Assamese to English Statistical Machine Translation Integrated with a Transliteration Module

Pranjal Das
Dept. of Information Technology,
Gauhati University,
Assam, India

Kalyanee K. Baruah
Dept. of Information Technology,
Gauhati University,
Assam, India

ABSTRACT

In this paper, it is described how an Assamese sentence is translated to English using statistical machine translation. Statistical Machine Translation is the paradigm where translations from source to target language are based on statistical models. Moses is used as a platform for Statistical Machine Translation. GIZA++ is also used for word-alignment and IRSTLM for language model training. A Transliteration model is also integrated into the system to deal with out of vocabulary (OOV) words.

General Terms

Machine Translation, Natural Language Processing.

Keywords

Assamese, English, Statistical Machine Translation, Transliteration, Corpus.

1. INTRODUCTION

Machine Translation is the process of translating a source language to a target language. It allows us to obtain the best possible translation without any human assistance. Machine translation is a part of Natural Language Processing. The field of Machine Translation is as old as digital computer [4]. On a basic level, Machine Translation performs simple substitution of words in one natural language for words in another. But Machine Translation is not limited to this, as only substitution of words cannot produce a good translation. Different languages have different word structures and that is why recognition of whole phrases and their closest counterparts in the target language is required to produce a better translation [11]. There are many approaches of Machine Translation. One of the famous and newest approaches is the statistical based approach. Statistical Machine Translation (SMT) is an approach of machine translation where a target sentence is generated on the basis of a large parallel corpus. A large parallel data composing of two different languages is trained and thus deriving some statistical parameters to translate a source sentence to another. Google Translate is one of the famous online machine translation services based on Statistical Machine Translation. In this paper, we are trying to build an Assamese to English translation system. Assamese is a language spoken among the people of North-eastern part of India. It is the official language of the people of Assam, a state of India. Assamese is also spoken by a huge amount of people from Arunachal Pradesh and Nagaland. Apart from the states in India, people speaking Assamese are also found in Bhutan and Bangladesh [1]. The use of Assamese language in the web has increased tremendously in the recent years. We hope that, by building a machine translation system for Assamese to English will benefit the common people. In the field of natural language processing, Assamese language is still at its developing stage unlike other Indian languages like Tamil, Hindi, Bengali, etc. We would like to contribute in this field and bring the Assamese language closer to the rest of the world. This paper describes how we tried to build a translation system using the statistical approach of machine translation. Language Model, Translation Model and Decoder are three modules that are used to train the system and produce the desire output. Various other tools like Moses, Giza++ and IRSTLM are also used. Finally we have tried to integrate a Transliteration module into our system to work with Out of Vocabulary (OOV) words which are not translated with the Moses decoder.

2. RELATED WORKS

Machine Translation is not a new process in India. Many researches are going on Machine Translation. India is in touch with the study of machine translation from the mid 80s and early 90s. Since then, several institutes like Centre and Development of Advanced Computing (CDAC), Ministry of Communications and Information Technology, etc. are working in the field of machine translation [8]. Some of the machine translation projects in India are:

- *Anusaaraka:* Anusaaraka was started in 1995 at IIT Kanpur with the aim of translation from one Indian language to another. It gives translations from Telegu, Kannada, Bengali, Punjabi and Marathi to Hindi [8]. Anusaaraka uses the principle of Paninian Grammar (PG).
- *Anglabharati*: Anglabharati project was started in 1991 at IIT Kanpur for Machine aided translations from English to Indian languages (primarily Hindi). Anglabharati uses the rule based pseudo-interlingua approach for translation [8].
- *Shiva and Shakti Machine Translation*: Shiva and Shakti are two machine translation systems for English to Hindi. Shiva is an Example-based machine translation system while Shakti is a hybrid system composed of Rule-based and Corpus-based approaches [8].
- *MaTra*: MaTra is a human aided machine translation tool developed by CDAC, Mumbai for English to Hindi translations of news stories [8]. MaTra uses Rule-based Transfer approach for generating the translations.

