A parametric approach to Bangla to English Statistical Machine Translation for complex Bangla sentences -Step 1

Mohammad Gias Uddin, Mahbub Murshed, Muhammad Abul Hasan
Dept. of CSE, International Islamic University Chittagong, Chittagong, Bangladesh
me_gias2003@yahoo.com, mmurshed@gmail.com, mah.cse@gmail.com

Abstract

A Bangla to English Statistical Machine Translation(SMT) Engine design for complex Bangla sentences requires considering different types of Bangla sentences and making decision based on them. There cannot be a unique solution for all types of Bangla sentences, but the different steps taken while solving the different Complex Bangla sentences can be combined later which can solve this mammoth task. In this paper we have used some complex Bangla sentence types and tried to solve them by creating some new parameters in addition to the IBM models and previously generated parameters. The types discussed here include Bangla sentences having doubly occurred words, sentences having implicit words and sentences having multiple subjects. The parameters discussed here are fully compatible with the previously generated methods and parameters.

Keywords: Statistical Machine Translation(SMT), Parameters, IBM Models, Complex Bangla Sentences, Translation Engine.

I. INTRODUCTION

This is the second approach to Bangla to English Statistical Machine Translation after [1]. In the first attempt simple Bangla sentences were used and processed for Statistical Machine Translation. Here we have extended the approach into some more complex Bangla sentences. A complex Bangla sentence may contain a number of verbs, subjects and so its translation to its English counterpart is considerably different from simple Bangla to English Sentence SMT. Machine Translation [2] Engine is mainly based on the Natural Language Process [2]. IBM have devised some model based parametric approaches [3] to solve this problems generally. This approach is more delicately handled in the workbook of Kevin Knight [4]. Besides these, a number of syntactical approaches have been proposed for Bangla to English SMT which include Modified Parsing methodology for Bangla Sentences [5], Morphological Analysis of Bangla words [6], Parsing Method for Bangla Natural Languages [7], Bangla Sentence Parser [8], Desing principle of Automatic Translation system for Bangla to other Natural Languages [9], Development of Bangla Machine Translation Dictionary [10], Bangla Conversion Processor [11], Design implementation of bilingual natural language parser [12], Computer parsing of Bangla verbs [13], Corpus oriented Bangla language parsing [14], Corpus based study of Bangla language [15], Reversibility of NLP [16], Computational linguistic analysis of Bangla using the GB Theory [17] and Semantic approach to Bangla sentences [18].

This paper focuses on some complex Bangla sentences which cannot be translated using the method proposed in [1]. We have introduced some type oriented parameters which will be used along with the existing parameters to go with the translationa process.

II. STATISTICAL MODELS AND PREVIOUSLY PROPOSED PARAMETERS

A sentence of English words e, can be translated into a sentence of Bangla words in various ways. In statistical translation, we take the view that every Bangla sentence **b**, is a possible translation of **e**. We assign to every pair of sentences $(\mathbf{e}, \mathbf{b})$ a number $P(\mathbf{b}|\mathbf{e})$, which we interpret as the probability that a human translator, when presented with **e** will produce **b** as his interpretation. Using Bayes' theorem,

$$P(\mathbf{e}|\mathbf{b}) = \frac{P(\mathbf{e})P(\mathbf{b}|\mathbf{e})}{P(\mathbf{b})}. \quad (1)$$

The appropriate English sentence will be **e**, which can be measured as,

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{argmax}\ P(\mathbf{e})P(\mathbf{b}|\mathbf{e}). \quad (2)$$

For any Bangla to English translation there are different alignments and training involves collecting fractional counts [4] that are weighted by $P(a|\mathbf{b},\mathbf{e})$,

$$P(a|\mathbf{e},\mathbf{b}) = \frac{P(a,\mathbf{b}|\mathbf{e})}{P(\mathbf{b}|\mathbf{e})} = \frac{P(a,\mathbf{b}|\mathbf{e})}{\sum_a P(a,\mathbf{b}|\mathbf{e})}. \quad (3)$$

According to IBM Model 1,

$$P(a,\mathbf{b}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m} t(b_j|e_{a_j}). \quad (4)$$

$t(b|e)$ = translation probability of occurrence of b given e
l = length of English sentence in words
m = length of Bangla sentences in words
a_j = jth word of an alignment a
Neglecting spurious word generation, the probability for each alignment will be,

$$P(\mathbf{b}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m}\sum_{a_1=1}^{l}\cdots\sum_{a_m=1}^{l}\prod_{j=1}^{m} t(b_j|e_{a_j}). \quad (5)$$

