Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System

Anup Kumar Barman
Dept. of IT
Gauhati University
Guwahati, India
anupbarman.gu@gmail.com

Jumi Sarmah
Dept. of IT
Gauhati University
Guwahati, India
jumis884@gmail.com

Prof. Shikhar Kumar Sarma
Dept. of IT
Gauhati University
Guwahati, India
sks001@gmail.com

Abstract

Machine Translation is a task to translate the text from a source language to a target language in an automatic manner. Here, we describe a system that translate the English language to Assamese language text which is based on Phrase based statistical translation technique. To overcome the translation problem related with highly open word class like Proper Noun or the Out Of Vocabulary words we develop a transliteration system which is also embedded with our translation system. We enhance the translation output by replacing words with their most appropriate synonymous word for that particular context with the help of Assamese WordNet Synset. This Machine Translation system outcomes with a reasonable translation output when analyzed by linguist for Assamese language which is a less computationally aware language among the Indian languages.

1 Introduction

Machine Translation (MT) is a system that can automatically generate translation output from source language to target language. In country like India, the MT system are very much essential due to their language diversity. The highly increasing rate of digital information indicates the top level requirement of MT system so that every people irrespective to their language can acquire and utilize those information. There are basically three approaches to develop a MT system which are Rule based approach, Statistical approach and some Hybrid approaches. In Rule based approach, various naturally occurring phenomenon on source language text are investigated and then extract them as some rules and analyze them to fit to embed for generating target language text. By using some parser from the source text they produce some intermediate representation and then the target language text is generated from the intermediate representation. In Statistical approach, to train up various parameters large number of parallel corpus are required. A Statistical approach derives with a better result when the size of the corpus is increased. Here, the best translation are performed based on some decision. Some Example based approaches are also used to translate the source language text to target language text. Due to the less amount of digitized documents as the parallel corpus some tries to correct their statistical translation output by using some Linguistic Rules. Such type of approaches are considered to be the hybrid approaches.

Assamese is a language spoken by the North-East people of India. It is one of the less computationally aware Indian language belonging to the Indo-Aryan Family. It is spoken by approximately 14 million people of the North-East region of India. Unfortunately, this language has less amount of computational linguistic resources. The linguistics researches are still in traditional mode. But, recently some researchers have made a deliberate attempt to study Assamese language from technological perspective. They have started to work in the development and enrichment of the language of Assamese in the field of Natural Language Processing (NLP). The Machine Translation task for Assamese language is very difficult as the amount of parallel corpus is very less.

WordNet, a lexical database was developed by Prof George A Miller for English language in 1985. Based on English WordNet structure, Indo-WordNet was being developed. Assamese WordNet was developed as a part of the Indo-WordNet project. Assamese WordNet, a lexical database was first developed in Gauhati University, 2009 by (Sarma et al., 2010). Assamese

WordNet comprises of contents that are linked to both English and Hindi WordNet. A combination of dictionary and thesaurus, Assamese WordNet comprises of four major components. They are ID which act as a primary key for identifying any synset in WordNet, CAT indicates the Parts Of Speech category, SYNSET lists the synonymous words in a most used frequency order and GLOSS describes the concept of any synset. GLOSS consist of Text-Definition and Example-Sentence. Text Definition contains concepts denoted by synset and Example shows the use of any synset entry. There are various semantic relation that occur between synsets in WordNet. They are Hypernymy-Hyponymy(IS-A/Kind of), Entailment-Troponymy (Manner-of for verbs), Meronymy-Holonymy (HAS-A/ PART-WHOLE). Synset, the basic building block of WordNet can explore the semantically related terms. For instance these words খাৰু (kharu: Bangles), কংকণ (kankan: Bangles), কঙ্কণ (kangkan: Bangles) describes the same concept হাতত পিন্ধা এবিধ গহনা (haatat pindhaa ebidh gahanaa:*A hand wearing ornament*).This structure of WordNet helps in automatic text analysis and various artificial intelligence applications as a combination of dictionary and thesaurus. Assamese WordNet has been used for a number of different purposes in text analysis such as Automatic document classification (Sarmah et al., 2012), Automatic text summarization (Kalita et al., 2012) etc. Here, we tried to use the Assamese WordNet basically the synsets for fine tuning the translated output by replacing words with their most appropriate synonymous word for that particular sentence.