Statistical Machine Translation is the most widely used approach. This is because building statistical based models are relatively quick and simple process. Some of the statistical based approaches of machine translation are:

- *Moses Statistical MT:* Moses was designed and developed by Philipp Koehn and Hieu Hoang at the University of Edinburgh. It allows us to automatically train translation models for any language pair. It just requires parallel texts that are used in training the system [7].

- *Prolog Statistical MT:* Prolog Statistical Machine Translation (PSMT) is a statistical machine translation program written in Prolog. PSMT is not much implemented for practical use [12].
- *Phramer Statistical MT:* Phramer is a phrase-based statistical machine translation system written in java [3]
- *EGYPT:* EGYPT is another statistical machine translation tool developed by the Statistical Machine Translation team [14].

Among these, only Moses Statistical Machine Translation provides complete set of training and decoding program as open source software.

3. METHODOLOGY

Statistical Machine Translation depends of huge amount of parallel text to generate translations from source to target language. A large corpora of parallel texts is trained and based on some statistical models it outputs the best possible translation. We have trained our system respectively with 4000, 6000 and 8000 Assamese-English parallel sentences. We have observed that a reasonable change in the translation is obtained when we increase the amount of corpus. More amount of data leads to more accuracy in the translations. Consider a given Assamese sentence *f*, the SMT system provides us with the probability $p(e \mid f)$ of an English sentence *e*. Now the Bayes' rule is applied to separately model the translation probability $p(f \mid e)$ which makes sure that the English generated is the appropriate translation of the source Assamese sentence and that of the English sentence $p(e)$, which guarantees fluent English output:

$$p(e \mid f) = \frac{p(e)p(f \mid e)}{p(f)} \qquad (1)$$

The probability of the Assamese sentence is dropped as it is a constant and would have zero effect on finding the target English sentence *ê*, which maximizes the equation $p(e)p(f \mid e)$:

$$\hat{e} = \arg\max_e p(e)p(f \mid e) \qquad (2)$$

3.1 Implemented System Architecture

The system architecture of our SMT system is shown in Figure 1. Our system is composed of three modules:

1) Language Model
2) Translation Model
3) Decoder

The Language Model gives the probability of the target language $p(e)$. The Translation Model $p(f \mid e)$ gives the probability of the source sentence with respect to the target sentence. The decoder then maximizes both the probabilities and outputs the most probable sentence.

3.1.1 Language Model

The purpose of the Language Model (LM) is to give the most fluent output by determining the probability of the target sentence. Here IRSTLM tool is used to develop the Language Model. IRSTLM estimates, represents and computes statistical language models. The Language Model is used for improving the target sentence in various ways. For example, it checks the word order. The sequence of words having more probability among a huge corpus is chosen. Consider the following Assamese sentence to be translated:

ৰাম এজন ভাল ল'ৰা (*Raāma ējana bhāla la' raā*)

The Language model checks for most probable target sentence in English. Consider the following probable target sentences:

$P(LM_1)$: *Ram is a good boy*

$P(LM_2)$: *good boy is a Ram*

After analyzing a large data of English text, the LM computes that $P(LM_1) > P(LM_2)$. Thus $P(LM_1)$ is fed to the decoder.

The Language Model also chooses the correct word when there is an ambiguity of choosing words. For e.g. consider an Assamese sentence:
ৰাহুল বুদ্ধৰ পুত্ৰ (*Raāhula bud'dhara putra*)

Here 'পুত্ৰ' can be: son, child, lad. So, the probable target sentences would be like these:

$P(LM_1)$: *Rahul is the son of Buddha*

$P(LM_2)$: *Rahul is the child of Buddha*

$P(LM_3)$: *Rahul is the lad of Buddha*

Analyzing a large corpus $P(LM_1)$ would be the most probable sentence.