According to Model 2, the reverse distortion parameter r which gives the probability of positioning of an English word given position of Bangla word,

$$r(a_j|j,l,m) = P(a_j|a_1^{j-1},b_1^{j-1},l,m). \quad (6)$$

a = particular English alignment for a given Bangla sentence which satisfy the constraints,

$$\sum_{i=1}^{l} r(i|j,l,m) = 1. \quad (7)$$

for each triple j, l and m.
The (5) will be handled as,

$$P(\mathbf{b}|\mathbf{e}) = \epsilon\sum_{a_1=1}^{l}\cdots\sum_{a_m=1}^{l}\prod_{j=1}^{m} t(b_j|e_{a_j})r(a_{i_j}|j,l,m). \quad (8)$$

IBM Model 3 introduces two new parameters Fertility Probability(n) and Distortion Probability(d) measured as,

$n(\phi_i|e_i)$= Probability of any word e_i will have ϕ_i Bangla words in translation.

$$d(b_{a_i}|i,l,m) = P(b_{a_i}|b_{a_1}^{i-1},e_1^{j-1},l,m). \quad (9)$$

$\mathbf{b}_a$ = particular Bangla sentence alignment for a given English sentence.

IBM Model 3 treats the spurious English words as being generated from $0th$ position. We will not consider the $0th$ position here, hence the modified Model 3 equation is,

$$P(a,\mathbf{b}|\mathbf{e}) = \prod_{i=1}^{l} n(\phi_i|e_i)\prod_{j=1}^{m} t(b_j|e_{a_j}) \prod_{j:a_j\neq 0}^{m} d(j|a_j,l,m)\prod_{i=0}^{l}(\phi_i)! \quad (10)$$

To compute $P(\mathbf{b}|\mathbf{e})$,

$$P(\mathbf{b}|\mathbf{e}) = \sum_a P(a,\mathbf{b}|\mathbf{e}). \quad (11)$$

[1] proposed some new parameters in addition to these parameters for simple Bangla to English Statistical Machine Translation. They are the Reverse Rranslation parameter $rev_t(e|b)$, the Reverse Fertility parameter $rev_n(\phi_i|b)$, the Dual Translation parameter $dual_t(b_1,b_2|e)$ and the Bigram probabilities $bigram(e2|e1)$, where
b = Bangla word,
e = English word,
b_p = Bangla word position,
e_p = English word position,
b_l = Bangla sentence length in words,
e_l = English sentence length in words.

The Reverse Translation parameter is estimated as follows,

$$rev_t(e|b) = \frac{P(e,b)}{P(b)}. \quad (12)$$

This parameter is identical to the Translation parameter $t(b|e)$ except that the execution phase is different. The Reverse Fertility parameter is also identical to Fertility parameter $n(\phi_i|e)$ except that it measures the fertility of English word. The Dual Translation parameter is used for those English words which generate two Bangla words using the following estimating equation,

$$dual_t(b_1,b_2|e) = \frac{P(b_1,b_2,e)}{P(e)}. \quad (13)$$

The Bigram probabilites are used to choose the correct verb with respect to the subject. The estimating equation is,

$$bigram(e2|e1) = \frac{P(e_2,e_1)}{P(e_1)}. \quad (14)$$

Here, e_1 will always be the first word in a English sentence, i.e., the subject.

III. NEW PARAMETERS FOR COMPLEX BANGLA SENTECE STRUCTURES

Complex Bangla sentences have more complex relationship between subjects and verbs. Here we will consider only a few types of Complex Bangla sentences which we will statistically translate into their English Counterpart by applying some new parameters in addition to previously generated parameters. The new parameters are,

1. The Bi-Occurred parameter $Bi - occ(e|b_1 b_2)$,
2. The Bi-Distribution parameter $Bi - dist(e_p|b_{p1} b_{p2}, l, m)$,
3. The Absent-Distribution parameter $D - absent(e|b, e_p, l, m)$,
4. The Subject-Chk parameter $Sub - chk(e_{n+1|e_1,e_2,\ldots,and e_n})$.

here,
b_1, b_2 = Bangla words,
$e, e_1, \ldots, e_{n+1}$ = English words,
b_{p1}, b_{p2} = Bangla word positions,
e_p = English word position,
m = Bangla sentence length in words,
l = English sentence length in words.
These parameters are more or less type dependent, so we will discuss them with the specified sentence types in the following sections,

