English Translation of Bangla Simple Sentences Using Bilingual Corpus

Conference Paper · February 2011

CITATIONS
0

READS
19

2 authors:

Md Musfique Anwar
Swinburne University of Technology
21 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Md. Al-Amin Bhuiyan
King Faisal University
72 PUBLICATIONS **595** CITATIONS
SEE PROFILE

English Translation of Bangla Simple Sentences Using Bilingual Corpus

Md. Musfique Anwar, Nasrin Sultana Shume and Md. Al-Amin Bhuiyan
Dept. of Computer Science & Engineering, Jahangirnagar University, Dhaka, Bangladesh
Email: musfique.anwar@gmail.com, shume_sultana@yahoo.com, alamin_bhuiyan@yahoo.com

Abstract
Transfer in machine translation (MT) plays an important role for producing correct output. This paper presents a technique to analyze and implement a corpus based automatic Bangla machine translator. The study is based on a bilingual corpus of Bangla and English texts and translation unit alignment. A bilingual dictionary contains the translation probability of English word. Our proposed MT system can be extendable to paragraph translation.

Keywords:
Machine Translation, Bilingual Corpus, Bilingual Dictionary, Translation Probability etc.

1. Introduction

Machine translation (MT) refers the translation from one natural (source) language to another (target language). It is an important area of Natural Language Processing (NLP). MT is a challenging job due to building up a successful translator for producing exact target language output from a source language. At a minimum, transfer systems require monolingual modules to analyze and generate sentences, and transfer modules to relate equivalent translation representations of those sentences [1]. We use statistical approach for machine translation. The Statistical Machine Translation (SMT) constructs a general model of the translation relation, and then let the system acquire specific rules automatically from the bilingual and monolingual text corpora [2].

Many factors make transfer system of MT an attractive issue [3]. These are:

- Many systems are bilingual, or their principal use for translation in one direction between a limited numbers of languages.
- Where full multilinguality is required it is possible to have a hub language into and out of which translation is done.
- Portions of transfer modules can be shared when closely related languages are involved.

2. Statistical Machine Translation (SMT)

The Statistical approach is the use of statistics in computational linguistics. The most established SMT system is based on word for word substitution although some experimental SMT systems employ syntactic processing.

Statistical approaches to MT means:

- Approaches which does not use explicitly formulated linguistic knowledge to perform MT or,
- The application of statistical techniques on calculating probability to aid parts of the MT task (example word sense disambiguation).

The idea behind SMT approach is to let a computer learn automatically how to translate text from one language to another by examining large amounts of parallel bilingual text, i.e. documents which are nearly exact translation of each other. The Statistical MT approach uses statistical data to perform translation. This statistical data is obtained from an analysis of a vast amount of bilingual texts. Different probabilities are extracted from the bilingual texts automatically by a computer and these are:

i) The probability of a source sentence to occur in the texts.
ii) The probabilities of a source word to be translated as one, two, three etc. target words.
iii) The translation probabilities of each word in each language, and
iv) The probabilities of the position of each word in the source language sentence which is not in the same position of the target language word in the target sentence (i.e. the probability of distortion).

These probabilities are vital to the translation process as these are the sole information for calculating how the source language sentence should be translated to the target language form.

2.1 Basic Probabilities

Let us consider that an English sentence *e* may translate into any Bangla sentence *b*. The basic probabilities are given below:

Priori probability, P (*e*): The probability that *e* happens. For example, if *e* is the English string "I eat rice", then P (*e*) is the probability that a certain person at a certain time will say, "I eat rice" as opposed to saying something else.

Conditional probability, P (*b* | *e*): The probability of *b* given *e*. For example, if *e* is the English string "The boy drinks tea" and if *b* is the Bangla string

বালকটি চা পান করে

“ ”, then P (b | e) is the probability that upon seeing e, a translator will produce b.

Joint probability, P (e , b): The probability of e and b both happening. If e and b do not influence each other, then we write P (e , b) = P (e) * P (b). For example, if e stands for “the first roll of the die comes up 5” and b stands for “the second roll of the die comes up 3”, then P (e , b) = P (e) * P (b) = (1/6) * (1/6) = 1/36. If e and b do influence each other, then we had better write P (e , b) = P (e) * P (b | e). That means the probability that “e happens” times the probability that “if e happens then b happens”. If e and b are strings that are mutual translations, then there is definitely some influence.

2.2 Translation process

Even though text alignment is not really a [art of the actual translation process, it serves, as a necessary tool for creating dictionaries and grammars, thus improving the quality of MT [4]. Text alignment is the first step to make bilingual corpora useful. Word alignment means a group of sentences in one language that corresponds in content to some group of sentences in the other language, where either group can be empty so as to allow insertions and deletions.