The probability of a sentence $P(S)$ is broken down to probability of individual words $P(w)$ using the Markov Chain Rule:

$$P(S) = P(w_1, w_2, w_3...., w_n)$$

$$= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1w_2)...P(w_n \mid w_1w_2..w_{n-1}) \qquad (3)$$

The Language Model computes the probability of the target sentence using the *n-gram* model. An *n-gram* model simplifies the task of approximating the probability of a word given all the previous words. An *n-gram* of size 1 is known as a *unigram*; size 2 is a *bigram*, also known as *digram*; size 3 is a *trigram*; size 4 is a *four-gram* and size 5 or more is simply called *n-gram* [5]. Let us show how the *n-gram* is calculated using a *bigram* model.

.

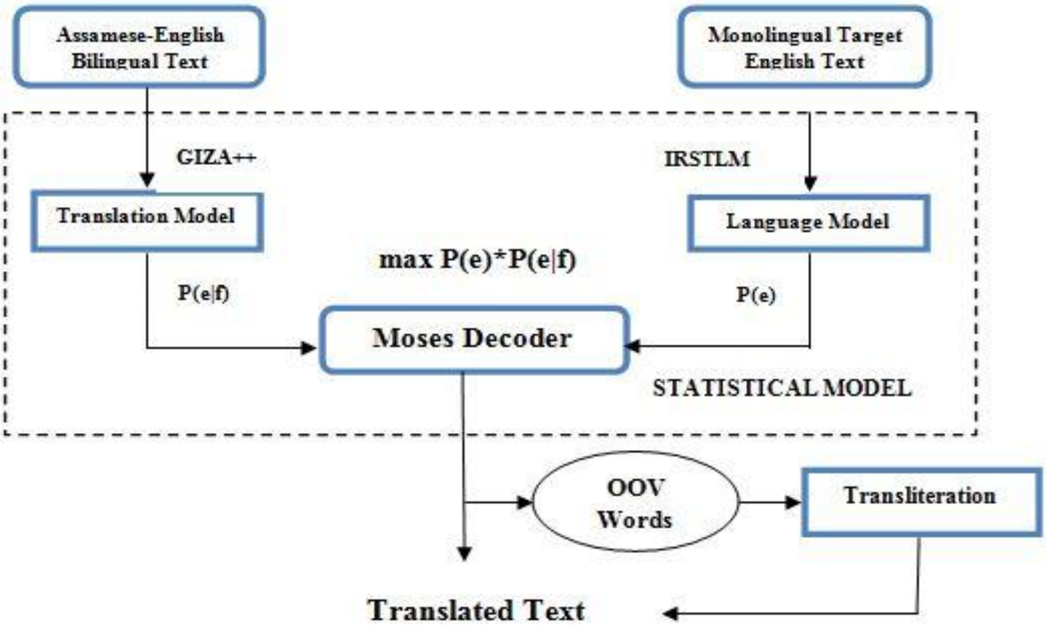

Fig 1: SMT System Architecture

The *bigram* probability is calculated using the following formula:

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})} \quad (4)$$

Suppose for a large amount of corpus we have got the following bigram probabilities:

Table 1. Bigram Probabilities

a good	.341	Sam is	.066	good girl	.11
a better	.132	Goa is	.003	good man	.02
a lovely	.0023	love is	.028	<start>Ram	.29
is the	.223	bad boy	.223	<start>Sam	.02
is with	.002	bad people	.021	<start>Goa	.27
is a	.43	good boy	.362	<start>Jaipur	.17
Ram is	.098	strong boy	.311	<start>India	.109

So, the probability of a sentence "*Ram is a good boy"* is:

P(*Ram is a good boy*)

= P(*Ram*|<*start*>)P(*is*|*Ram*)P(*a*|*is*)P(*good*|*a*)P(*boy*|*good*)

= (.29)*(.098)*(.43)*(.341)*(.362)

= 0.362

3.1.2 Translation Model

The Translation Model produces a target language sentence *e* from a source language sentence *f* by assigning probabilities to both source and target sentences. Giza++ is used to develop the translation model. It is tool used to align words in statistical machine translation systems [2]. The Translation Model uses phrases as well as single word models as fundamental units of translation. A word-based model is shown in Figure 2.