A. The Bi-Occurred and Bi-Distribution Parameter

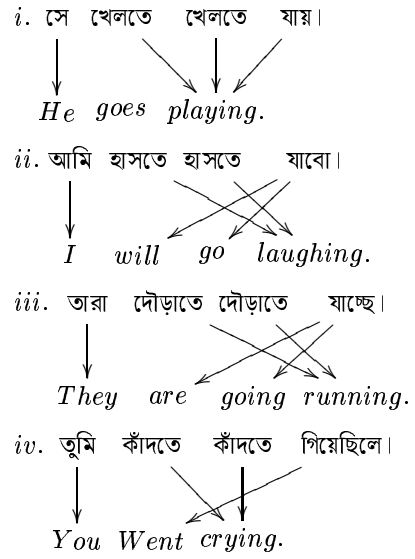

Fig. 1: Example application of Bi-Distribution and Bi-Occurred parameters.

In Fig. 1 one particular type of Bangla sentences is shown where the same Bangla word is coming doubly in every sentence along with the main verb. This doubly occurrance cannot be solved using the existing parameters. Here খেলতে খেলতে is connected to the single English word 'playing', হাসতে হাসতে is connected to 'laughing', দৌড়াতে দৌড়াতে is connected to 'running', কাঁদতে কাঁদতে is connceted to 'crying'. The Bi-Occurred parameter estimation is activated when these type of sentences come. The estimating equation is,

$$Bi - occ(e|b_1, b_2) = \frac{P(e, b_1, b_2)}{P(b_1, b_2)}. \qquad (15)$$

This parameter works along with Bi Bi-Distribution parameter $Bi - dist(e_p|b_{p1}, b_{p2}, l, m)$ which says in which position to place the appropriate English word with respect to the doubly-occurred verbs. The estimating equation is,

$$Bi - dist(e_p|b_{p1}, b_{p2}, l, m) = P(a_j|j, j+1, l, m)$$
$$= P(a_j|a_1^{j-1}, b_1^{P_1}, b_1^{p_2}, l, m). \qquad (16)$$

B. The Absent-Distribution Parameter

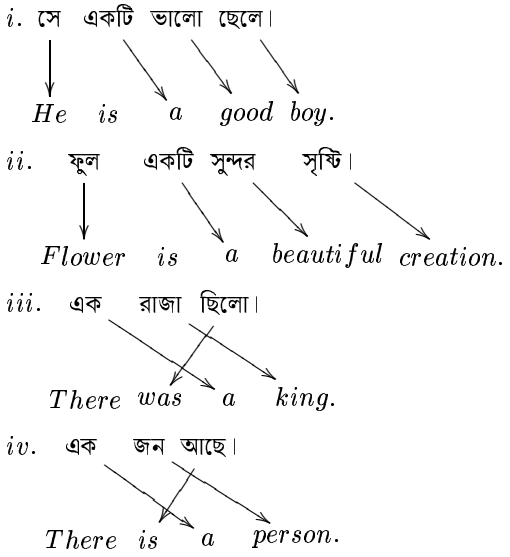

Fig. 2: Example application of D-Absent parameter.

In Fig. 2 some Bangla Sentences are shown where their English counterparts have some implicit words not connected with any of the Bangla words. This type of sentences were shown as the only failure in [1]. In the first sentence 'is' is implicit, in the sencond example also 'is' is implicit whereas in the third and fourth example 'There' is implicit. Now, let us make a

hypothesis: the words implicit in the Bangla to English sentence Statistical Conversion are only are *am, are, was, were and There*. This hypothesis can be proved in almost all cases of Bangla to English SMT design. Based on this hypothesis we will generate the Absent-Distribution parameter $D-absent(e|b,p,l,m)$ which will be estimated as follows,

$$D{-}absent(e|b,e_p,l,m) = P(e_{a_j}|b_1,e_p,l,m). \quad (17)$$

This parameter will work with $rev_t(e|b), rev_n(\phi_i|b)$ and $n(\phi_i|e)$ which will tell the SMT Engine that an Englihs word is linked with any of the Bangla words of the specified Bangla Sentence. Then the $D-absent(e|b,e_p,l,m)$ parameter will check whether this almost implicit word is among the group of the implicit word list. If so, it will estimate the position of word registering it as a found word.

