A Study on Various Applications of NLP Developed for North-East Languages

Saiful Islam[I], Maibam Indika Devi[II], Prof. Bipul Syam Purkayastha[III]
[I, II, III] Department of Computer Science, Assam University, Silchar, PIN-788011, Assam, India
[I]sislam.mca@gmail.com, [II] indika.aus@rediffmail.com, [III] bipul_sh@hotmail.com

***Abstract*— Natural languages are the most important aspect of human life for communication. Natural Language Processing (NLP) is an important research area and application in computer science and computational linguistics, which explores how a computer can be used to understand and manipulate the text or speech of natural languages to do useful things. A small number of research or applications of NLP has been done for North-East (NE) languages in India. In this paper, we discuss the five most important as well as one of the major applications of NLP which are developed for NE natural languages and frequently used in human life namely, Electronic Dictionary (E-dictionary), Machine Translation (MT), Part of Speech Tagging (POST), Wordnet and Word Sense Disambiguation (WSD). These applications are very helpful and form an intermediate and necessary step towards implementation of higher applications of NLP. Since NE is a multilingual region in India; therefore these applications will be basically helpful for NE people as well as other people of India and abroad. The main purpose of this paper is to explore the importance, techniques, features of E-dictionary, MT, POST, Wordnet and WSD and to review the existing works of these applications developed for NE natural languages in India.**

Keywords- E-Dictionary; MT; NLP; POS; Wordnet; WSD

I. INTRODUCTION

Language is the most important communication media for all human beings. The languages which are prominently used by humans for speaking and communication purpose are called human languages or natural languages. Natural language processing is one of the interdisciplinary research areas in Computational linguistic, Computer Science and Artificial Intelligence. The NLP is a very interesting area which deals with interactions between computers and natural languages. It is used to design and develop computer programs that will analyze, understand and produce speech or text of natural languages [7]. At present, it is a very demandable research task in computer science that explores how computers can be used to understand and manipulate text or speech of natural language to do useful things. A large number of various applications of natural language processing have been developed in the world as well as in India. The most commonly used applications of NLP or research tasks in NLP are [19] Automatic Text Summarization, Discourse Analysis, Electronic Dictionary, Grammar Checking, Information Extraction, Information Retrieval, Machine Translation, Machine Transliteration, Morphological Segmentation, Named Entity Recognition, Natural Language Generation, Natural Language Understanding, Optical Character Recognition, Part of Speech Tagging, Parsing, Question Answering, Speech Processing, Speech Recognition, Speech Segmentation, Spelling Checker, Wordnet, and Word Sense Disambiguation.

In this paper, we discuss only the five most important applications of NLP which are developed in India for NE languages, namely Electronic Dictionary, Machine Translation, Part of Speech Tagging, Wordnet and Word Sense Disambiguation.

II. NORTH-EAST LANGUAGES

North-East is one of the most linguistically and ethnically diverse regions of India. The North-East India (NEI) consists of eight states and the states are Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Tripura and Sikkim [2]. Each of the states has their own culture, language and tradition. The people of NEI belonging to different communities have different languages. The languages act as a bridge amongst the people and help in creating a bond among them. There are about 220 spoken languages in NEI and the languages are divided mainly into three language families, namely Indo-Aryan, Sino-Tibetan and Austro-Asiatic. 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. Therefore, the NE region is also known as multilingual and multicultural region of India. The different languages of the eight states of NEI are shown in Table 1[1].

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, Bengali, Bodo, English	Assamese, Bengali, Bishnupriya Manipuri, Bodo, Dimasa, Hindi, Karbi, Mishing, Nepali, Rabha
Manipur	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, Kok borok	Bengali, Bishnupriya Manipuri, Garo, Halam, Hindi, Kokborok (Tripuri), Manipuri, Mogh

III. STUDY ON THE APPLICATIONS OF NLP

A large number of various applications of NLP have been developed in the world as well as in India for natural languages. But, a small number of applications of NLP have been developed for NE languages in India. In this paper, we discuss the following five most important applications of NLP which are developed for NE natural languages in India.

