English to Bangla Statistical Machine Translation using A* Search Algorithm

Hoque Mohammed Moshiul[1], and Mohammad Kamrul Hassan[2]

[1]Department of Computer Science & Engineering, Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh, E-mail: moshiul_240@cuet.ac.bd
[2]Department of Humanities, Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh, E-mail: mkhasan_cuet@yahoo.com

Abstract

A statistical translation model is a mathematical model in which the process of human language translation is statistically modeled. Model parameters are automatically estimated using a corpus of translation pairs. This paper presents a statistical model for translating the different kinds of English sentences into Bangla. A lot of alignments are possible between translation of English and Bangla sentences. In this paper, we have used a fertility model that finds the possible meanings of English word into Bangla word with corresponding probabilities. For calculating the probabilities of all possible alignments we have used distortion model and chosen the best one that have the highest probability. For searching process, an efficient algorithm called A search algorithm has been used in this paper that provides a good result. We have tested our proposed model using several English sentences with different word lengths and got successful translation for most of the test cases.*

Keywords: *Statistical Machine Translation, Heuristic Function, Fertility, Distortion, Language Model, and Alignment.*

1. Introduction

Machine Translation (MT) refers to the application of computers for the task of translating automatically from one natural language to another. The differences between languages and especially the inherent ambiguity of language make MT a very difficult problem. Traditional approaches to MT have relied on human's supplied linguistic knowledge in the form of rules to transform text into one language to another. Given the vastness of language, this is a highly knowledged intensive task. Statistical MT is a radically different approach that automatically acquires knowledge from large amounts of training data. This knowledge, which is typically in the form of probabilities of various language features, is used to guide the translation process. MT is a hard problem, because natural languages are highly complex, many words have various meanings. Therefore, different translations are possible for same sentence. Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability. The best translation, of course, is the sentence that has the highest probability. Translation model have been used for statistical machine translation in Ures [1], word alignment of a translation corpus in Melamed [2], multilingual document retrieval in Ward [3], automatic dictionary construction in Melamed [4], and data preparation for word sense disambiguation programs in Roossin [5]. Few research works have been done on SMT for Bangla to English translation. One of these works has been presented in Ali [6] where simple Bangla sentences were used and processed. In Wasif [7] presents an MT technique from English to Bangla using parsing and English to Bangla translation using rule-based parser has been presented in Mondol [8]. Besides these, Berwick [9] describes computation linguistic analysis of Bangla using GB theory. Considering the structure of the sentence does now-a-days most of the translation. The major problem in structure based translation is that if one can translate a specific structured sentence then it is not able to translate all other sentences that are in different structure. To resolve this problem we have to use SMT, which does not consider the structure of the sentence. Another advantage of SMT is that it has no grammatical-complexity. In this paper, we have proposed an SMT model to

translate English to Bangla sentences and verify the model with various kinds of English sentence for translating corresponding Bangla sentences with statistical manner.

2. Statistical Machine Translation (SMT)

A statistical model is a mathematical model in which the process of human language translation is statistically modeled. In SMT, a text in source language is translated into a target text of another language. In translation model, a source string of one language, $e_1^J = e_1 ... e_i ... e_J$, which is to be translated into a target string $\hat{b}_1^I = b_1 ... b_i ... b_I$ of another language. Among all possible target strings, we will choose the string with the highest probability:

$$\hat{b}_1^I = \arg\max_{b_1^I} \{\Pr(b_1^J \left| e_1^I \right.)\} = \arg\max_{b_1^I} \{\Pr(b_1^I).\Pr(e_1^J \left| b_1^I \right.)\} \quad (1)$$

The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. $\Pr(b_1^I)$ is the language model of the target language, whereas $\Pr(e_1^J \left| b_1^I \right.)$ denotes the translation model [10]. The $\Pr(b_1^I)$ factor helps b_1^I output to be natural and grammatical, while $\Pr(e_1^J \left| b_1^I \right.)$ factor ensures that b_1^I is normally interpreted as e_1^J. Many statistical models have been used word-to-word mappings between source and target words. These mappings are called alignment [11, 12]. The alignment mapping is $j \to i = a_j$ from source position j to target position $i = a_j$. In statistical alignment, $\Pr\left(b_1^J, a_1^J \left| b_1^I \right.\right)$ the alignment is used as a hidden variable.

