ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2017,
Vol. 10, No. (2):
Pgs. 429-437

A Review on Electronic Dictionary and Machine Translation System Developed in North-East India

SAIFUL ISLAM* and BIPUL SYAM PURKAYASTHA

Department of Computer Science, Assam University, Silchar, PIN-788011, Assam, India
Corresponding author e-mail: sislam.mca@gmail.com

http://dx.doi.org/10.13005/ojcst/10.02.25

(Received: May 04, 2017; Accepted: May 12, 2017)

ABSTRACT

Electronic Dictionary and Machine Translation system are both the most important language learning tools to achieve the knowledge about the known and unknown natural languages. The natural languages are the most important aspect in human life for communication. Therefore, these two tools are very important and frequently used in human daily life. The Electronic Dictionary (E-dictionary) and Machine Translation (MT) systems are specially very helpful for students, research scholars, teachers, travellers and businessman. The E-dictionary and MT are very important applications and research tasks in Natural Language Processing (NLP). The demand of research task in E-dictionary and MT system are growing in the world as well as in India. North-East (NE) is a very popular and multilingual region of India. Even then, a small number of E-dictionary and MT system have been developed for NE languages. Through this paper, we want to elaborate about the importance, approaches and features of E-dictionary and MT system. This paper also tries to review about the existing E-dictionary and MT system which are developed for NE languages in NE India.

Keywords: Electronic Dictionary, Machine Translation, NE languages, NLP.

INTRODUCTION

Natural language is the most important communication media for all human beings. Natural language processing is one of the interdisciplinary research areas in Computer Science, Computational Linguistic and Artificial Intelligence. It is a very attractive method for interactions between computers and natural languages. The purpose of NLP is to design and develop software that will analyze, understand, and produce speech or text of natural languages[21]. At present, it is a very demand able research area in computer science that explores how computers can be used to understand and manipulate text or speech of natural languages to do useful things. A

huge amount of different applications of NLP has been developed in the world as well as in India. The most commonly used applications of NLP are Electronic Dictionary, Machine Translation, Machine Transliteration, Information Retrieval, Morphological Segmentation, Named Entity Recognition, Optical Character Recognition, Part of Speech Tagging, Parsing, Question Answering, Speech Processing, Speech Recognition, Speech Segmentation, Spelling Checker, Wordnet and Word Sense Disambiguation.

The North-East is one of the most linguistically and ethnically diverse regions of India. The NE region is situated between the two great traditions of the Indic Asia and the Mongoloid Asia. The North-East India (NEI) consists of eight states which are Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura. Each of the states has own culture, language and tradition[1,2]. The people of NEI belong to different communities and each community has different languages. Therefore, NE is also known as multilingual and multicultural region of India. In this situation, E-dictionary and MT system are very important for communication among people of the different communities. A small number of E-dictionary and MT system have been developed for NE languages in India.

In this paper, we discuss the two most important applications of NLP which are developed for NE languages in NE India, namely E-Dictionary and Machine Translation. The E-dictionary and MT systems are tremendous potential and frequently used in human daily life.

North-East Languages

The natural languages act as a bridge amongst the people and help in creating a bond among their cultures. There are about 220 spoken languages in NEI and these languages are divided into mainly three language families, namely Indo-Aryan, Sino-Tibetan and Austro-Asiatic [2]. The languages of the Indo-Aryan family are Assamese, Bengali and Nepali. The languages of the Austro-Asiatic family are Khasi and Jaintia. The languages of the Sino-Tibetan family are Abor, Adi, Angami, Ao, Apatani, Biate, Bodo, Chakm, Deuri, Garo, Hmar, Hrusso, Karbi, Konyak, Lotha, Meiteilon, Misimi, Mishing, Mizo, Nocte, Nyishi, Rabha, Sema, Tanee, Tangkhul, Tiwa etc. The most commonly speaking languages in NE region are Assamese, Bengali, English, Hindi, Manipuri and Nepali. The Assamese, Bengali, Hindi, Manipuri and Nepali are also five of the 22 recognized languages and English is the associate official language of India [1]. The different languages of the eight states of NEI are shown in the Table 1.