This paper presents a MT system for English-Assamese which is based on Statistical Phrase based translation approach. Here we first developed a MOSES based translation system which we consider as the baseline translation system. For linguistically open class Proper Noun or some other Out of vocabulary words we implemented a MOSES based transliteration system which transliterate the English word to Assamese word in Character level. Then we embed this transliteration system with our Base line Translation system. The output of the new system was enhanced by mapping the words with Assamese WordNet synset so that we can put the most appropriate synonymous word for that particular sentence. This will give us a more relevant translation output when reviewed by some linguistic persons.

This paper further continues with a description of Previous Notable Work done while implementing a MT system for other Indian Languages in Section2, Section3 portrays our methodology to implement a English-Assamese MT system . This section starts with a description of tools used in implementing our system, an explanation of our English-Assamese parallel corpus, a system architecture where it gives us a overview of our baseline translation system, an elaboration of our transliteration system and the process of enhancing the translation output through mapping with the Assamese WordNet synsets. Section4 analyzes the result of our system. Finally conclusions are drawn in section5.

2 Related Study

Though spoken by major population of North-East India, Assamese language is still behind in computational perspective, basically processing the MT system. No work on MT system was researched or developed for Indian language like Assamese till date. To develop any MT system for a specific language requires collaboration among computer researchers, linguistics and expert manual translators. Although in languages like English or some other foreign language, the MT task is very well processed, the Machine Translation in Indian languages is still an open problem. MT in Indian languages was developed using various approaches.

(Devi et al., 2010; Goyal and Lehal, 2009) used Direct Machine Translation Systems for languages Hindi and Punjabi respectively. Another one MT system was developed based on the Paninian Grammar (Goyal and Lehal, 2010) using this approach. The other approach found while developing an MT system for Indian Languages is the Rule based approach. In rule based approach Transfer based machine translation is used in (Saha, 2005; Bandyopadhyay S, 2000) where there are three modules - analyzed, transfer and generation module. One another Transfer based MT system was developed at Resource Centre for Indian Language Technology solution (Dwivedi and Sukhadeve, 2010) to translate English to Canada Text. Pseudo Interlingua approach of Rule Based was used to develop Anglabharti MT system for translating English to Indian languages. To reduce the human labour than the Rule based approach a Corpus-based MT approach was used. In statistical approach, the open source software like GIZA++ can be used to align the parallel corpus and then the aligned corpus is processed to generate the Phrase based Translation model. Tool like SRILM may be used to generate the statistical language model. The Phrase based MOSES decoder can be used to translate the sentences after finding the translation model and language model. Statistical MT system for English-Hindi (Ahsan et al., 2010) and English-Malayalam was developed by using English-Hindi and English-Malayalam parallel corpus. Some Example based MT system was developed where the hypothesis is that a translation will be considered as most appropriate if it was occurred previously. Anubharti is an example based MT system which was developed by IIT kanpur (Sinha, 2004) where some grammatical analysis was also performed to reduce the size of the parallel corpus.

3 Our Approach

This section describes verious software tools used for developing our proposed Machine Translation system, portrays our system architecture and description of each modules of our system.

3.1 Used Tools

In order to develop an English -Assamese Machine Translation System we used various open source software tools. The phrase based machine translation system MOSES (Koehn et al., 2007) is used to perform the translation task. Through this statistical machine translation tool we train up a translation module by using English-Assamese Parallel corpus. MOSES implies an efficient search algorithm called beam-search which can quickly find the highest probability translation from huge numbers of choices. Another statistical machine translation toolkit GIZA++ (Och and Ney, 2003) was used to align our parallel corpus. For alignment task, GIZA++ is used to train IBM models 1-5 and HMM word alignment model. To generate the word classes which is necessary for training the aligned models this machine translation toolkit uses mkcls tool. A bilingual dictionary can be produced from that parallel corpus using GIZA++. We use SRILM toolkit to develop a statistical language model which has been under development in the SRI Speech Technology and Research Laboratory since 1995. In SRILM (Stolke, 2002) a set of C++ class libraries are available to implement a language model. To accomplish some standard task like training a language model or testing on data there are also a set of executable programs and some auxiliary scripts which are built on top of these class libraries. We run all these toolkits on LINUX platform.