3. Statistical MT Model 4

Various statistical translation models have been introduced for SMT. We used the Model 4 for translating of English sentence into Bangla. It is consisting four sub-models; translation model, fertility model, language model and distortion model. The final probability $\Pr(e_1^J, a_1^J \left| b_1^I \right.)$ for Model 4 is obtained by multiplying the probabilities of the sub-models for all words [13]. The structure of SMT Model 4 is shown in Fig. 1.

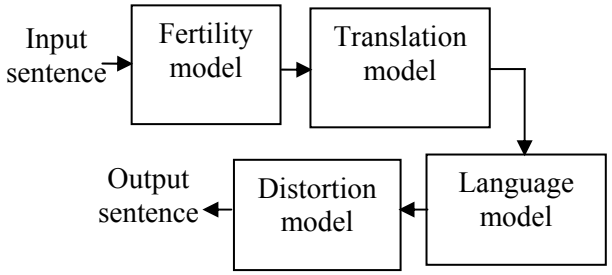


Fig. 1: General structure of SMT Model 4

3. 1 Fertility Model

The fertility (F) model $p(\phi|b)$ for the probability that a target language word b is aligned to ϕ source language words.

3. 2 Translation Model

The lexicon model $p(e|b)$ for the probability that the source word e is a translation of the target word b. It is also called the translation (T) model.

3. 3 Language Model

The language (L) model $p(b)$ for the probability that a target sentence is how much grammatical.

$$P^L(b) = \max_{u,v} p(b|u,v) \quad (2)$$

3. 3. 1 Trigram Language Model

The easiest way to break a string down into its components is to consider substrings. An n-word substring is called n-gram. If $n = 2$, we say it bi-gram. Similarly for $n = 3$ called trigram and $n = 1$ is called unigram or word. In this project, we use tri-gram language model.

$$b(z|xy) = \text{number-of-occurrences}("xyz") / \text{number-of-occurrences}("xy") \quad (3)$$

The process of determining the probability of a sentence using tri-gram language model according to the equation (3) are shown in Table 1, based on the parallel corpora for 100 sentences. Here 'start' and 'end' indicates start-of-sentence and end-of-sentence.

Table 1: Finding probability of a sentence using tri-gram language model

Input sentence	Sub-string	Probability	Overall probability of input sentence
p(I eat rice every	*b* (I | start start)*	0.15 (15/100)	0.0000099
	b (eat | start I)*	0.20 (3/15)	
	b (rice | I eat)*	0.33 (1/3)	
	b (everyday | eat rice)*	0.05 (5/100)	
	b (end| rice everyday)*	0.02 (2/100)	

3. 4 Distortion Model

In the distortion model (D), the probability $\Pr(j|i,m,n)$ that the translations of a target word in position i generating some source word in position j, when the respective lengths of the sentences are m and n.

$$\hat{b}_1^J = \arg\max_{b_1^I}\left\{\Pr\left(b_1^I\right)\sum_{a_1^J}\Pr\left(e_1^J, a_1^J \middle| b_1^I\right)\right\} = \arg\max_{b_1^I}\left\{\Pr\left(b_1^I\right)\max_{a_1^J}\Pr\left(e_1^J, a_1^J \middle| b_1^I\right)\right\} \quad (4)$$

Here a_1^J denotes the alignment that may contain $a_j = 0$. When source word is not aligned to any target word. Alignment a_1^J is used as a hidden variable in the distortion model. The search space consists of the set of all possible target language strings b_1^I and all possible alignments a_1^J [14].

4. A* Search Algorithm

Now we have a way of finding the best translation given a set of candidate translations using equation (5). We cannot practically consider each sentence in the target language. Therefore, we need a heuristic search method that can efficiently find the sentence with the highest probability. Every partial hypothesis consists of a prefix of the target sentence and a corresponding alignment. A partial hypothesis is extended which yields an extension score that is computed by taking into

account the lexicon, distortion, and fertility probabilities involved with this extension. A partial hypothesis is called open if more source words are to be aligned to the current target word in the following extensions. Every extension of an open hypothesis will extend the fertility of the previously produced target word and an extension of a closed hypothesis will produce a new word [15]. The basic A* search can be described by the following steps:

Step1: Initialize priority queue with an empty hypothesis.
Step 2: Remove the hypothesis with the highest score from the priority queue.
Step 3: If this hypothesis is a goal hypothesis output this hypothesis and terminate.
Step 4: Produce all extensions of this hypothesis and put the extensions to the queue.
Step 5: Go to step 2.