A corpus is a large and structured set of texts. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe [5]. This is the basic training corpus used to train the alignment template Language Model. Aligning a corpus means making each translation unit of the source corpus correspond to an equivalent unit of the target corpus. In this case, the term "translation unit" covers both larger sequences such as chapters or paragraphs and shorter sequences such as sentences, syntagms or simply words [6].

The language model provides us with probabilities for string of words. We need to build a machine that assigns a probability P (e) to each English sentence e. It concerns with the probabilities of the occurrence of a word with its neighboring words to form a string of words. The language model use a bi-gram model which takes into account every two neighboring words for calculating P (e). In order to calculate the source language probabilities, a large amount of monolingual data is required, since of course the validity, usefulness or accuracy of the model will depend mainly on the size of the corpus.

3. Practical Implementation

3.1 Bilingual corpus

The bilingual corpus contains both an English sentence and a Bangla sentence for each aligned pair. This is the basic training corpus used to train the alignment template language model. The translation model uses both the English and Bangla sentences to estimate the translation probability of each Bangla word. The language model uses only the English sentences to set the grammatical structure of the expected English sentence as output. **Fig. 1** shows a sample bilingual corpus.

I eat rice. আমি ভাত খাই। I eat rice. আমি খাই ভাত।
I go home. আমি বাড়ি যাই। I go home. আমি যাই বাড়ি।
he eats. সে খায়। he goes home. সে যায় বাড়ি।
karim goes home. করিম যায় বাড়ি। he writes.
সে লেখে। he is poor. সে হয় গরিব। rahim goes home.
রহিম যায় বাড়ি। he is an engineer. সে হয় একজন
প্রকৌশলী। he is an engineer. সে প্রকৌশলী হয় একজন।
I am an artist. আমি হই একজন শিল্পী। I am an artist.
আমি একজন শিল্পী হই।

Fig. 1 Sample Bilingual Corpus

3.2 Bilingual Dictionary

Dictionaries are the largest component of an MT system in terms of information it holds. In a bilingual dictionary various types of entries are possible. Normally “paper dictionaries” are collection of entries. That is, they are basically lists of words, with information about various properties of the word. But in this model, bilingual dictionary contains only two information of each Bangla word. According to the alignment in the bilingual corpus each Bangla word contains the translation probability of the connected English word. The formation of this bilingual dictionary is as,

I আমি 0.93456243. I বাড়ি 0.068763.
eat ভাত 0.0796715. eat খাই 0.87034.

Which represents that the translation probability of “I” is 0.93456243 and 0.068763 to be the translation of “আমি” and “বাড়ি” respectively and so on.

3.3 Training Procedure

Training the bilingual corpus by means of translation modeling is a matter of inducing the translation probability table. For given English word e, we pretend that all Bangla words connected to each English word are equally likely translations. For a given sentence pair, all alignments will therefore look equally likely as well. Also for language modeling each English sentence is traversed to form the

sentence linguistically. A Bangla sentence is taken as input from the user and for each Bangla word the highest translation probability is taken as the translation of that Bangla word.

4. Experimental Result

For implementing the language model, bi-gram model is considered. First the exact translation of each Bangla word is chosen from the bilingual dictionary and initially an English sentence is made. Then the probability of each English word is calculated to come first of the English sentence by training the bilingual

corpus. The exact word is chosen from among these probabilities for which the probability is highest. Then from the rest of the words the probability is calculated to come next and so on. And again form these probabilities the next word of the highest probability is chosen. In this way, the final English sentence is constructed which follows the training corpus. **Fig. 2** illustrates the snapshot of the implemented method.

