Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 6; October-December, 2016, pp. 477-480
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Machine Translation for North Eastern Indian Languages: The Current Scenario

Aiusha V. Hujon

Department of Computer Science, St. Anthony's College
E-mail: avhujon@gmail.com

Abstract—*India being a country with many languages, and communication between the various regions are in Hindi or English. English language is still playing a major role in most government documents, official letters and other forms of communications among the people. There is a need to be able to translate English language to the regional languages. Although plenty of research and developments has been going on for decades to use machine translation from English to a particular Indian language or from one Indian language to another, the languages in the North Eastern region of India are still undergoing through a growing process of Initial development in the field of machine translation at present. This paper surveys the developments and work that has been done for the languages in the North Eastern region of India in the recent years. It also gives a brief description of the methods used for machine translation and the systems that has been developed using these tools.*

Keywords*: Machine translation, Natural language processing, MT systems, machine translation approach, Indian languages*

1. INTRODUCTION

India is a country with many languages. There are 22 languages in the Eight schedule which include Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu, Bodo, Santhali, Maithili and Dogri. Of these languages Manipuri, Assamese, Bodo are languages of the north eastern states of India. There are many other languages which have not been included in the Eight schedule like Mizo, Nagamese, Khasi, Garo and others. The following list shows the languages spoken in the North East states of India.

Assam-Assamese, Bodo

Tripura -Kokborok

Meghalaya-Khasi, Garo

Mizoram-Mizo

Manipur-Manipuri, Meitei

Nagaland-Nagamese (Tangkhul-naga)

Arunachal Pradesh-Tibetan, Adi, Bodo, Mikir, Monpa Nishi/Daa, Nock, Tanga, Wansho, Nefamese (Arunachalese)

Sikkim -Sikkimese (Bhutia), Lepcha and Nepali (the lingua franca of Sikkim)

A development of machine translation for the languages that has been included in the Eight schedule has been going on for more than a decade. Some of these are ANUSAARAKA developed by Prof. Rajeev Sangal and Team, IIT Kanpur in 1995, MANTRA developed by Dr. Hemant Darbari and Dr. Mahendra Kumar Pandev and Team, C-DAC, Bangalore in 1999, MATRA by Dr. Durgesh Rao and Team, C-DAC, Mumbai, in 2004, ANGLABHARTI developed by Prof. R.M.K. Sinha and Team, IIT Kanpur in 1991, ANUBHARTI by Prof. R.M.K. Sinha and Team in 2004 at IIT Kanpur, SHIVA And SHAKTI by Shiva and Shakti Machine Translation Team , IIIT Hyderabad and Institute of Science at Bangalore, ANUBAAD[7] by Dr. Sivaji Bandyopadhyay and team, Jadavpur University, Kolkata in 2004, SAMPARK by Sampark machine translation team, Consortium of Institutions in 2009. However, the development of machine translation for the north eastern languages of India has just started to progress.

2. APPROACHES IN MACHINE TRANSLATION

There are three paradigms in machine translation

- Rule based machine translation (RBMT)
- Statistical machine translation (SMT)
- Example based machine translation (EBMT)

Rule based machine translation (RBMT): RBMT is used based on the linguistic information of the pair of languages i.e., for the source and the target language. The linguistic information includes the morphology, syntax and semantic regularities of the source and target language. Rule based machine translation is divided into three approaches

- Direct
- Interlingua
- Transfer based

Direct approach: This method directly translates the source language into the target language aided by a dictionary and the performance depends very much on the quality and quantity of the source and target dictionaries.

Interlingua approach: This method involves two stages. In the first stage the source language is first converted into an intermediate form called Interlingua. The Interlingua form is a language independent representation. In the second stage the Interlingua is converted into the target language.

Transfer approach: This method involves three phases. The first phase is the Analysis, where the source language text is converted into an abstract source language oriented representation using a parser. The second phase is the Transfer, where the source language representations are converted into corresponding target language representation. Finally in the third phase, the Synthesis, morphological analysis is performed on the target language representation and the final target text is created.

Statistical machine translation (SMT): Statistical machine translation uses huge parallel text corpora of both the source and target language. Using statistical parameters it translates the source language to the target language. Most SMT uses either word based translation or phrase based translation.

Example based machine translation (EBMT): Example based machine translation also uses parallel corpora of the pair of languages. It reuses the example translated sentences to translate new sentences. It involves three stages, Matching, Alignment and Recombination.

3. EXISTING TRANSLATION SYSTEMS FOR INDIAN LANGUAGES

A few of the renowned Machine Translation Systems in India are mention below:

ANUBHARTI: It uses example based approach to translate Hindi in to English and it was developed at IIT Kanpur in 2004.