Electronic Dictionary

A dictionary is a very important language learning book which contains enormous words of one or more particular languages and the words are arranged alphabetically with their meaning, part-of-speech (POS), synonyms, phonetics and examples. It is a very helpful tool for students, research scholars, teachers, travellers and other people to improve their knowledge about the various languages. The dictionary is divided basically into two broad categories, namely Paper dictionary and Electronic dictionary. The Paper dictionary is also known as hard or printed dictionary and Electronic

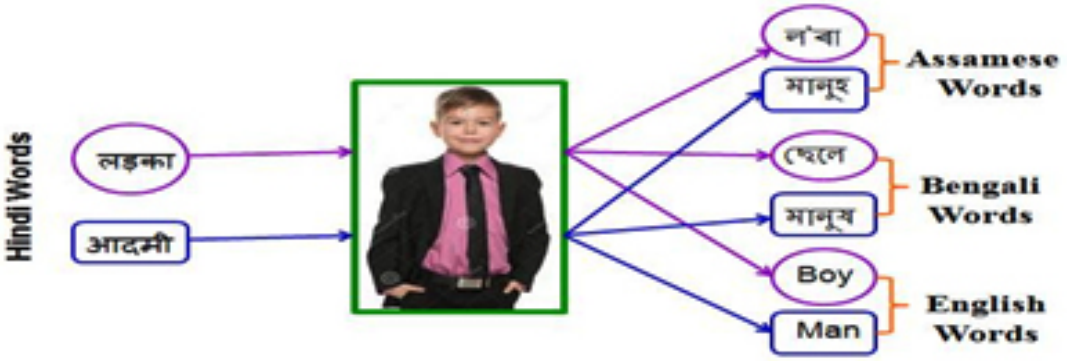

Fig. 1: Example of H-ABE E-dictionary

dictionary is also known as soft dictionary. The Electronic dictionary (E-dictionary) is one kind of dictionary whose data exist in digital form and can be used through different media. The E-dictionary is the most powerful tool for human to learn about the specific natural languages from anywhere place and any time using computers, smart phone and PDA. It is very convenient to use and tremendously better than paper dictionary. The E-dictionary is an important application of NLP and can be used to implement the other research tasks in NLP like machine translation, wordnet, etc.

Different types of E-dictionary

Generally, the E-dictionary can be divided into two types as discuss as follows:

Online E-dictionary

The online E-dictionary can be accessed in digital form through the Internet using web browsers from anywhere place in the world. Therefore, this dictionary is also known as Internet dictionary. It is very convenient to use if there is an Internet connection and large numbers of user can be accessed simultaneously on online.

Offline E-dictionary

The offline E- dictionary can be accessed through digital computer, personal data assistant and smart mobile phone. This dictionary can be carried and keep backup using compact disk, digital versatile disk, hard disk and pen drive. Therefore, this dictionary is also known as a

Table 1: Different languages of the states of NEI

Name of the States	Official languages	Major Spoken Languages
Arunachal Pradesh	English	Adi, Assamese, Bengali, Hindi, Mishing, Monpa, Nepali, Nyishi, Tangsa, Wancho
Assam	Assamese,	Assamese, Bengali, Bishnupriya Manipuri, Bodo, Dimasa, Hindi, Karbi, Mishing, Nepali, Rabha
Manipur	Bengali, Bodo, English English, Meiteilon (Manipuri)	Bengali, Hindi, Kabui, Kuki, Hmar, Manipuri, Nepali, Paite, Tangkhul, Thadou-Kuki
Meghalaya	English, Garo, Khasi	Assamese, Bengali, Garo, Hajong, Hindi, Khasi, Koch, Nepali, Rabha
Mizoram	English, Mizo	Bengali, Chakma, Hindi, Hmar, Lakher (Mara), Mizo, Nepali, Pawi, Paite, Tripuri
Nagaland	English	Angami, Ao, Assamese, Bengali, Chakru, Chang, Hindi, Garo, Kheza, Konyak, Kuki, Lotha, Nagamese, Phom, Rengma, Sangtam, Sumi, Yimchungre
Sikkim	English, Nepali	Hindi, Lepcha, Limbu, Nepali, Rai, Sherpa, Sikkimese (Bhutia), Tamang
Tripura	Bengali, English,	
Kokborok	Bengali, Bishnupriya Manipuri, Garo, Halam, Hindi, Kokborok (Tripuri), Manipuri, Mogh	

portable digital dictionary. User can download this type of dictionary through the Internet in his/her own computer or smart phone and can be accessed without the Internet.