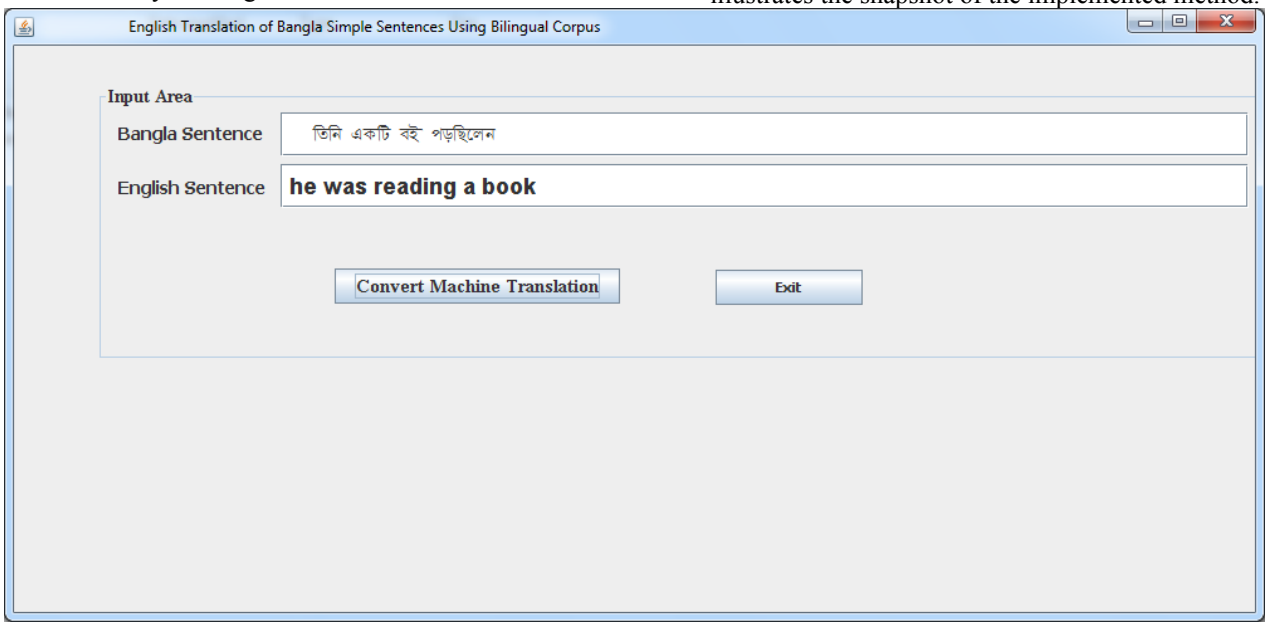

Fig. 2 Sample output of the implemented MT model of English translation of Bangla sentences

5. Conclusion

From a methodological point of view, combining a linguistic approach with a statistical approach makes it possible to fine-tune the alignment and enhance processing of bilingual corpora with a view to machine translation. The primary goal is to develop an MT system for English-Bangla integrating proper linguistic analysis and syntactic transfer into a data-driven approach. This paper focuses on the improvement of translation quality and the adaptability of the system to the user's requirements. We have tried to set up an appropriate model to adapt SMT system of English translation of Bangla simple sentences. The model can be extended to perform machine translation of Bangla complex and compound sentences to English in future.

References

[1] A. Trujillo, "Translation Engines: Techniques for Machine Translation", Springer-Verlag, London, (1992).

[2] K. Knight, "Automatic Knowledge Acquisition for Machine Translation", AI Magazine 18(4), (1997).

[3] M. M. Asaduzzaman and M. M. Ali, "Transfer Machine Translation – An Experience with Bangla English Machine Translation System", Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 265-270 (2003).

[4] T. Watanabe and E. Sumita, "Example-based Decoding for Statistical Machine Translation", ATR Spoken language, Translation research Laboratories 2-2-2, Keihanna Science City Kyoto 619-0288 Japan, (2001).

[5] M. M. Anwar, M. Z. Anwar and M. A. Bhuiyan, "Structural Analysis of Bangla Sentences for Machine Translation", Proceedings of International Conference On Computational Intelligence Applications, India, pp. 230–237 (2010).

[6] M. Guidère, "Toward Corpus-Based Machine Translation for Standard Arabic", Translation Journal, Vol. 6 No. 1, (2002).

Md. Musfique Anwar completed his B.Sc (Engg.) in Computer Science and Engineering from Dept. of CSE, Jahangirnagar University, Bangladesh in 2006. He is now a Lecturer in the Dept. of CSE, Jahangirnagar University, Savar, Dhaka, Bangladesh. His research interests include Natural Language Processing, Artificial Intelligence, Image Processing, Pattern Recognition, Software Engineering and so on.

Nasrin Sultana Shume completed her B.Sc (Engg.) in Computer Science and Engineering from Dept. of CSE, Jahangirnagar University, Bangladesh in 2006. She is now a Lecturer in the Dept. of CSE, Green University of Bangladesh, Mirpur, Dhaka, Bangladesh.Her research interests include Artificial Intelligence, Neural Networks, Image Processing, Pattern Recognition, Database and so on.

Md. Al-Amin Bhuiyan received his B.Sc (Hons) and M.Sc. in Applied Physics and Electronics from University of Dhaka, Dhaka, Bangladesh in 1987 and 1988, respectively. He got the Dr. Eng. Degree in Electrical Engineering from Osaka City University, Japan, in 2001. He has completed his Postdoctoral in the Intelligent Systems from National Informatics Institute, Japan. He is now a Professor in the Dept. of CSE, Jahangirnagar University, Savar, Dhaka, Bangladesh. His main research interests include Image Face Recognition, Cognitive Science, Image Processing, Computer Graphics, Pattern Recognition, Neural Networks, Human-machine Interface, Artificial Intelligence, Robotics and so on.