4. 1 Admissible Heuristic Function

The success of A* rests heavily on the heuristic function chosen. To guarantee that we will find the optimal solution, if one exists, the heuristic must be "admissible". The heuristic function estimates the probability of a completion of a partial hypothesis. We have used a good admissible heuristic function taking into account distortion probabilities and the coupling of lexicon, fertility and language model probabilities. The simplest realization of a heuristic function, denoted as $h^T(j)$, takes into account only the translation probability $p(e|b)$:

$$h^T(j) = \max p(e_j|b) \tag{5}$$

This heuristic function can be expressed by introducing the fertility probabilities of a target word b:

$$h^{TF}(j) = \max\left\{\max_{b\neq b_0,\phi} p(e_j|b)\sqrt[\phi]{p(\phi|b)}, p(e|b_0)\right\} \tag{6}$$

A coupling between translation and fertility probabilities has been achieved. By taking the ϕ-th root, we can avoid that fertility probability of a target word whose fertility is larger than one computed for every source word aligned to it. The language model can be incorporated by considering that for every target word there exists an optimal language model probability for trigram model.

$$P^L(b) = \max_{u,v} p(b|u,v) \tag{7}$$

Therefore, a heuristic function including a coupling among translation, fertility, and language model probabilities can be expressed as equation (8).

$$h^{TFL}(j) = \max\left\{\max_{b,\phi} p(e_j|b)\sqrt[\phi]{p(\phi|b)\,p^L(b)}, p(e|b_0)\right\} \tag{8}$$

The heuristic function for distortion probabilities depend on the model. For Model 4, we have to use equation (9).

$$h^D(j) = \max p(j|i,m,n) \tag{9}$$

Here, a target word in position i generating some source word in position j, when the respective lengths of the sentences are m and n. The model makes use of the fact that longer sentences in one language tend to be translated into shorter sentence [15]. This yields the following heuristic functions taking into account translation, fertility, language, and distortion model probabilities.

$$H^{TFLD}(n) = \prod h^{TFL}(j).h^D(j) \tag{10}$$

3. Proposed Statistical Translation Process

The translation process can be divided into several steps. The overall translation process can be depicted as in Fig. 2.

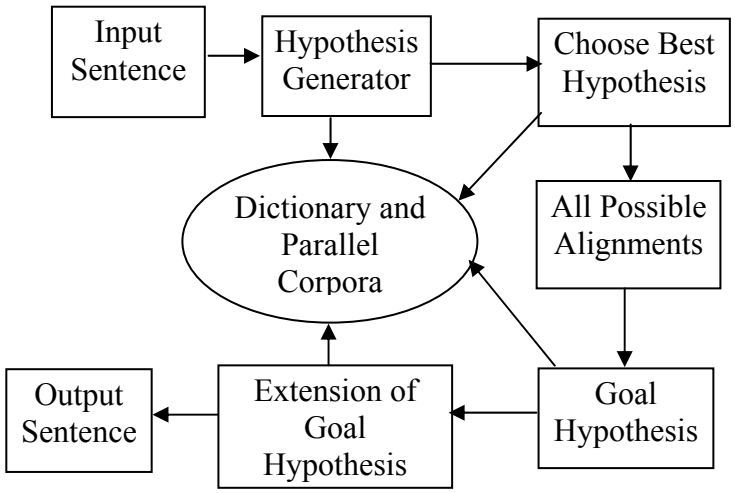


Fig. 2: Systematic process of SMT

Input sentence is taken which is in source language. Considering all target words corresponds to each of the input word generates all possible target strings. Alignments for each hypothesis are considered and among them considering the highest probability of heuristic function identify the goal hypothesis and finally all extensions of goal hypothesis are produced to get the output sentence in target language.

3. 1 Input

At first, an input sentence is taken which have to be translated in Bangla. For example, consider an input sentence "**you play in the field.**"

3. 2 Dictionary and Bi-lingual Corpora

In the word dictionary all possible meanings of a target word are given. A small fragment of dictionary is shown in Table 2. Here 'Examples' column is used to understand under which condition corresponding meaning of a word will be taken.

Table 2: A small fragment of dictionary

Words	**Meaning**	**Examples**
You	Zzwg	**You** eat
	‡Zvgv‡K	I shall give **you**
	‡Zvgvi	**You** have a book
Play	‡Lwj	I **play**
	†Lj	You **play**
	†L‡j	They **play**
In	G	**in** field
	†Z	**in** river
The	wU	**The** boy
Field	gvV	in **field**
I	Avwg	I eat
	Avgvi	**I** have a book
Eat	LvB	I **eat**
	LvI	You **eat**
	Lvq	They **eat**

Rice	fvZ	I eat **rice**

The Bi-lingual parallel corpora of two languages will also be given as Table 3.

