 11.70. Table 5.9 shows sample translation by Anubadok and our system. None of the translation candidates generated by either of the systems are close to the reference translation. Anubadok's translation contains some English words which are OOV words. Our system transliterated these words except *sold*; *sold* is not an OOV word in our system. We will discuss this issue later in the Discussion section. Finally, we can conclude that the translation by our system looks better than the translation by Anubadok.

[2] Anubadok is available at: http://bengalinux.sourceforge.net/cgi-bin/anubadok/index.pl

Table 5.8: Comparison Between Our system and Anubadok Online

System	BLEU (MOSES)	BLEU (NIST)	NIST score	TER
Anubadok	0.84	1.60	1.46	1.03
Our System	10.80	11.70	4.27	0.76

Table 5.9: Sample Translation by Anubaok and Our System

English Senteces	**Anubadok**	**Our System**
A shopper‘s guide	একটি shopper‘s পরিচালনা করে	একটি স্পার এর নির্দেশিকা
Your legal rights	আপনার আইনি অধিকার	আপনার আইনগত অধিকার
Office of fair trading	উজ্জ্বল বাণিজ্যের অফিস	অফিসে ন্যায্য বাণিজ্য
This is not as difficult as it sounds and just the threat of it could be enough to resolve matters.	এইটি এইটি হিসেবে যত কঠিন শব্দ করে না এবং এইটির ভয়ের কারণ মাত্র বিষয় স্থিরসঙ্কল্প করতে যথেষ্ট।	এটি কঠিন নয় এবং এটা কেবল এই গণ্ডগোলের হুমকি এটা হতে পারে যথেষ্ট অভিযোগটির মীমাংসা করার জন্য ।
Mahmoud Ahmadinejad has denied the holocaust, describing it a ”myth”.	Mahmoud Ahmadinejad খান্ডবদহণ, বর্ণনা করা হচ্ছে এইটি একটি অস্বীকার করেছে ” myth ”	মাহমুদ আহমেদিনাজাদ খান্ডবদহণ যে , বর্ণনা করা হচ্ছে এটা করা হলোকাস্ট একটি " মিথ " ।
Ban Ki-Moon	নিষেধ Ki-Moon	কি-মোন নিষেধ
In some cases it may be necessary to go to court to get the matter settled.	কিছু ঘটনা এইটিতে বিষয় পেতে কোর্টতে যেতে প্রয়োজনীয় মীমাংসা করতে পেরেছিল।	কোন কোন ক্ষেত্রে প্রয়োজনীয় হতে পারে কোর্টে যে সেটেল্ড পেতে পারেন ।
This offers the additional benefit to consumers of a 14 day cooling-off period on most goods sold by members of the direct selling association.	এইটি সরাসরি selling সমিতির সদসবৃন্দের মধ্যে ভাল সর্বাপেক্ষা একটি ১৪ দিন cooling-off পর্যায়কালের ভোক্তাতে অতিরিক্ত সুবিধা প্রস্তাব দেয় বিক্রি করে।	আপনার ক্রেতাদের যে অতিরিক্ত বেনিফিট কোন সময়ের ১৪ দিন কলিঙ্গ-অফফ ওপর বেশীর ভাগ মালবাহী sold সদস্যদের বিক্রি করে সরাসরি অ্যাসোসিয়েশন ।

Chapter 6

Conclusions

In this chapter, we will discuss shortcoming of our system, summarize this work and point out directions for future work.

6.1 Discussion

SMT systems require a significant amount of parallel corpora to achieve satisfactory translations. There are not enough parallel corpora available between English and Bangla to achieve high quality translation using only a statistical MT system. Our system works reasonably well under the circumstances; and something line transliteration module should be further improved. After transliterating all the OOV words, there are some English words in the translated text. We have used *english.vcb* as a vocabulary list. Any word not in the vocabulary list is considered as an OOV word. Table 4.9 and Table 5.9 show that *sold*, is an English word in the translated text. As a reason, we have found 53 phrase-table entries where *sold* is the part of the phrase but, there is no phrase-table entry where *sold* is a single phrase and also no entry where *sold* is following or preceding adjacent words in the test sentences. Table 6.1 shows some phrase-table entries of *sold.* We have experimented with different alignment options available in MOSES. We used *grow-diag-final*, *grow* and *intersection* alignment options. All cases some English words were left in the translated text after transliterating all OOV words.

6.2 Summary

In this thesis work, we presented an English to Bangla phrase-based machine translation system. We incorporated two modules with the baseline translation system to improve the translation accuracy and quality. We also showed that an automatic transliteration system can be built from Phrase-based SMT system and its performance is comparable to the state-of the-art transliteration system designed for transliteration [Jiang et al., 2009].The transliteration module's *ACC -Top 1* score is 0.18 but the transliterated words are very close to reference translation when considering the mean F-score. There is no concept of preposition in Bangla. Bangla uses postposition and inflectional attachment

Table 6.1: Phrase-table Entry for *sold*

English Phrase	**Bangla Bangla Phrase**	$\phi\,(e\vert f)$
be **sold on the streets**	এর কোন কোনটি বাইরে বিক্রিও	0.111111
be **sold on the**	এর কোন কোনটি বাইরে বিক্রিও	0.000191987
be **sold on**	এর কোন কোনটি বাইরে বিক্রিও	0.111111
even be **sold on the**	এর কোন কোনটি বাইরে বিক্রিও করা	0.111111
may be **sold**	বিক্রি হতে পারে	0.5
may be **sold**	হতে পারে	0.00209205
sold using the designs as trade names	বাণিজ্যিক নাম ব্যবহার করে '	0.333333

with head nouns instead of English preposition. The preposition handing module is also effective to improve translation accuracy. Even though there is not much improvement after combing preposition handling module, but it shows that preposition should be handled in the English ⇔ Bangla machine translation.For short sentences, the BLEU score of our system is 21.67 and TER is 0.63 which are quite reasonable for a low density language. The overall system performance is BLEU (MOSES): 10.80, BLEU (NIST): 11.70, NIST: 4.27 and TER: 0.76.