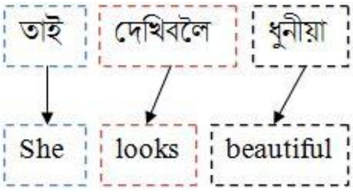

Fig 2: Word-Based Translation

For Assamese to English, majority of the sentences are translated using the phrase-based model. The Phrase-based translation is carried out in three steps [10]. First, it groups each Assamese word into phrases $f_1, f_2, ..., f_I$. Then it translates each Assamese phrase f_1^I to its corresponding English phrase e_1^J. Finally, it reorders each target English phrase using the Language Model. Figure 3 shows an example of a phrase-based translation.

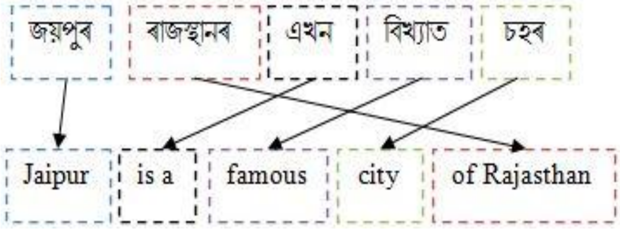

Fig 3: Phrase-Based Translation

The probability model of a phrase based translation depends on the translation probability $\phi(e_1^J \mid f_1^I)$ and the distortion probability *d*. The translation probability $\phi(e_1^J \mid f_1^I)$ generates English phrase e_1^J from the corresponding Assamese phrase f_1^I [9]. The distortion probability *d* is used to reorder the phrases of the target language i.e. English phrases. The distortion is parameterized by $d(start_j - end_{j-1})$ where $start_j$ is the start position of the English phrase generated by the i^{th} Assamese phrase f_i and end_{i-1} is the end position of the English phrase generated by $(i-1)^{th}$ Assamese phrase f_{i-1}. After applying Bayes' rule and decomposing the translation model, the translation probability now becomes:

$$p(f_1^I \mid e_1^J) = \prod_{i,j=1}^{I,J} \phi(f_i \mid e_j) d(start_j - end_{j-1} - 1) \qquad (5)$$

3.1.3 Decoder

The decoder computes the best probable translation $\hat{e}$ using the output of Language Model and Translation Model [9]. The output obtained from Language Model and Translation Model is fed to the decoder and the decoder maximizes this probability and gives the final translation output.

$$\hat{e} = \arg\max_{e \in English} \{p(f_1^I \mid e_1^J).p(e_1^J)\} \qquad (6)$$

Machine Translation decoders use best-first search based on heuristics [6]. Here we have used the Moses decoder to work with our system.

3.2 Transliteration

We have integrated an external module to our system, as we hope that this module will help to improve the quality and accuracy of our Machine Translation system. A Transliteration module is added to improve the translation by identifying out the Out of Vocabulary words (OOV) and transliterating those words to avoid the presence of Assamese words in the target English sentences. Some proper nouns which are not in our corpus are not translated. These proper nouns may be name of a person, place, etc. Transliteration is a process of mapping phonemes and graphemes of the source language to phonemes and graphemes of the target language. Transliteration systems find wide applications in Machine Translation systems and Cross Lingual Information Retrieval Systems (CLIR) [9]. The difficulties of MT systems arise due to huge number of OOV words which may be proper names, technical words and foreign words. Transliteration is very important when two different languages use different writing scripts, for e.g. Assamese-English, Chinese-English etc. We have used a perl script for the retrieval of the OOV words. In the script, each Assamese character is stored with its corresponding phonetic English character. For example,

ক → k, খ → kh, গ → g

Our Transliteration module is shown in Figure 4. But there are also some problems associated with our transliteration module. Till now, we are only concentrating on phonetic transcription. That is, for some words we may not get the correct spelling when it is transliterated. For example, the word 'কানাডা' is transliterated into 'kanada'. But the actual word should have been 'canada'.