According to the languages involve, the E-dictionary can be divided into three categories as below[5]:

Monolingual E-dictionary

The monolingual E-dictionary is one kind of dictionary where users can look up the meaning of word and other related information of the word like POS, synonyms and examples from one natural language to itself. For example, Assamese-Assamese, English-English, Manipuri-Manipuri and so on are monolingual E-dictionary.

Bilingual E-dictionary

The bilingual E-dictionary is one kind of dictionary where users can look up the meaning of word and other related information of the word like POS, synonyms and examples from a source natural language to a target natural language. For example, Assamese to English, English to Bengali and so on are bilingual E-dictionary.

Multilingual E-dictionary

The multilingual E-dictionary is one kind of dictionary where users can look up the meaning of word and other related information of the word like POS, synonyms and examples from one natural language to two or more natural languages. For example, Assamese to English and Bengali, English-Manipuri and Nepali, etc. are multilingual E-dictionary.

Different Techniques of E-dictionary

There are many word search techniques are available to develop the E-dictionary. Different developers use different word search techniques to look up (search) the words from the E-dictionary on both online and offline. The most commonly used word search techniques to develop the E-dictionary are as follows[22]:

- Sequential Search Technique
- £ Index Based Search Technique
- £ Binary Search Technique
- £ Incremental Search Technique

Wildcard Search Technique

Example of an Electronic Dictionary

Let us consider, Hindi to Assamese, Bengali and English (H-ABE) a multilingual E-dictionary. In this dictionary, the user can look up the meaning of Hindi word into corresponding Assamese, Bengali and English words as well as other related information of the given word like POS, synonyms and examples. Two examples of Hindi words and their corresponding meanings in Assamese, Bengali and English words are shown in figure 1.

E-dictionary Developed in NEI

In this section, we discuss about the existing electronic dictionaries which are developed in NEI for NE languages. From the literature survey, it has been found that a large number of paper dictionaries have been compiled by many lexicographers for the maximum numbers of major languages of NEI. At present, due to expansion of computer and Internet, a small number of E-dictionary has been developed on both online and offline for the languages of NEI. In this paper, some of the electronic dictionaries which have been developed for NE languages in NE India are shown in Table 2.

Machine Translation

Machine Translation (MT) is one of the most important applications and research tasks of NLP which investigates the use of software to translate text or speech from one natural language to another natural language using computers with or without human assistance. The MT system which generates translation between two specific languages are called bilingual MT systems. The bilingual MT system may be either one direction or both directions. The machine translation is as old as that of computers and it was the first computer based applications related to NLP. The MT system generally started in the year 1950, although work can be found from earlier periods. The first non-military computers were developed in 1947, from that time the idea was proposed to translate text from a source language (SL) to a target language (TL) using a computer [14]. At present, it is a very challenging research tasks in the area of computational linguistics and NLP in the world as well as in India. The research scenario in India is relatively young and machine

translation gained momentum in India only from 1980 onwards with institutions like IIT Kanpur, IIT Bombay, IIIT Hyderabad, University of Hyderabad, NCST Mumbai. The Technology Development for Indian Languages (TDIL), Centre for Development of Advanced Computing (CDAC) and Ministry of Communications and Information Technology are playing a major role in developing the MT systems [16]. The MT system is very important for human nowadays for the following reasons:

- Huge amount of text can be translated from one natural language to another natural language using a MT system which is not possible by human translators.
- It can be used to reduce the human efforts and to give the translation results quickly.
- The use of MT system can increase the volume and speed of translation throughput.
- Manual translation for translating the huge amount of text document is not only time consuming, but also need a more expense. Therefore, MT system can be used to save time and reduce cost.