C. The Subject-Check Parameter

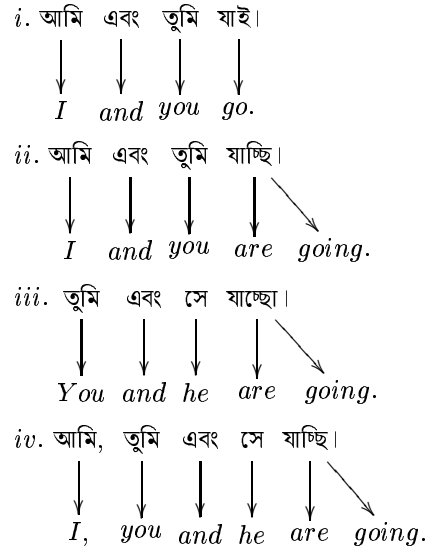

Fig. 3: Example application of Subject-Check parameter.

In Fig. 3 we have introduced another type of Bangla sentences where there are multiple subjects. The change of verbs in an English sentence is directly related to its suject(s). In the first example, verb 'go' is connected to both 'I' and 'you', in the second example 'are going' is related to both 'I' and 'you', in the third example 'are going' is related to both 'you' and 'he', and in the fourth example 'are going' is related to three subjects 'I', 'you', and 'he'. We cannot set limit to the number of subjects in a subjects. Also, we cannot directly identify a subject or a verb using statistical method. But, we can make a common assumption: all subjects appear before the verbs and in the first half of a sentence, if it is a normal language. Now, we can modify the bogram probabilities $bigram(e_2|e_1)$ parameter up to these multiple subjects. The desired equation will be Subject-chk parameter $Sub-chk(e_{n+1|e_1,e_2,\ldots,e_n})$ which will be activated if there is a comma(,) or a 'and' in the English sentence right after the first English word. This will finish its checking only after it has found the second next word(e_{n+1}) after 'and'. The estimating equation will be,

$$\begin{aligned} &Sub-chk(e_{n+1}|e_1,e_2,\ldots,ande_n) \\ &= \frac{P(e_1,e_2,\ldots,ande_ne_{n+1})}{P(e_1,e_2,\ldots,ande_n)}. \end{aligned} \quad (18)$$

IV. TRAINING

Statistical based scheme is dependent mainly on two things, good algorithm and good training data. So, this training scheme should be trained up using a good and huge reference data for a good success rate. The conversion between distortion and reverse distortion parameter will be as shown in [1]. The estimating equation will be,

$$d(\mathbf{b}_p|\mathbf{e}_p,\mathbf{e}_l,\mathbf{b}_l) = \frac{r(\mathbf{e}_p|\mathbf{b}_p,\mathbf{e}_l,\mathbf{b}_l)}{r(\mathbf{e}_p)}, \quad (19)$$

$$r(\mathbf{e}_p) = \sum_{\mathbf{b}_p} r(\mathbf{e}_p|\mathbf{b}_p,\mathbf{e}_l,\mathbf{b}_l). \quad (20)$$

Here,
$\mathbf{e}_l$ = English sentence length,
$\mathbf{b}_l$ = Bangla sentence length,
$\mathbf{e}_p$ = English word position in a particular sentence having length $\mathbf{e}_l$,
$\mathbf{b}_p$ = Bangla word position in a particular sentence having length $\mathbf{b}_l$,
The parameters $rev_t(e|b)$ and $rev_n(\phi_i|b_i)$ will be estimated by the following Model 3 formula,

$$\begin{aligned} P(a,\mathbf{e}|\mathbf{b}) \quad = \quad & \prod_{i=1}^{m} rev_n(\phi_i|b_i) \prod_{j=1}^{l} rev_t(e_j|b_{a_j}) \\ & \prod_{j:a_j\neq 0}^{l} r(j|a_j,l,m) \prod_{i=0}^{m} (\phi_i)! \end{aligned} \quad (21)$$

$$P(\mathbf{e}|\mathbf{b}) = \sum_{a} P(a,\mathbf{e}|\mathbf{b}) \quad (22)$$

The introduced parameter here will be estimated whenever the particular types of sentences will arrive. So, their estimation will not cost so much to the total process and the whole system will work as better as before.

V. PERFORMANCE EVALUATION

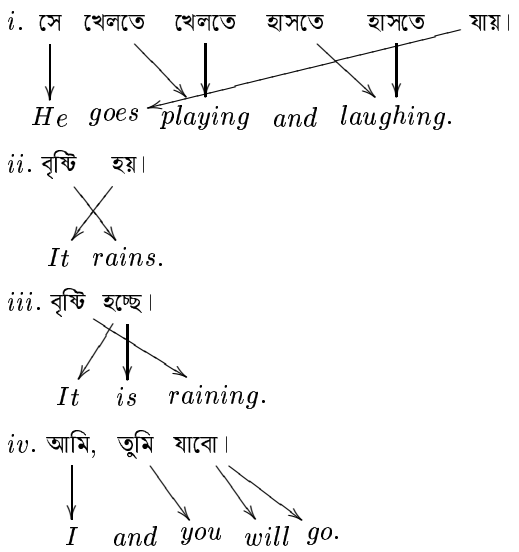

Fig. 4: Constraints in Performance Evaluation.