Table 3: Parallel Corpora

English sentences	**Bangla sentences**
I play	Avwg †Lwj
You play	Zzwg †Lj
You catch fish in the river	Zzwg b`x‡Z gvQ ai
Is he playing football	†m wK dzUej †Lj‡Q
What are you doing now	Zzwg GLb wK KiQ
Do the work	KvRwU Ki
May you live long	Zzwg `xN©Rxwe nI
How nice the girl is	evwjKvwU wK my›`i

3. 3 Hypothesis Generator

At first we use the fertility model on the input sentence and then we generate all hypothesis i.e. all target language strings.

3. 3. 1 Application of Fertility Model

By using fertility model, we find out the meaning of which input English word will be available in the Bangla output sentence by using word dictionary and parallel text dictionary. If one Bangla word meaning is available for an input English word then '1' is placed. Similarly, '2' is placed if two Bangla word is available for an English word and '0' is placed if none of the Bangla meaning is available. For the given input English sentence, for example, "You play in the field", Table 4 shows the fertility value for each English word using the equations (3) and (6). The √ sign indicate that its probability is high and the corresponding fertility will be taken.

Table 4: Fertility measurement

Words	Condition (trigram)	Fertility		Probability
You	You | start start	0		0
		1	√	1 (40/40)
		2		0
Play	Play | start you	0		0
		1	√	1 (20/20)
		2		0
In	In | you play	0		0
		1	√	1 (12/12)
		2		0
The	The | play in	0	√	1 (22/22)
		1		0
		2		0
Field	Field | in the	0		0
		1	√	1 (15/15)
		2		0

Therefore, we have got the following fertility representation.

	you	play	in	the	field
	↓	↓	↓	↓	↓
Fertility:	1	1	1	0	1

Then the word of fertility **0** will be removed and the word of fertility **1** will be placed one time and similarly the word of fertility **2** is placed two times in the sentence. So we get the sentence "**You play in field".**

3. 3. 2 Generation of Target Language Strings

The possible meanings of a word and generate all possible sentences in target language that are shown in Table 5. Here for the sentence "**you play in field**" all possible sentences in target language are generated.

Table 5: Possible target strings

Input sentence	**Possible target strings**
You play in field	Zzwg ‡Lwj G gvV
	Zzwg ‡Lwj †Z gvV
	Zzwg ‡Lj G gvV
	Zzwg ‡Lj †Z gvV
	Zzwg ‡L‡j G gvV
	Zzwg ‡L‡j †Z gvV
	‡Zvgv‡K ‡Lwj G gvV
	‡Zvgv‡K ‡Lwj †Z gvV
	‡Zvgv‡K ‡Lj G gvV
	‡Zvgv‡K ‡Lj †Z gvV
	‡Zvgv‡K ‡L‡j G gvV
	‡Zvgv‡K ‡L‡j †Z gvV
	‡Zvgvi ‡Lwj G gvV
	‡Zvgvi ‡Lwj †Z gvV
	‡Zvgvi ‡Lj G gvV
	‡Zvgvi ‡Lj †Z gvV
	‡Zvgvi ‡L‡j G gvV
	‡Zvgvi ‡L‡j †Z gvV

3. 4 Choosing the Best Hypothesis

Applying the translation model on the generated target language strings we find out the best hypothesis with the highest probability.

3. 4. 1 Application of Translation Model

Using translation model, we find the proper translation for each word. For the given English sentence we find the Bangla words using equation (3) and (5). Table 6, show the possible translation of English word into Bangla with probability.

Table 6: Translation Probability measurement

Words	**Condition (tri-gram)**	**Translation**		**Probability**
You	You | start start	Zzwg	√	0.85 (34/40)
		‡Zvgv‡K		0.025 (1/40)
		‡Zvgvi		0.125 (5/40)
Play	Play | start you	‡Lwj		0
		†Lj	√	1 (20/20)
		†L‡j		0
In	In | you play	G	√	0.533 (8/15)
		†Z		0.277 (4/15)

The	The | play in	wU		0
Field	Field | in the	gvV	√	1 (15/15)

Here √ indicates higher probability. So, the Bangla words with the highest probability are

you	play	in	field
↓	↓	↓	↓
Zzwg	†Lj	G	gvV

Since the probability of the words Zzwg, †Lj, G, gvV are the highest, so among all sentences in table 6, probability of the sentence " Zzwg †Lj G gvV" will be the highest and that is why, "Zzwg †Lj G gvV" will be the best hypothesis.