Problems with Machine Translation

The machine translation is a very difficult research task in NLP due to some problems with it like Word Order (WO), Word Sense Ambiguity (WSD), Part-Of-Speech (POS) and Idioms. These problems are different between different languages. The problems of MT are discussed as below for English and Bodo languages:

Word Order

Word order is different between English and Bodo languages.
English WO: subject(S)-verb(V)-object(O)
Bodo WO: subject(S)-object(O)-verb(V)
For example:

English: I(S) eat(V) rice(O). WO=SVO

Bodo : आं (S) औंखाम(O) जायौ(V)। WO=SOV

Word sense ambiguity

The same word may have different meaning or sense when being translated to another language.
For example:

English: Plant means **Tree** or **Factory**

Bodo: Plant means बिफां or कारखाना

POS

The POS is different between English and Bodo languages. Pre-position is used in English and Post-position is used in Bodo language.
For example:

English: Anil has sat **on** the table (O).
Bodo: अनिला आर्सा (O) **सायाव** जिरायबाय।

Idioms

Meaning of idiom is different between English and Bodo languages.
For example:

English: 'as well as'
Bodo: 'आरो'

Different approaches of MT

There are various approaches of machine translation. Generally, the approaches of MT can be divided into main three categories [12, 13], namely Human Translation with Machine Support, Machine Translation with Human Support and Fully Automated (Automatic) Machine Translation. Again the Fully Automated (Automatic) MT can be divided into seven categories, namely Empirical (or Corpus Based)

MT, Knowledge Based MT, Hybrid MT, Rule Based MT (RBMT), Principle Based MT and Online Interactive MT. The Empirical (or Corpus Based) MT can be divided into two categories, namely Statistical Machine Translation (SMT) and Example Based Machine Translation (EBMT). The SMT approach can also be divided into three categories, namely, Word Based Translation, Phrased Based Translation and Hierarchical Phrased Based Translation. The RBMT can also be divided into three categories, namely Direct MT, Transfer MT and Interlingua MT.

Example of a MT System

Let us consider, English to Assamese (E-A) is a bilingual MT system. In this system, users can translate the huge amount of text of English

Table 2: E-dictionary developed for NE languages

Title of the E-dictionary	Languages	Developers	Place/ Institution	Year
English-Assamese-Bodo Trilingual E-dictionary (http://www.iitg.ernet.in /rcilts/dictionary.html)	Assamese, Bodo, English	Resource Centre for Indian Language Technology Solutions	IIT Guwahati, Assam	2004
Multilingual Online Dictionary (English to NE languages) [www.xobdo.org]	Assamese, Bishnupriya Manipuri,Bodo, Dimasa, English, Garo, Hmar, Karbi, Khasi, Meiteilon, Mising	Bikram M Baruah	Assam, India	2006
Bilingual Dictionary of Words & Phrases [www.bengali-dictionary.com]	Bengali, English	Subhamay Ray	Tripura, India	2006
Nepali to Hindi Bilingual Electronic Dictionary [9]	Hindi, Nepali	Shantanu Kar, Alok Chakrabarty	Silchar, Assam	2011
Building Multilingual Lexical Resources [10]	Assamese, Bodo, Hindi	Shikhar Kr. Sarma et al.	Gauhati University, Assam	2012
Manipuri-English Bilingual E-dictionary [4]	English, Manipuri	S. Poireiton Meitei et al.	Assam University, Silchar	2012
Mizo-English-Mizo Online Dictionary [http://www.freelang.net /online/mizo.php]	English, Mizo	Renato B. Figueiredo	Mizoram, India	2014
Web Enabled Multilingual Manipuri Dictionary [3]	English, Hindi, Manipuri	Yumnam Bablu Singh	Assam University, Silchar, Assam	2015
Kokborok-English Bilingual Electronic Dictionary [8]	English, Kokborok	Partha Sarkar, Bipul Syam Purkayastha	Assam University, Assam	2015
Axomi: Assamese-English and English-Assamese E-dictionary (App for Android OS)	Assamese, English	Manabendra Gogoi	Assam	2015
Multilingual Assamese Electronic Dictionary (Assamese to Bengali, English and Hindi) [7]	Assamese, Bengali, English, Hindi	Saiful Islam, Bipul Syam Purkayastha	Assam University, Silchar	2015
English to Assamese, Bengali and Hindi Multilingual Electronic Dictionary [5]	Assamese, Bengali, English, Hindi	Saiful Islam	Assam University, Silchar, Assam	2016
English to Nepali Online Dictionary [www.englishnepalidictionary.com]	English, Nepali	NA	Sikkim, India	2016
Multilingual Bengali Electronic Dictionary (Bengali to Assamese,	Assamese, Bengali,	Saiful Islam, Bipul	Assam University,	2016