A. Electronic Dictionary

A dictionary is the most important language learning book which contains an enormous collection of words of one or more particular language(s) and the words are arranged alphabetically with their meaning, Part of Speech (POS), synonyms, phonetics and examples. The dictionary is one of the most important tools to assist students and teachers in understanding as well as enlightening the skill of reading [5]. It is also very helpful for students, research scholars, teachers, travelers and other people to improve their knowledge about the known and unknown human languages. The basic categories of dictionary are shown in figure 1 as below:

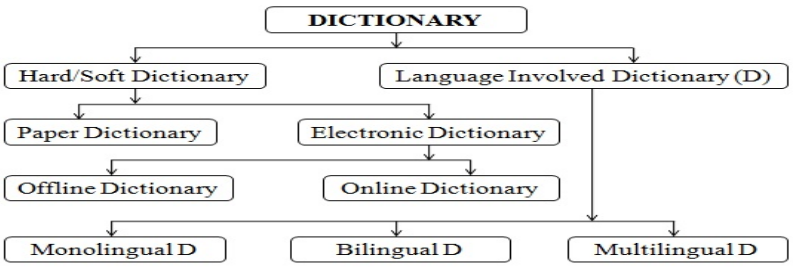

Figure 1. Basic categories of dictionary.

An E-dictionary is a very demandable dictionary nowadays whose data is found in digital form and can be accessed through different media. The E-dictionary is the most powerful tool for a human to learn about the various natural languages on both online and offline using a computer, mobile phone and Personal Digital Assistant (PDA). It is very convenient to use and much better than a paper dictionary. The E-dictionary is an important application of NLP and can be used to implement other research tasks of NLP. There are many word search techniques available to develop the E-dictionary. The developers use different word search technique like Sequential Search Technique (SST), Index Based Search Technique (IBST), Binary Search Technique (BST), Incremental Search Technique (IST) and Wildcard Search Technique (WST) to look up the words from the dictionary on both online and offline mode [20]. An example of English to Assamese and Bengali (E-AB) multilingual E-dictionary is shown in figure 2.

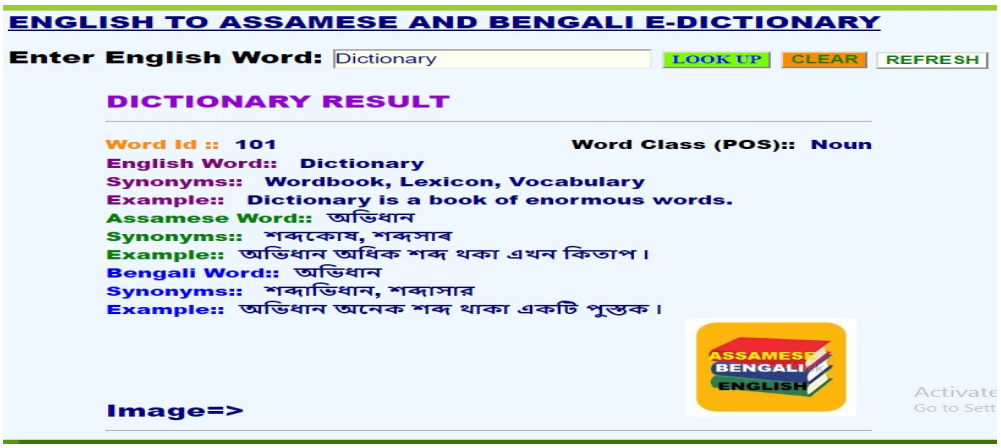

Figure 2. Snapshot of E-AB multilingual E-dictionary.

Large numbers of paper dictionaries have been compiled by many lexicographers for the maximum numbers of major natural languages of NEI. At present, due to the expansion of computer and Internet, a very small number of E-dictionary has been developed on both online and offline mode for the languages of NEI. In this paper, some of the existing electronic dictionaries which have been developed for NE languages in India are mentioned in Table 2.