3. 5 Possible Alignment

Generally, one defines an alignment as a relation between the words in the English sentence and the words in the Bangla sentence [11]. It is a generative process through which a string from a source alphabet is mapped to a rooted tree whose nodes are labeled from a target alphabet. Alignments for the hypothesis can be shown as Fig. 3.

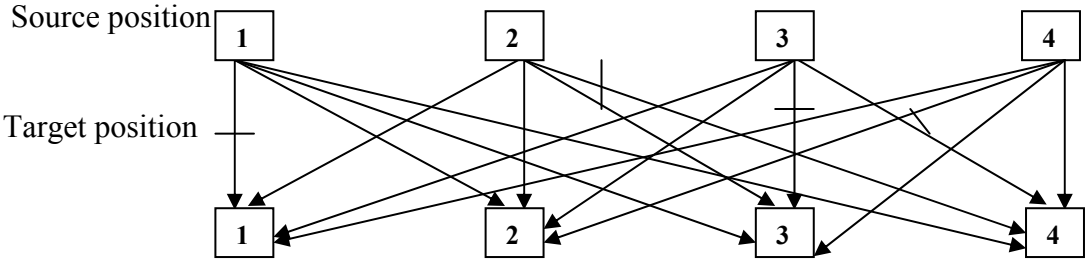


Fig. 3: Possible alignments of the hypothesis

3. 6 Goal Hypothesis

From all possible alignments we have to find out the correct one for which the generated target sentence is more grammatical and for which the probability will also be large.

3. 6. 1 Application of Distortion Model

We use the distortion model to get perfect alignment. It considers the length of both English and Bangla sentence and searches in the parallel text dictionary for each English word's position to the corresponding Bangla word's position. Using the equation (9) the probability of all possible alignments is generated which is shown in Table 7.

Table 7: Distortion Probability measurement

Length of source text	Length of target text	Source position	Target position		Probability
5	4	1	1	√	0.56(28/50)
			2		0.16(8/50)
			3		0.2(10/50)
			4		0.08(4/50)
		2	1		0.16(8/50)
			2		0.2(10/50)
			3		0.04(2/50)
			4	√	0.6(30/50)
		3	1		0.2(10/50)
			2		0.24(12/50)
			3	√	0.5(25/50)

			4		0.06(3/50)
		4	1		0.04(2/50)
			2	√	0.7(35/50)
			3		0.16(8/50)
			4		0.1(5/50)

Here the length of English sentence is 5 (since 5 words in "**you play in the field**") and the length of Bangla sentence is 4 (since 4 words in "Zzwg †Lj G gvV"). By sorting the position and replacing the Bangla words we get as: Zzwg gvV G †Lj. The mapping among sorting position of the Bangla words can be depicted in Fig. 4. Sometimes the distortion model may not be able to get perfect alignment. To resolve this problem we use language model.

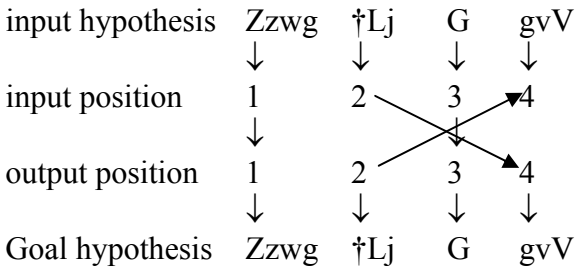
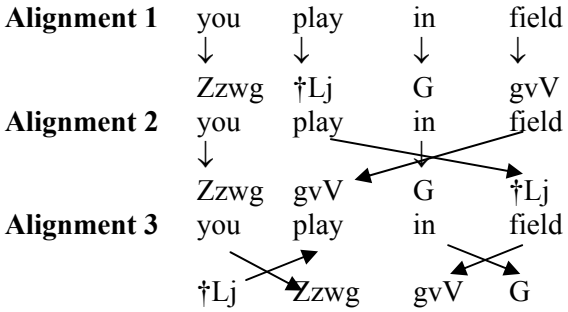


Fig. 4: Mapping of input and output positions.

3. 6. 2 Application of Language Model

Using language model we find out which alignment is grammatical. For this, we first consider all possible alignments of the generated hypothesis and determine which alignment's probability is large. We can generate several sentences in target language by considering different alignments according to Fig. 3 that can describe in the following way:



In this way, we can generate many sentences in target language that are shown in Table 8 by considering all possible alignments of Fig. 3. Using the equation (7), the probability for all sentences in Table 8 is determined. Then the goal hypothesis is chosen by considering the highest probability, which is "Zzwg gvV G †Lj".