ANGLABHARTI: It uses pattern directed approach to translate English into Indian languages. It is develop at IIT Kanpur. Many other organizations have participated to translate from English to their respective regional language by adding another layer in the AnglaBharti system such as Anglamalayam, Anglamarathi etc.

ANGLABANGLA: Which translates English to Bangla develop at CDAC, Kolkata using the AnglaBharti technology.

MANTRA: Translates English to Hindi, Telegu, Gujarati develop at CDAC, Pune in 1995.

SAMPARK: A machine translation system among Indian languages. An experimental version is available online. It translates from Hindi to Bengali, Hindi to Urdu, Hindi to Malayalam, Hindi to Punjabi, and Hindi to Tamil. It can also translate form these languages to Hindi.

4. DEVELOPMENTS AND PROGRESS OF THE NORTH EAST INDIAN LANGUAGES IN MACHINE TRANSLATION

Some systems have been developed for two languages spoken in the north east India which has been included in the Eight schedule, i.e., Assamese and Manipuri. For some other languages research is still going on and very few of these works has been reported. The following discusses some of the systems and other works developed related to machine translation.

4.1 Developments for the Assamese language

English to Assamese: Assamese or 'ASAMIYA' is spoken originally in the state of Assam. A lot of research has been going on in translating English to Assamese and Assamese to English. One such translation has been designed using Statistical machine translation which translates from English to Assamese by Singh, Borgohain and Gohain, Dibrugarh University, Assam. It uses N-Grams for the Language model and phrase based Translation model [9]. The system is built with a corpus of 5000 sentences in bilingual text corpus of Assamese and English. It uses various statistical parameters to compute the Assamese text sequence which had the maximum probability value of being translated from the corresponding English language.

SMT to translate from Assamese to English: Assamese sentence is translated to English using statistical machine translation. The language model was built with the help of IRSTLM. The Translation model was built with the help of GIZA++ and Moses is used as a platform for Statistical Machine Translation[10]. The system is trained respectively with 4000, 6000 and 8000 corpus which gives a much better translation compared to the smaller size corpus. The system was also tested with BLEU and gives a score of 11.32 which is good enough for a small corpus size.

Part of speech tagger for Assamese: A Part of Speech tagging system was developed by Navanath Saharia et. Al. It uses Hidden Markov model to develop the POS tagger. A tag set of 172 tags was developed [14]. It uses a corpus of 10000 words and the system was tested and gives an accuracy of around 87%.

4.2 Developments for the Manipuri language

Part of speech tagging for Manipuri: A POS tagger for Manipur was developed by Singh, et al., here the POS tagger uses three dictionaries containing root words, prefixes and suffixes has been designed and implemented using the affix information irrespective of the context of the words. The inputs of 3784 Manipuri sentences of 10917 unique words as input to the tagger engine [1]. The POS tagger shows an accuracy of 69%.

Manipuri-English Example based machine translation: A Parallel corpus is used where phrase alignment is applied

where POS tagging, morphology analysis, NER and chunking are applied on the parallel corpus[2]. In this system evaluation was performed using BLEU and NIST which gives an accuracy of 0.137 BLEU and 3.361 NIST.

English - Manipuri SMT: A Manipuri based machine translation using morpho-syntactic and semantic information. Here the morphology and dependency relation was developed by Singh and Bandhopadhyay. In this system the important translation factors considered is the role of suffixes and dependency on the source side. The system was trained with 10350 sentences and 500[13]. After evaluation the system proves to performed better with shorter sentences than the larger sentences in terms of fluency and adequacy.

Morphological analyzer for Manipuri: A Manipuri Morphological analyzer, which can find out the word morphemes from the raw text, has been developed by Singh and Bandopadhyay[15]. It uses a Manipuri – English dictionary that stores the Manipuri root words and their associated information. This Morphological analyzer can handle five types of words. A word without any affix, word with a prefix, word with one or more suffix, compound word, semantic reduplicative words.

Another such system is the Morphological analyzer which has been developed by Choudhury, Sirajul Islam, et al [4].

Part of speech tagging for Manipuri: A Manipuri language rule based POS tagging using rule based approach was developed by Kh Raju Singha. It used the ILPOST framework; a total number of 97 tags are used for testing. It applied 25 rules in this system and gives an accuracy of 85% is obtained for 1000 words [5]. Another Manipuri POS tagging using Conditional random fields (CRF) and Support vector machine (SVM) was developed by Singh and Bandhopadhyay [3].

4.3 Developments for the Mizo language

Resource building and POS tagging for Mizo language: This work has been developed by Partha Pakray et al. A Mizo to English dictionary was developed and a 24 item POS tagset was generated to be used for the automatic POS tagger [2]. The dictionary also consists of a POS tag for each synonym. The dictionary was generated by a combination of automatic and manual techniques. It contains 26,407 entries.