English and Hindi) [6]	English, Hindi	Syam Purkayastha	Silchar, Assam	
Assamese to Bengali Bilingual E-Dictionary Using Sequential Search Technique [20]	Assamese, Bengali	Saiful Islam. Bipul Syam Purkayastha	Assam University, Silchar, Assam	2016

Fig. 2: Example of sentences in E-A MT system

Table 3: MT system developed in NEI

Title of the MT system	Languages	Technique	Developers	Place/Institution	Year
English to Assamese and Manipuri MT system [14]	Assamese, English, Manipuri	EBMT, RBMT	IIT Guwahati	Assam, India	2004
Assamese to English MT system [11]	Assamese, English	SMT	Pranjal Das, Kalyanee K. Baruah	Gauhati University, Assam	2014
English to Assamese MT system [15]	Assamese, English	SMT	M. T. Singh et al.	Dibrugarh University, Assam	2014
Machine Translation for Assamese-English Using Apertium [17]	Assamese, English	RBMT	Pranjal Das et al.	Gauhati University, Assam	2014
Bengali to Assamese MT system using Moses [12]	Assamese, Bengali	SMT	Nayan Jyoti Kalita, Baharul Islam	Royal School of Engineering and Technology, Assam	2015
English to Nepali Statistical Machine Translation System [18]	English, Nepali	SMT	Abhijit Paul, Bipul Syam Purkayastha	Assam University, Silchar, Assam	2016
English to Bodo Phrase Based Statistical Machine Translation [19]	English, Bodo	SMT	Saiful Islam, Bipul Syam Purkayastha	Assam University, Silchar, Assam	2016

language (SL) into Assamese language (TL) using computers. Some examples of sentences in English to Assamese MT system are shown in figure 2.

Machine Translation System Developed in NEI

In this section, we discuss about the existing machine translation systems which are

developed in NEI for NE languages. From the literature survey, it has been found that a very small number of MT system has been developed in NEI using different approaches by different developers. In this paper, some of the MT systems which have been developed for North-East languages in NEI are shown in Table 3.

CONCLUSION

Electronic dictionary is a powerful dictionary whose data is found in digital form and can be accessed through online and offline from anywhere place using a computer, smart phone and PDA. Through this dictionary, a user can look up the meaning of word and other related information of the word like POS, synonyms and examples from a source natural language to the target natural languages. Machine translation is the process of automatic translation of text from a source natural language to another natural language using computers on both online and offline.

Through the MT system, used can translate huge amount of words or phrases or sentences from a specific natural language to another natural language. The E-dictionary and MT system are the tremendously helpful for people to extend their knowledge about the known and unknown natural languages. These applications are also very important in NLP to implement other research tasks related to NLP. The main purpose of this paper is to focus on the existing E-dictionary and MT system which are developed for NE languages in the NE India. A small number of E-dictionary and MT system have been developed for

NE languages in the NEI. Nowadays, some research scholars are working on E-dictionary and MT system for NE languages in India as well as in the NE region. Since, NE is a multilingual region in India. Therefore, the E-dictionary and MT system will be helpful for NE people as well as other people of India and abroad.