3. 7 Extension of Goal Hypothesis

In the extension process of the goal hypothesis we have to add or remove one or two words from the goal hypothesis if require to get an appropriate target sentence. Then we have to check whether this sentence in target language can produce the given input sentence in source language. If it can generate the input sentence then we have got the output else we have to take all other meanings except the one that has already taken for each word of the goal hypothesis and by repeating all the steps we have to generate the real goal hypothesis. For the given input sentence we don't need to add or remove any word with the goal hypothesis and the goal hypothesis "Zzwg gvV

G †Lj" can generate the input sentence "**you play in the field**". So "Zzwg gvV G †Lj" is the required output sentence in target language.

Table 8: Probability of Language model

Input hypothesis	Possible Hypothesis	$p(b)$ Language model	Order	Goal hypothesis
Zzwg †Lj G gvV	Zzwg †Lj gvV G	(0.4)	2	Zzwg gvV G †Lj
	Zzwg †Lj G gvV	(0.01)	5	
	Zzwg G gvV †Lj	(0.000001)	12	
	Zzwg G †Lj gvV	(0.000001)	11	
	Zzwg gvV †Lj G	(0.0001)	9	
	Zzwg gvV G †Lj	**(0.8)** √	**1**	
	†Lj Zzwg G gvV	(0.0000001)	13	
	†Lj Zzwg gvV G	(0.1)	4	
	†Lj G gvV Zzwg	(0.0000001)	15	
	†Lj gvV G Zzwg	(0.001)	6	
	†Lj G Zzwg gvV	(0.00001)	10	
	†Lj gvV Zzwg G	(0.0000001)	16	
	gvV G Zzwg †Lj	(0.2)	3	
	gvV G †Lj Zzwg	(0.001)	7	
	gvV Zzwg †Lj G	(0.0001)	8	
	gvV †Lj Zzwg G	(0.0000001)	14	

3. 8 Output

Finally we get the correct output sentence in target language. For the given input sentence the corresponding output sentence in target language is "Zzwg gvV G †Lj|".

3. 9 Translation of Different Types of English Sentences

Using this SMT procedure it is possible to translate all kinds of English sentences into Bangla sentences; including assertive sentences, interrogative sentences, imperative sentences, and

exclamatory sentences respectively. Some example translations have been shown in Table 9. The fertility model generates fertility value for the given input English sentences. The translation model removes the words of 0 fertility and generates corresponding Bangla words for the remaining English words in that sentence. Finally, the distortion model performs the alignment among the target sentences. Table also represented the probability of each type of sentences.

Table 9: Translation of different types of input sentence

Type of input sentence	Fertility	Translation	Alignment	Probability
Interrogative	what class do you read in 1 1 0 1 1 1	What class you read in †Kvb K¬vm Zzwg co G	†Kvb K¬vm Zzwg co G Zzwg †Kvb K¬vm G co	2.854323×10^{-11}
Imperative	Do the work quickly 1 1 1 1	Do the work quickly Ki wU KvR vovZvwo	Ki wU KvR ZvovZvwo KvR wU ZvovZvwo Ki	4.0773×10^{-11}
Exclamatory	How nice the bird is 1 1 1 1 0	How nice the bird wK my›`I wU cvwL	wK my›`I wU cvwL cvwL wU wK my›`I	1.68639×10^{-11}

4. Experimental Results

We have tested our proposed model for different types of English sentences that are taken from different articles of English newspaper and English textbooks respectively. During experimentation, we got successful translation for most of the test cases. Different kinds of translations from English to Bangla have been shown as example in Fig. 5, 6, 7 and 8 respectively.

Example 1
Input Sentence: I eat rice.
Output Sentence: Avwg fvZ LvB|
Probability: $6.17741536940944 \times 10^{-7}$

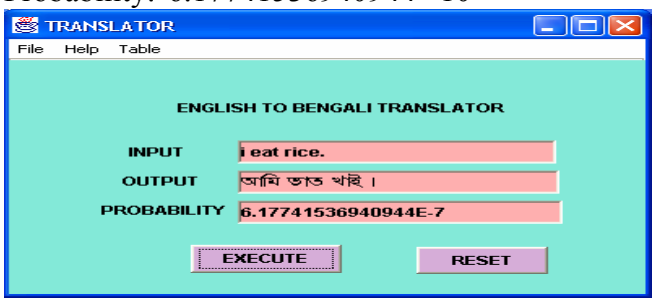


Fig. 5: Output for the sentence "I eat rice. (Avwg fvZ LvB|)"