Table 2: Existing E-dictionary for NE languages.

Title of the E-dictionary	Used Languages	Developers	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	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	2006
Bilingual Dictionary of Words & Phrases [www.bengali-dictionary.com]	Bengali, English	Subhamay Ray	2006
Nepali to Hindi Bilingual Electronic Dictionary [9]	Hindi, Nepali	Shantanu Kar, Alok Chakrabarty	2011
Building Multilingual Lexical Resources [10]	Assamese, Bodo, Hindi	S. K. Sarma, D. Sarmah, B. Brahma, M. Mahanta, H. Bharali, U. Saikia	2012
Manipuri-English Bilingual E-dictionary [4]	English, Manipuri	S. P. Meitei, S. Ningombam, H. M. Devi, B. S. Purkayastha	2012
Mizo-English-Mizo Online Dictionary [http://www.freelang.net/online/mizo.php]	English, Mizo	Renato B. Figueiredo	2014
Web Enabled Multilingual Manipuri Dictionary [3]	English, Hindi, Manipuri	Yumnam Bablu Singh	2015
Kokborok-English Bilingual Electronic Dictionary [8]	English, Kokborok	Partha Sarkar, Bipul Syam Purkayastha	2015
Axomi: Assamese-English and English-Assamese E-dictionary (App for Android OS)	Assamese, English	Manabendra Gogoi	2015
Multilingual Assamese Electronic	Assamese, Bengali,	Saiful Islam, Bipul	2015

Dictionary (Assamese to Bengali, English and Hindi) [7]	English, Hindi	Syam Purkayastha	
English to Assamese, Bengali and Hindi Multilingual Electronic Dictionary [5]	Assamese, Bengali, English, Hindi	Saiful Islam	2016
English to Nepali Online Dictionary [www.englishnepalidictionary.com]	English, Nepali	Nepalayas	2016
Multilingual Bengali Electronic Dictionary (Bengali to Assamese, English and Hindi) [6]	Assamese, Bengali, English, Hindi	Saiful Islam, Bipul Syam Purkayastha	2016
Assamese to Bengali Bilingual E-Dictionary Using Sequential Search Technique [18]	Assamese, Bengali	Saiful Islam. Bipul Syam Purkayastha	2016

B. Machine Translation

Machine Translation is a process of automatic translating a large amount of text from a particular natural language to another natural language using computers. It is one of the most important applications and challenging research tasks in NLP and was the first application of computer-related applications to NLP. In India, it is relatively young and gained momentum from 1980 onwards in 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 also playing a major role in developing the MT systems [16]. The MT is a very hard research task due to some problems with it like Word Order (WO), Word Sense Ambiguity, Idioms, Pre-position and Post-position. There are many approaches of MT to solve these problems. Nowadays, the most commonly used MT approaches include Rule Based MT (RBMT), Statistical MT (SMT), Example Based MT (EBMT) and Hybrid MT [12, 13]. The RBMT approach can be classified into three categories, namely Direct MT, Transfer MT and Interlingua MT. The SMT approach can also be classified into three categories, namely Word Based Translation, Phrase Based Translation and Hierarchical Phrase Based Translation. The main advantage of MT system is that it can be used to reduce the human efforts and to translate the enormous text from a language to another language quickly which is not possible by the human translators. It can also be used to save time and reduce cost. An example of English to Manipuri MT system is shown in figure 3.

English [SL]	Machine [Computer]	Manipuri [TL]
I live at Manipur.		ঐ মনিপুরদা লৈ।
She is a good student.		মহাক অফবা ছাত্রি অমনি।
Imphal is the capital of Manipur.		ইম্ফাল অসি মনিপুরগী কোনুংনি।
Man is mortal.		মীওইবসি শিবা নাই।

Figure 3. Example of English-Manipuri MT system.