We have tested our proposed SMT scheme for a number of Bangla Sentences and the outcome was quite satisfactory. But there should be some condition to be maintained strictly for the successful application of the proposed parameters. The Bangla sentences shown in Fig. 4 cannot be translated using the proposed parameter. In the first and fourth example the reason is clear, there is an implicit 'and' which is not used in the Bangla sentences. We will not consider these types of sentences as valid for our proposed scheme. The reason of failure for the second and third examples is the mapping of বৃষ্টি to both 'rains' and 'raining'. This is a unique problem which should be addressed and solved by some modified approach. Our proposed scheme will work efficiently only when we can express the full sentences perfectly.

VI. CONCLUSION

This paper tries to solve some complex Bangla to English sentence Translation statistically. The parameters shown here are quite efficient implementing the translation scheme perfectly based on the condition that there should not be any kind double specification for any Bangla sentences. The main advantage of the scheme proposed here is that the parameter estimation time and space will not be in any circumstances, as they are activated only at a certain moment. Statistical approach is fully based on the training data, so the parameters will behave as they are trained. Hence, training data should be created with proper planning. We have stepped only a bit forward to Complex Bangla Sentece conversion. There is huge step ahead for Idioms and Phrses and that should be the biggest challenge for Bangla to English SMT design. Also, the algorithm should be designed eyeing at the proper utilization of memory for its complete success.

References

[1] M. G. Uddin, H. Ashraf, A. H. M. Kamal, and M. M. Ali, "New parameters for bangla to english statistical machine translation," in *International Conference on Electrical and Computer Engineering, ICECE* (I. C. on Electrical and I. Computer Engineering, eds.), vol. 1, pp. 545–548, December 2004.

[2] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*. Pearson Education, Inc., 2nd ed., 2004.

[3] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of machine translation: Parameter estimation," *Computational Linguisticts*, vol. 19, pp. 263–311, June 1993.

[4] K. Knight, "A statistical mt tutorial workbook," in *A Statistical MT Tutorial Workbook*, April 1999.

[5] M. M. Asaduzzaman and M. M. Ali, "A knowledge based approach to bangla-english machine translation for simple assertive sentences," *International Journal of Translation*, vol. 15, pp. 77–97, July-Dec 2003.

[6] M. M. Asaduzzaman and M. M. Ali, "Morphological analysis of bangla words for automatic machine translation," in *ICCIT 2003*, vol. 1, pp. 271–276, December 2003.

[7] M. M. Hoque and M. M. Ali, "A parsing methodology for bangla natural language sentences," in *ICCIT 2003*, vol. 1, pp. 277–282, December 2003.

[8] D. Arnold, L. Balkan, S. Meijer, R. Humphreys, and L. Sadler, *Machine Translation - An Introductory Guide*. NCC Blackwell Ltd., 1994.

[9] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*. Pearson Education Inc., 2001.

[10] M. I. A. Khan, A. K. M. A. Hossain, and R. C. Debnath, "A bangla conversation processor using natural language processing," in *ICCIT 2002*, pp. 262–266, 2002.

[11] M. M. Ali and M. M. Ali, "Development of machine translation dictionaries for bangla language," in *ICCIT 2002*, (Dhaka, Bangladesh), pp. 267–271, 2002.

[12] M. Kay, *Machine Translation*. Xerox-PARC, Palto Alto, CA and Stanford University.

[13] M. M. Murshed, "Parsing of bengali natural language sentences," in *ICCIT 98*, (Dhaka, Bangladesh), pp. 185–189, 1998.

[14] M. R. Selim and M. Z. Iqbal, "Syntax analysis of phrases and different types of sentences in bangla," in *ICCIT 99*, (Sylhet, Bangladesh), pp. 175–186, 1999.

[15] N. S. Dash, "Corpus oriented bangla language processing," *Jadavpur Journal of Philosophy*, vol. 11, no. 1, pp. 1–28, 1999.

[16] G. V. Noord, *Reversibility in Natural Language Processing*. University of Utrecht, Netherlands, 1993.

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

[18] M. Bhattacharyya, *Syntactic and Semantic Juncture: A Key for MT Grammar Formalism*. Calcutta, India, 1999.