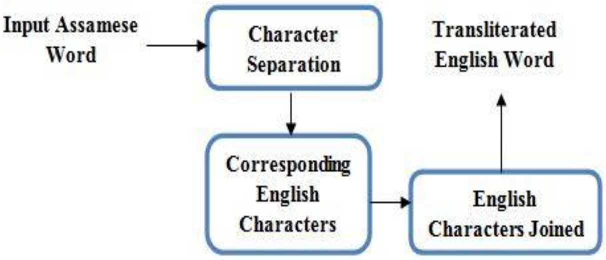
Input Assamese Word
Character Separation
Corresponding English Characters
English Characters Joined
Transliterated English Word

Fig 4: Transliteration Module

Also some proper nouns when transliterated cannot produce the desired output. The following table shows some of them:

Table 2. Non-Transliterated Proper Nouns

Assamese	English
প্ৰশান্ত মহাসাগৰ (prashaântâ mahaâsaâgârâ)	Pacific Ocean
ভাৰত(Bhaârâtâ)	India
চীন (chin)	China

So these problems should be dealt with some other methods, which we are trying to figure out. One solution is to increase the number of words in our corpus which will minimize the load of the transliteration module.

3.3 Data

Using a huge amount of parallel corpus is the fundamental need of a Statistical Machine Translation system. Parallel corpus for many languages is freely available on the internet. We have used Assamese-English parallel corpora of about 8000 sentences. Our data is based on travel and tourism in India. A sample parallel corpus is shown in the following table:

Table 3. Assamese-English Parallel Corpora

Assamese	English
কনক বৃন্দাবন হৈছে জয়পুৰৰ এখন জনপ্ৰিয় বনভোজ স্থান।	Kanak Vrindavan is a popular picnic spot in Jaipur.
জয়পুৰ মাৰ্বলৰ মূৰ্ত্তি, নীলা কলহ আৰু ৰাজস্থানী জোতাৰ বাবেও বিখ্যাত।	Jaipur is also famous for marble statues, blue pottery and the Rajasthani shoes.
অম্বৰ পেলেচটটো হৈছে মোগল আৰু হিন্দু স্থাপত্য বিদ্যাৰ আদৰ্শ উদাহৰণ।	The Amber Palace is a classic example of Mughal and Hindu architecture.

To prepare the data for training some preprocessing has to be done with the data. The preprocessing was as follows:

1) Tokenize the Assamese and English corpus.

2) Lowercase of the English corpus.

3) Cleaning the data, i.e. removing extra spaces, empty lines and lines that are too short or too long.

The data is then trained and a configuration file *moses.ini* is obtained which we use to run the decoder [13].

4. RESULTS AND EVALUATION

We have trained our system with different number of parallel sentences. We have seen that whenever we increase the quantity of our corpus, the quality of the translation is also improved. The system is trained respectively with 4000, 6000 and 8000 corpus. The following figure shows the differences obtained while training with the respective number of corpora.

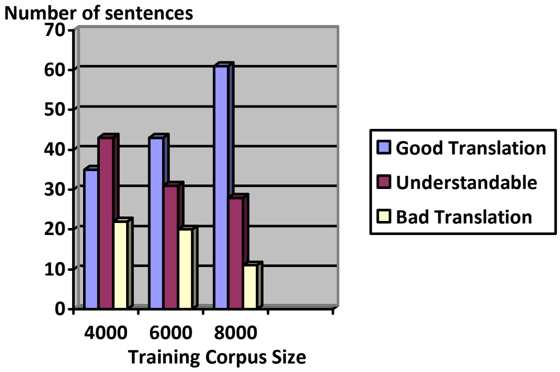

Fig 5: Test Results

From the above figure it is clear that increasing the corpus size results in much better translations. We have tested our system by translating 100 random sentences and found that some sentences are exactly translated as desired, some are understandable and some gives rough translations. The results obtained from the final training with 8000 sentences are shown in the following table:

Table 4. Translation Results

Source Assamese Sentence	Target English Sentence
দিল্লী ভাৰতৰ ৰাজধানী।	Delhi is the capital of India.
জয়পুৰ চহৰ ভ্ৰমণৰ উত্তম সময় হৈছে অক্টোবৰৰ পৰা মাৰ্চলৈ	Jaipur city of the best time to October to March
হায়দৰাবাদ এখন বিখ্যাত চহৰ দক্ষিণ ভাৰতৰ	Hyderabad is a famous city of Southern India
কানাডা এখন বিশাল দেশ।	kanada is a vast country
তাজমহল আগ্ৰাত অৱস্থিত ।	the Taj Mahal , Agra .
পৰিভ্ৰমণকাৰীৰে ভৰি থকা চিমলা এখন জনপ্ৰিয় ঠাই।	Shimla is a popular city with tourists.