3.2 System Architecture

In our English-Assamese MT system, we integrated the baseline statistical MT model with a statistical transliteration model. The transliteration model helps us to improve the translation output by providing the transliteration basically for Proper Noun or some other Out of vocabulary(OOV) words. Mapping the translated output with WordNet synset gives us more suitable synonymous word for that specific sentence. Below given a architecture diagram for our MT system.

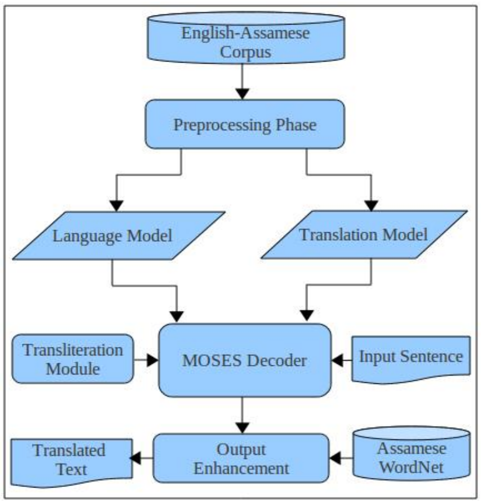

Figure 1: System Diagram

3.3 Data

Since digitized documents for Assamese language are very less in amount so to collect the parallel corpus for MT task was very difficult. For our approach, we prepare a parallel corpus of English-Assamese at Gauhati University NLP lab. This parallel-corpus contains data basically of tourism domain. In our parallel-corpus there were 100 files for each English and Assamese language. This parallel-corpus contains a collection of 14,371 English sentences with their respective translations in Assamese language. The corpus contains 326804 and 267224 words for English and Assamese language respectively. To fit this parallel-corpus in our translation model we follow the below pre-processing steps-

File Format Conversion:- We converted the file format of each file from UTF-16 to UTF-8 and we convert the line encoding from Windows to Unix/Linux.

Sentence Extraction:- Sentences were extracted from XML mark-up text.

Corpus Cleaning:- We clean-up the whole parallel-corpus by removing all unwanted characters, junk values and blank spaces etc.

Sentence Alignment:- We align each English sentence with respective Assamese sentences.

3.4 Baseline Translation System

To set-up our baseline translation system for English-Assamese we use the English-Assamese parallel corpus. By using the GIZA++ toolkit, we convert this sentence aligned parallel corpus to word aligned corpus and then processed them to fit in our phrase based translation model. We use the SRILM tool to develop our statistical language model from our corpus. For phrase based translation, we use MOSES decoder after getting the translation model and language model. To make this MOSES system learn we use this English-Assamese parallel corpus. Sample output of our Baseline Translation system is given below:

SNo.	English Text	Assamese Text
1.	Tiger is India's national animal	ব্যাঘ্ৰ ভাৰতৰ ৰাষ্ট্ৰীয় জন্তু
2.	Tiger is a violent animal	ব্যাঘ্ৰ হিংসুক জন্তু
3.	Mumbai is an ancient city	Mumbai আদিম চহৰ
4.	Assam is a beautiful state	Assam ধুনীয়া ৰাজ্য
5.	Times of India	সময় ভাৰত

Table 1: Output of Baseline Translation

3.5 Transliteration System

Since Assamese is a less computationally aware language, the parallel corpus is very less in size. Moreover, for some linguistically open word class like Proper Noun are very less available in the parallel corpus. The statistical MT system, acquires knowledge from the trained English-Assamese parallel corpus. From the above Table 1 which shows the results of our baseline translation system, we found that some words which are translated as source input word is. Those non-translated words are not found in the trained parallel corpus of English-Assamese. To overcome the translation problem basically related with Proper Noun word we develop a Statistical transliteration system. But, for other Out of vocabulary words we cannot implement the transliteration system since transliteration cannot represent the concept of those word in target language. To implement our transliteration system, we collect nearly 0.1 million unique Proper Noun in English and we transliterate them to Assamese. For transliteration, we consider each Proper Noun as a sequence of characters separated by a space. Then we create the language model by using SRILM tool. For alignment purpose , we use the same GIZA++ tool. Then we train up the MOSES decoder by using the Name Entity parallel corpus for English Assamese. We take the best output from n numbers of output from our statistical transliteration system. Output are also generated with a space in between each character. Finally, we combine this characters to get our transliterated output. Following Table 2 shows the sample result of our statistical transliteration system.