Example 2
Input Sentence: Where are you going now?
Output Sentence: Zzwg GLb †Kv_vq hv"Q ?
Probability: $5.391569181704293 \times 10^{-10}$

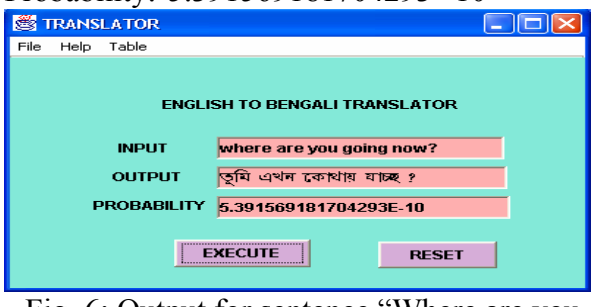


Fig. 6: Output for sentence "Where are you going now? (Zzwg GLb †Kv_vq hv"Q ?)"

Example 3
Input Sentence: May Bangladesh live long.
Output Sentence: evsjv‡`k `xN©Rxwe †nvK|
Probability: $5.555555555555556 \times 10^{-5}$

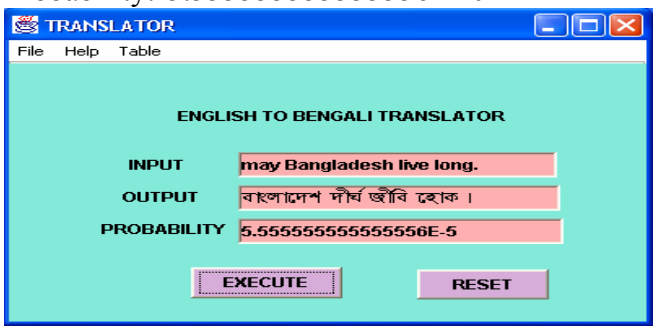


Fig. 7: Output for the sentence "May Bangladesh live long. (evsjv‡`k `xN©Rxwe †nvK|)"

Example 4
Input Sentence: How nice the bird is!
Output Sentence: cvwLwU wK my›`i!
Probability: $1.6863224410394223 \times 10^{-11}$

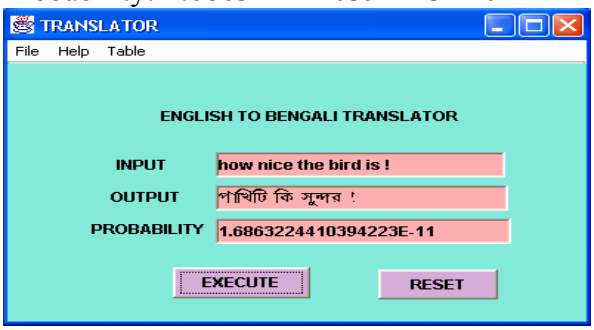


Fig. 8: Output for the sentence, how nice the bird is! (cvwLwU wK my›`i!).

4. 1 Performance Analysis

The corpus statistics that we have used is given in Table 10. We have taken about 2700 sentences of English and Bangla as corpora with about 24300 words of English and 26000 words of Bangla.

Table 10: Training corpus statistics

Types of Language	English	Bangla
No. of input Sentences	2700	2700
Words	24300	26000
Average sentence length	9	8
Vocabulary size	400	550

We have tested our corpora with 6, 8, 10, 12, and 14 word length input English sentences of 50 sentences each category. Total English words and corresponding Bangla words needed for each different length sentence category is given in the Table 11.

Table 11: Test corpora statistics

Word length of input English sentence	Input English sentences	Words	
		Total English words	Corresponding Bangla words
6	50	300	320
8	50	400	410
10	50	500	515
12	50	600	605
14	50	700	710

A comparison between several SMT models is given in Table 12. It is seen that search success rate is decreased with the increase of sentence length and also seen, SMT Model 4 i.e. TFLD search success rate is higher than all other SMT model.

Table 12: Search Success Rate (%)

Word length	Success Rate of Translation			
	T	TF	TFL	TFLD
6	100	100	100	100
8	100	100	100	100
10	88	92	94	98
12	60	74	85	94

The average search time needed for TFLD model using A* Search Algorithm for the parallel corpora of Table 10 is shown in Table 13. With the increasing of word length of sentence the corresponding search time is also increased.