We have evaluated the quality of our system using the BLEU (Bilingual Evaluation Understudy) metric. BLEU is an evaluation technique used in Machine Translation. It uses *n-gram* precision to compare a target translation with multiple reference translations [9]. The evaluation results obtained are shown in Table 5.

Table 5. Evaluation

Corpus	Source/Target	BLEU
Tourism Data	Assamese/English	11.32

The score is not amusing as the amount of corpus we have used is very small. For a corpus with millions of sentences, the BLEU score would relatively improve.

5. CONCLUSION AND FUTURE WORK

Statistical based systems require significant amount of corpus to achieve good translations. We have used a very small amount of data (about 8000 parallel sentences) to train our system. This statistic is very small compared to a good translation system which uses millions of sentences for training. There are not enough parallel corpora available between Assamese and English. But still, the results obtained by us are quite satisfactory. We have added a Transliteration module into the system which improves the translation quality and also the BLEU score. Among the many approaches of Machine Translation, Statistical Machine Translation is the most widely used. Many of the researches on Machine Translation in India are Statistical-based. In the future we would like to increase the amount of corpus as it would further improve our system. Also we would like to improve the transliteration module while dealing with OOV words.

6. ACKNOWLEDGMENTS

The authors are thankful to the *Department of Information Technology*, *Gauhati University* for providing us the Assamese-English parallel corpus and also the *Department of Natural Language Processing*, *Gauhati University* for their immense support.

7. REFERENCES

[1] Dr Shikhar Kr. Sarma et al, “Foundation and Structure of Developing an Assamese Wordnet”, In Proceedings of5th International Conference of the Global WordNet Association.

[2] F.J. Och and H.Ney, “Improved statistical alignment models”, In the Proceedings of ACL, 2000.

[3] Marian Olteanu et al, “Phramer: Open Source Statistical Phrase-Based Translator”, In the Proceedings of the Workshop of Statistical Machine Translation, June 2006, pp. 146-149.

[4] Peter F. Brown et al., “A Statistical Approach to Machine Translation” Computational Linguistics Volume 16, Number 2, June 1990, pp. 79-85.

[5] Philipp Koehn et al, “Statistical Phrase-Based Translation”, In the Proceedings of HLT-NAALC, May-June 2003, pp. 48-54.

[6] Philipp Koehn, “Pharaoh: A beam search decoder for phrase based statistical machine translation models”, In the proceedings of AMTA, 2004.

[7] Philipp Koehn et al, “Moses: Open Source Toolkit for Statistical Machine Translation”, In the Proceedings of the ACL, June 2007, pp. 177-180.

[8] Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve, “Machine Translation System in Indian Perspectives”, Journal of Computer Science, Volume 6, Issue 10, pp. 1111-1116.

[9] Md. Zahurul Islam, “English to Bangla Statistical Machine Translation”, Master Thesis, Universitat des Saarlendes, August 2009.

[10] Philipp Koehn, “Noun Phrase Translation”, PhD Thesis, University of Southern California, 1993.

[11] “Machine Translation”, Available: http://en.wikipedia.org/wiki/Machine_translation.

[12] “PSMT”, Available: http://psmt.sourceforge.net/

[13] Statistical Machine Translation System User Manual and Code Guide”, Available: http://www.statmt.org//moses/manual/manual.pdf.

[14] “The EGYPT Statistical Machine Translation Toolkit”, Available: http://old-site.clsp.jhu.edu/ws99/projects/mt/toolkit/.