SNo.	Input Term	Transliterated Term
1.	Mumbai	মুম্বাই
2.	Assam	অসম
3.	Times of India	টাইমচ্ অফ ইণ্ডিয়া
4.	Rajasthan	ৰাজস্থান
5.	Brahmaputra	ব্ৰহ্মপুত্ৰ

Table 2: Output of Transliteration System

3.6 Combined System

A combined system was formed by combining the statistical transliteration with the baseline translation system. The statistical transliteration system is only for Proper Noun. We use one Named Entity dictionary comprising 1 lakh English Named Entities to recognize the Named Entities. The transliterated form was XML marked up. These XML files later were provided as an external knowledge to MOSES decoder for decoding. Combined system gives us the output provided by the baseline system with the Transliterated System. Following Table 3 shows the result of our combined system.

SNo.	English Text	Assamese Text
1.	Tiger is India's national animal	ব্যাঘ্ৰ ভাৰতৰ ৰাষ্ট্ৰীয় জন্তু
2.	Tiger is a violent animal	ব্যাঘ্ৰ হিংসুক জন্তু
3.	Mumbai is an ancient city	মুম্বাই আদিম চহৰ
4.	Assam is a beautiful state	অসম ধুনীয়া ৰাজ্য
5.	Times of India	টাইমচ্ অফ ইণ্ডিয়া

Table 3: Output of Combined System

SNo.	English Text	Assamese Text
1.	Tiger is India's national animal	বাঘ ভাৰতৰ ৰাষ্ট্ৰীয় জন্তু
2.	Tiger is a violent animal	ব্যাঘ্ৰ হিংস্ৰ জন্তু
3.	Mumbai is an ancient city	মুম্বাই প্ৰাচীন চহৰ
4.	Assam is a beautiful state	অসম ধুনীয়া ৰাজ্য
5.	Times of India	টাইমচ্ অফ ইণ্ডিয়া

Table 4: A sample of Final Output

3.7 Enhancement Using WordNet

The representation of a single concept using various words (synonymous words) in one language made influence in MT task. All synonymous words are not equally appropriate for all sentences in terms of their context. In a statistical MT system, for a source language term the most weighted target language term is always selected for every sentence containing that term without considering the appropriateness of that sentence. But, in natural language one individual concept is represented by using various synonymous term in various sentences. To overcome this statistical MT problem, we take help of the lexical resource Assamese WordNet where the set of synonymous terms to represent each concept are available. We enhance the output of our Combined System by selecting the appropriate synonymous term for that sentence through mapping each term to their respective WordNet synset. Selection of the appropriate synonymous terms in context of various sentences was done by checking manually through some Linguist fellows. Then we replace each term by the using the selected synonymous term so that our statistical MT output becomes more relevant in terms of Assamese language. Following Table 4 shows the final translation output after enhancement of the output produced by the combined system using Assamese WordNet.

4 Result Analysis

To evaluate our system performance, we take 500 English sentences for testing our statistical MT system. In the above tables, we show a sample output of each modules. Table 1 shows the baseline system's output where some of the Proper Nouns like India was translated to ভাৰত(Bharat:*India*) and some others remain the same like Mumbai and Assam. In Table 2 we show a sample output of our Transliterated system. The statistical transliterated system outcomes with a state-of-art accuracy. Then we mixed the Statistical Baseline System and Transliteration system to produce a combined system and a sample output of the system is shown in Table 3. Here the translation problem related with Proper Noun was solved with the proper transliteration form of them. As shown in Table 3 the term Assam, Mumbai, Times Of India was transliterated to অসম(*assam*),মুম্বাই(*Mumbai*), টাইমচ্ অফ ইণ্ডিয়া(*Times of India*) respectively which were not correct in Table1. Our combined system translation output was more or less correct but there are several words which are not appropriate for that specific sentence. To handle that type of inappropriateness problem we enhance our combined system output by mapping each term to their respective synset i.e, if a synset s have n entries for a word w in a sentence st then we have n number of possibilities to replace that particular word. Now we discover all possible such sentences which was later judged by the linguist to determine the most appropriate one. As in Table 3, we have seen that for the English sentence -Mumbai is an ancient city, the output is মুম্বাই আদিম চহৰ (*mumbai aadim sahar*)but the term আদিম(*aadim*)is not relevant to that particular sentence. In Assamese WordNet, there is a synset পুৰণি, পুৰণা, প্ৰাচীন, পুৰাতন, পুৰাকালীন, প্ৰাক্-কালীন, আদিম. Id 1661 where including this word আদিম there are seven synonymous words. Among these, the term পুৰণি(*puroni*) is the most appropriate as per linguist judgement. In our 1st and 2nd example sentences the words like ব্যাঘ্ৰ and হিংসুক are replaced with the most appropriate synset terms বাঘ and হিংস্ৰ respectively. In this way, we found the enhance result of our combined system's translation output which is depicted in Table 4.