REFERENCES

1. Government of India: 47th Report (June 2008 to July 2010) of the Commissioner for Linguistic Minorities: Ministry of Minority Affairs, India, 2011.
2. T. Raatan: History, Religion and Culture of North East India, Isha book, Delhi, 2006.
3. Singh, Y. B.: Corpus and wordnet based Multilingual Manipuri Dictionary, academia. edu.
4. Meitei, S. P., Ningombam,S., Devi, H. M., Purkayastha, B. S.: A Manipuri-English Bilingual Electronic Dictionary: Design and Implementation. *International Journal of Engineering and Innovative Technology*, Vol.**2**, No.1, 2012.
5. Islam, S.: An English To Assamese, Bengali and Hindi Multilingual E-Dictionary. *International Journal of Current Engineering and Scientific Research*, Vol.**3**, No. 9, 2016.
6. Islam, S., Purkayastha,B. S.: Multilingual Bengali Electronic Dictionary Using Sequential Search Technique. *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. **5**, No. 3, pp. 3307-3314, 2016.
7. Islam, S., Purkayastha,B. S.: Development of Multilingual Assamese Electronic Dictionary. *International Journal of Computer Science and Information Technologies*, Vol. **6**No. 6, pp. 5446-5452 , 2015.
8. Sarkar, P., Purkayastha,B. S.:Morphological Analyzer in the Development of Bilingual Dictionary (Kokborok-English) - An Analysis for Appropriate Method and Approach. *International Journal of Engineering and Innovative Technology*, Vol. **4**, No.10, 2015.
9. Kar, S., Chakrabarty, A.: Expansion of the First Hindi-Nepali Word-Net based Bilingual Dictionary and the advancement of the Human-Machine Interface. Special Issue of *International Journal of Computer Applications*, 2011.
10. Sarma , S. K., Sarmah, D., Brahma , B., Mahanta, M., Bharali, H., Saikia, U.: Building Multilingual Lexical Resources Using Wordnets: Structure, Design and

Implementation. Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pp. 161–170, COLING, Mumbai, 2012.

11. Das, P., Baruah, K. K.: Assamese to English Statistical Machine Translation Integrated with a Transliteration Module. *International Journal of Computer Applications*, Vol. **100**, No. 5, pp. 20-24, 2014.
12. Kalita, N. J., Islam, B.: Bengali to Assamese Statistical Machine Translation using Moses (Corpus Based). Proceedings of the International Conference on *Cognitive Computing and Information Processing*, 2015.
13. Godase, A., Govilkar, S.: Machine Translation Development for Indian Languages and its Approaches. *International Journal on Natural Language Computing*, Vol. **4**, No. 2, pp. 55-74, 2015.
14. Antony, P. J.: Machine Translation Approaches and Survey for Indian Languages. Computational Linguistics and Chinese Language Processing, Vol. **18**, No. 1, pp. 47-78, 2013.
15. Singh, M. T., Borgohain, R., Gohain, S.: English-Assamese Machine Translation System. *International Journal of Computer Applications*, Vol.**93** No. 4, pp. 1-6 , 2014.
16. Sanyal, S., Borgohain, R.: Machine Translation system in India: Annals of Faculty Engineering Hunedoara – *International Journal of Engineering*, pp. 137-142, 2013.
17. Das, P., Baruah, K. K., Hannan, A., Sarma, S. K.: Rule Based Machine Translation for Assamese-English Using Apertium**.** *International Journal of Emerging Technologies in Computational and Applied Sciences*, Vol.**8**, No.5, pp. 401-406, 2014
18. Abhijit, P., Purkayastha, B. S.:English to Nepali Statistical Machine Translation System. Proceedings of the International Conference on Computing and Communication Systems, NEHU, Meghalaya, Springer, 2016 (Accepted and presented, waiting for publish).
19. Islam, S., Purkayastha,B. S.: English to Bodo Phrase Based Statistical Machine Translation. Proceedings of the 10Th International Conference On Advanced Computing & Communication Technologies, APIIT, India, Springer, 2016 (Accested and presented, waiting for publish).
20. Islam, S., Purkayastha,B. S.: Assamese to Bengali Bilingual E-Dictionary Using Sequential Search Technique. Proceedings of the DST Sponsored National Seminar On Computational Research And Its Development In Experimental Sciences. *Journal of Science forum*, Vol.**5** NO.1 pp. 90-98, 2016.
21. Kumar, E.: Natural Language Processing (Book), 2011. Lew, R.: Online dictionary skills, Adam Mickiewicz University, 2013.