Table 13: Average search time for A*: TFLD with different word length

No. of words	6	8	10	12
Time (in sec)	0.80	0.90	0.94	1.15

6. Conclusion

Traditional MT techniques require large amounts of linguistic knowledge to be encoded as rules. Statistical MT provides a way of automatically finding correlations between the features of two languages from a parallel corpus, overcoming to some extent the knowledge bottleneck in MT. The proposed translation architecture not only successful in translating all types of English sentences to corresponding Bangla sentences, but also does it in most optimal way without any grammatical knowledge and only based on the bi-lingual corpus which is a general procedure of learning a language used by human being. The main advantage of SMT is to reducing grammatical complexity and dependency on structure of the sentence, which are the key obstruction in English to Bangla translation. The limitation of SMT model is its execution time that takes quite more time than structure based translation method to learn from corpora before performing translation. But calculating and storing the searching result previously before translation process start execution and also by using faster system can overcome this limitation. It is often possible to faster compute acceptable results using a beam search approach. Since SMT is in some sense word alignment (with probabilities), it can be used for lexicon acquisition also and A* search algorithm is useful during the development of a SMT module. Translation of large sentences that include phrase and idioms and group verbs may be carry on. Introducing linguistic knowledge can further strengthen the statistical model. Such knowledge may be in the form of morphological rules, rules about word order, idiomatic usages, known word correspondences and so on. For the translation model to work well, the corpus has to be large enough that the model can derive reliable probabilities from it, and representative enough have the domain or sub-domain. Sentence in the source language is often split into multiple sentences, multiple sentences are clubbed into one, and the same idea is conveyed in words that are not really exact translations of each other. In such situations, sentence-alignment itself might be a big challenge.

References

[1] Berger, A., Brown, P., Della, Pietra, S., Della, Pietra, V., Gillett, J., Lafferty, J., Mercer, R., Printz, H. and Ures, L., "Language Translation Apparatus and Method Using Context-Based Translation Models", U.S. Patent 5510981,1996

[2] Melamed, I., "Models of translational equivalence among words", Computational Linguistics, 2000, volume, 26, no. 2, pp. 155-163.

[3] Franz, M., McCarley, J. and Ward, R., "Ad hoc, cross-language and spoken document information retrieval", 1999, IBM, TREC-8.

[4] Resnik, P. and Melamed, I., "Semi-automatic acquisition of domain-specific translation lexicons", 1997, NLP-97.

[5] Brown, P., Cocke, J., Della, Pietra, S., Jelinek, F., Mercer, R. and Roossin, P., "Word-sense disambiguation using statistical methods", 1991, ACL-91.

[6] Uddin, Gias, M., Ashraf, H. Kamal, Hena, A., and Ali, Masroor, M., "New parameters for Bangla to English statistical machine translation" Proceeding of the 2nd International Conference on Electrical and Computer Engineering, ICECE, Dhaka, 2004, volume 1, pp. 545-548.

[7] Dasgupta, S., Azam, S. and Wasif, A., "An Optimal way of Machine Translation From English To Bengali", Proceeding of 7th International Conference on Computer and Information Technology (ICCIT), Dhaka, 2004, pp. 220-234.

[8] Kabir, F. Alam, Shamsul, M. and Mondal, Islam, M., "Parsing for English sentence English to Bengali translation using rule based parser", Proceeding of 1st National Conference on Computer Processing of Bangla (NCCPB), Dhaka, 2003, pp.101-107.

[9] Khan, R. and Berwick, C., "A Computational Linguistic Analysis of Bangla using the GB theory", 1999, Calcutta, India.

[10] Knight, K. "A Statistical MT Tutorial Workbook" April 30, 1999.

[11] Och, J. and Ney, H., "A comparison of alignment models for statistical machine translation", The 18th International Conference on Computational Linguistics, Saarbr¨ucken, Germany, 2000, pp. 1086–1090.

[12] Vogel, S., Ney, H. and Tillmann, C., "HMM-based word alignment in statistical translation", The16th International Conference on Computational Linguistics, Copenhagen, 996, pp. 836–841.

[13] Brown, F., Della, Pietra, A., Della, Pietra, J. and Mercer, L., "The mathematics of statistical machine translation: Parameter estimation", Computational Linguistics, volume 19, no. 2, 1993, pp. 63–311.

[14] Gale, A. and Church, W., "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics, volume 19, no. 1, pp. 78-80.

[15] Och, J., Ueffing, N. and Ney, H., "An Efficient A* Search Algorithm for Statistical Machine Translation", Workshop 1999.

Amount of Figures: Eight (08)
Amount of Tables: Thirteen (13)

Article received: 2008-05-17