4.4 Developments for the Kokborok language

Kokborok Morphological Analyzer: The development of a Kokborok Morphological Analyzer has been developed by Khumbar Debbarma, et al. The analyzer uses three dictionaries of morphemes, the root, prefix and suffix[6]. The root dictionary stores the related information of the corresponding roots. A stemmer algorithm was used for the root dictionary. The analyzer has been tested on 56732 Kokborok words and showed an accuracy of 80% on a manual check.

4.5 Developments for the Bodo, Khasi, Garo, Nagamese, and Nefamese languages

Machine translation for the languages such as Bodo, Khasi, Garo, Nagamese and Nefamese or Arunachalese still requires a lot of effort from the government and the research community. Some research works is going on for Bodo[12] and Khasi language[16], but to my knowledge there is no such systems or tools till date that has been developed for these languages.

5. CONCLUSION

Research work for the languages which are spoken in the North East region of India still needs a lot effort and collaboration between Private Agencies, Government and the Research Community to bring about a level of progress that is required for languages in India. This paper has highlighted some of the major breakthrough in machine translation on some languages which are spoken by a majority of speakers in India like Hindi, Telegu, Malayalam, Marathi and Bengali. It also discusses the developments achieved by some researchers for Assamese, Manipur and Kokborok. Most of the major achievements are for Assamese and Manipur. However, as per my knowledge, there is no report found on the developments achieved on the other languages like Khasi, Bodo, Garo, Nefamese and Nagamese till date.

REFERENCES

[1] Singh, T. D, Bandyopadhyay, S. "Morphology Driven Manipuri POS Tagger." *IJCNLP*, 2008.

[2] Pakray, P, Pal, A, Majumder, G, Gelbukh, A, Resource Building and Parts-of-Speech (POS) Tagging for Mizo Language, *14th Mexican International Conference on Artificial Intelligence, MICAI 2015. IEEE CS*, ISBN 978-1-5090-0323-5 , 2016, 3–7.

[3] Singh, T. D., Ekbal, A., Bandyopadhyay, S. (2008). Manipuri POS tagging using CRF and SVM: A language independent approach, *Proceeding of 6th International conference on Natural Language Processing (ICON-2008)* pp. 240-245.

[4] Choudhury, Islam, S., et al. "Morphological analyzer for manipuri: Design and implementation", *Asian Applied Computing Conference. Springer Berlin Heidelberg*, 2004.

[5] Singh, T. D., Bandyopadhyay. S., "Manipuri-English Example Based Machine Translation System", *International Journal of Computational Linguistics and Applications (IJCLA)*, ISSN (2010): 0976-0962.

[6] Dipankar, Debbarma,K., Patra,B.G.,Das, Bandyopadhyay,S., "Morphological Analyzer for Kokborok", *24th International Conference on Computational Linguistics*, 2012.

[7] Bandyopadhyay, S., "ANUBAAD-the translator from English to Indian languages", *VIIth State Science and Technology Congress*, Calcutta, India, 2000.

[8] Singh, T. D., Bandyopadhyay, S., "Manipuri-english bidirectional statistical machine translation systems using morphology and dependency relations", *Proceedings of the Fourth Workshop on Syntax and Structure in Statistical Translation*, 2010.

[9] Singh, Tiken, M., Borgohain, R., Gohain, S., "An English-Assamese Machine Translation System", *International Journal of Computer Applications*,93.4, 2014.

[10] Das, Pranjal, Baruah, K.K., "Assamese to English Statistical Machine Translation Integrated with a Transliteration Module", *International Journal of Computer Applications*, 100.5, 2014.

[11] Sarkar, Partha, Purkayastha, P.S., "Morphological Analyzer in the Development of Bilingual Dictionary (Kokborok-English)-An Analysis for Appropriate Method and Approach"

[12] Boro, Bhatima, S.K.R., "Corpus Building of Literary Lesser Rich Language-Bodo: Insights and Challenges", *24th International Conference on Computational Linguistics*, 2012.

[13] Sing, T.D., Bandyopadhyay, S., "Statistical machine translation of English–Manipuri using morpho-syntactic and semantic information", *Proceedings of the Association for Machine Translation in the Americas (AMTA 2010)*, 2010.

[14] Saharia, Navanath, et al. "Part of speech tagger for Assamese text", *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics*, 2009.

[15] Singh, T.D., Bandyopadhyay, S., "Manipuri morphological analyzer", *Proceedings of the Platinum Jubilee International Conference of LSI, Hyderabad*, India, 2005.

[16] Tham, M.J., "Design considerations for developing a parts-of-speech tagset for Khasi", *Emerging Trends and Applications in Computer Science (NCETACS), 201, 3rd National Conference on, IEEE*, 2012.