6.3 Future Work

Although satisfactory results (for a low density language) were obtained with the current modules of the system and the architecture proposed in each one of these modules, there is still place for improvement in several parts of the system. Obviously there is no alternative other than adding more parallel data for training the system. The Center for Research on Bangla Language Processing (CRBLP) [1] at BRAC University, Bangladesh is currently developing a parallel corpus of 1 million words. Our plan is to incorporate the CRBLP corpus as training data.

A test set with more than one reference would be useful to evaluate our system. So, our plan is to develop a test set for English⇔ Bangla MT system with more than one reference sentences.

Only 18% translated names are correct. We need to add more names in the training corpus in order to improve the transliteration quality. So, we will collect more names and re-train our system.

In our system, we have considered only 19 postpositional words and 9 inflectional suffixes for the preposition handing module. Adding more postpositional words and inflectional suffixes would improve the system's accuracy. Both English and Bangla have many compound words, so another module that could

[1] Web address of CRBLP: http://crblp.bracu.ac.bd/

handles English compound words would be useful for English ← Bangla MT system.

Nowadays, linguistic features are being used to enhance phase-based SMT systems. So we will work towards integrating some linguistics features (i.e. : syntactic information, morphological information) in our MT system.

Bibliography

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. 19(2): 263–311, 1993.

P.F. Brown, J. Cocke, S. A. Della Peitra, V. J. Della Peitra, F. Jelinek, J. C Lai, and R. L. Mercer. Method and system for natural language translation, 1995.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL*, 2006.

Bob Carpenter. Character language models for chinese word segmentation and named entity recogntion. In *Proceedings of the 5th ACL Chinese Special Interest Group (SIGHan)*, 2006.

Sajib Dasgupta, Abu Wasif, and Sharmin Azam. An optimal way towards machine translation from english to bengali. In *Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT)*, 2004.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 2002.

Andrew Finch and Eiichiro Sumita. Phrase-based machine transliteration. In *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific SpeechTranslation (TCAST)*, 2008.

Xue Jiang, Le Sun, and Dakun Zhang. A sylleable-based name transliteration system. In *Proceedings of Name Entities Workshop(NEW) 2009, ACL-IJCNLP*, 2009.

K. Knight and J. Graehl. Machine transliteration. *Computational Linguistics*, 24(4):599–612, 1997.

P Koehn. *Noun Phrase Translation.* PhD thesis, University of Southern California, 1993.

P Koehn. Pharaoh: A beam search decoder for phrase based statistical machine translation models. In *Proceedings of AMTA*, 2004.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

David Matthews. Machine translation of proper names. Master's thesis, School of Informatics, University of Edinburgh, 2007.

Dan Melamed. A geometric approach to mapping bitext correspondence. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1996.

SK Naskar and S Bandyopadhyay. Handling of prepositions in english to bengali machine translation. In *Proceedings of the EACL 2006 Workshop*, 2006a.

SK Naskar and S Bandyopadhyay. A phrasal ebmt system for translating english to bengali. In *Proceedings of the Workshop on Language, Artificial Intelligence, and Computer Science for Natural Language Processing Applications (LAICS-NLP)*, 2006b.

F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

F.J. Och. An efficient method for determining bilingual word classes. In *Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics(EACL)*, 1999.

F.J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 4th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

F.J. Och and H. Ney. The alignment template approach to statistical machine translation. 30(4):317–449, 2004.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a method for automatic evaluation of machine translationt. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2001.

D Saha and S Bandyopadhyay. A semantics-based english-bengali ebmt system for translating news headlines. In *Proceeding of MT Summit X second workshopon example-based machine translation*, 2005.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhou. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, 2006.

A. Stolcke. SRILM–an extensible language modeling toolkit. In *Proceedings of the ICSLP*, 2002.

J. Tiedemann. ISA and ICA – two web interfaces for interactive alignment of bitexts. In *Proceedings of LREC 2006*, 2006.

Naushad UzZaman, Arnab Zaheen, and Mumit Khan. A comprehensive roman (english) to bangla transliteration scheme. In *Proceedings of International Conference on Computer Processing on Bangla*, 2006.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Proceedings of the RANLP Conference*, 2005.

Ashish Venugopal and Stephan Vogel. Considerations in maximum mutual information and minimum classification error training for statistical machine translation. In *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*, 2005.

D Vilar, JT Peter, and H Ney. Can we translate letters? In *Proceedings of ACL workshop on SMT*, 2007.

S Vogel. Smt decoder dissected: Word reordering. In *Proceedings on the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003.