5 Conclusions

We sum up our translation task by developing a English-Assamese translation system which is a combination of a statical phrase based translation and a statistical transliteration system and later the output was fine tuned by using Assamese WordNet. This introducing English-Assamese Statistical MT system gains a satisfactory output. The more strength of the parallel-corpus better the result of the Statistical MT system. Assamese is a less computationally aware language so the strength of English-Assamese parallel-corpus is weak. A statistical transliteration module is also embedded with our translation system so that we can overcome the translation problem related with Proper Nouns and some out-of vocabulary words. The transliteration module with a good accuracy helped in improving our translation system's performance. Also the lexical resource Assamese WordNet gives us a significant improvement in our translation output by providing the most accurate synonymous word for a specific context. A state-of-art translation results are generated by our Statistical MT system. As this is a first approach towards developing a English-Assamese Machine Translation system this will contribute significantly towards Assamese Natural Language Processing.

References

Arafat Ahsan, Prasanth Kolachina, Sudheer Kolachina, Dipti Misra Sharma and Rajeev Sangal. 2010. *Coupling Statistical Machine Translation with Rule-based Transfer and Generation*. In Proceedings of AMTA- The Ninth Conference of the Association for Machine Translation in the Americas. Denver, Colorado,.2010.

Sivaji Bandyopadhyay. 2000. *ANUBAAD - The Translator from English to Indian Languages*. In Proceedings of VIIth State Science and Technology Congress, Calcutta, India, 2000.

Sobha Lalitha Devi, Pravin Pralayankar, S. Maneka, T. Bakiyavathi, R.V.S. Ram and V. Kavitha. 2010. *Verb Transfer in a Tamil to Hindi Machine Translation System*. In Proccedings of International Conference of Asian Language Processing, 2010, Harbin, China, pp. 261-264.

Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve. 2010. *Machine translation systems in Indian perspective*. Journal of computer science, pp.1082-1087,2010.

Vishal Goyal and Gurpreet Singh Lehal. 2009. *Evaluation of Hindi to Punjabi Machine Translation System*. International Journal of Computer Science Issues, vol4 no1, 2009, pp. 36-39.

Vishal Goyal and Gurpreet Singh Lehal. 2010. *Web Based Hindi to Punjabi Machine Translation System*. Journal of Emerging Technologies in Web Intelligence, Vol.2, 2010, pp.148-151.

Chandan Kalita, Navanath Saharia and Utpal Sharma. 2012. *An Extractive Approach of Text Summarization of Assamese using WordNet*. In Proccedings of 6th International Global WordNet Conference (GWC 12) Japan, January 9-13 2012, pp. 149-154.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen,Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. In Annual Meeting of the Association for Computational Linguistics(ACL),2007.

Franz Josef Och and Hermann Ney. 2003. *A systematic comparison of various statistical alignment models*. Computational Linguistics, 29(1):19-51,2003.

Goutam Kumar Saha. 2005. *The EB-Anubad translator: A hybrid scheme*. Journal of Zhejiang University SCIENCE 2005, ISSN 1009-3095, pp.1047-1050.

Shikhar Kr. Sarma, Moromi Gogoi, Utpal Saikia and Rakesh Medhi 2010. *Foundation and structure of Developing Assamese WordNet*. In Proceedings of 5th International Conference of the Global WordNet Association.

Jumi Sarmah, Navanath Saharia and Shikhar Kr. Sarma. 2012. *A Novel Approach for classification of Document using Assamese WordNet*. In Proccedings of 6th International Global WordNet Conference (GWC 12) Japan, January 9-13 2012, pp. 324-329.

R. Mahesh K. Sinha. 2004. *An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures*. In Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004), November 17-19, 2004.

Andreas Stolke. 2002. *SRILM-an extensible language modeling toolkit*. In Proceedings of the ICSLP,2002.