There are multiple numbers of spoken and official languages in NE India. Even then, a very small number of MT systems have been developed for NE languages. Some of the existing MT systems which have been developed in India for NE languages are shown in Table 3.

Table 3: Existing MT system for NE languages.

Title of the MT system	Used Languages	Techniques	Developers	Year
Anglabharati Machine Translation system [13][14]	Assamese, Bengali, English. Hindi, Nepali	RBMT	V.N. Shukla, R. M. K. Sinha	2003
English to Assamese and Manipuri MT system [14]	Assamese, English, Manipuri	EBMT, RBMT	IIT Guwahati Dept. of CSE Group	2004
Manipuri-English Machine Translation System [21]	English, Manipuri	EBMT	Thoudam Doren Singh, Sivaji Bandyopadhyay	2010
Assamese to English MT system [11]	Assamese, English	SMT	Pranjal Das, Kalyanee K. Baruah	2014
English to Assamese MT system [15]	Assamese, English	SMT	M. T. Singh, R. Borgohain, S. Gohain	2014
Machine Translation for Assamese-English Using Apertium [17]	Assamese, English	RBMT	P. Das, K. K. Baruah, A. Hannan, S. K. Sarma	2014
Bengali to Assamese MT system using Moses [12]	Assamese, Bengali	SMT	Nayan Jyoti Kalita, Baharul Islam	2015

C. Part of Speech Tagging

Part of Speech Tagging is the process of assigning a word, its word class based on its context. The POST is not as easy as it seems to be. Ambiguity, where a word depending on its context may represent different word class is a difficult task to handle in POST. It plays a very important role in Question-Answering Systems, Parsing and Machine Translation. POS taggers, based on how training and tagging are done, can be categorized into supervised and unsupervised. The supervised taggers make use of pre-tagged corpora as the basis for implementing the automated tagging process. The unsupervised taggers do not make use of any pre-tagged corpora; rather it employs sophisticated methods and tools to automatically tag word classes. A few of the commonly employed POST approaches include Rule Based, Stochastic and Transformation Based approaches. The Rule Based approach uses grammatical rules or context rules to assign POS tags to words. The Transformation Based Learning (TBL) algorithm is one such algorithm which uses linguistic rules to implement the tagging process. The Stochastic tagging approaches include any model or algorithm which relies on frequency or probability. N-gram approach, Hidden Markov Model (HMM), Conditional Random Field (CRF), Baum-Welch Algorithm, Maximum Entropy, Viterbi Algorithm falls in this approach. The Transformation Based approach uses transformation rules to transform a state to another state. The transformation eases the process of finding a suitable tag for each word. The suitability and adaptability of an approach depend on specific language's structure and morphology. Other than computational applications, POS tagging is also useful in determining authorship and can be adapted for mathematical text analysis also [61]. Therefore, POS tagging is useful not only from the computational linguistics point of view, it has a variant of applications over a broad range of fields. Let us consider a simple example of Bengali phrase, রিনা একটি সুন্দর মেয়ে ছিল. The POST of the Bengali phrase is shown in figure 4.

Figure 4. Example of Bengali POST.

Many POS taggers have been developed using varying techniques in the world as well as in India for natural languages. But, a small number of POST have been developed in India for NE natural languages. Some of the existing POST which has been developed for NE languages in India are shown in Table 4.

Table 4: Existing POST for NE languages.

Title of the POST	Used Languages	Techniques	Developers	Year
A Hybrid Model for Part-of- Speech Tagging and its Application to Bengali [47]	Bengali	Combination of supervised and unsupervised learning using HMM	S. Dandapat, S. Sarkar, A. Basu	2004
Bengali part of speech tagging using conditional random field [42]	Bengali	CRF	A.Ekbal, R. Haque, S. Bandyopadhyay	2007
POS Tagging Using HMM and Rule-based Chunking [43]	Bengali, Hindi	HMM and Rule Based	A. Ekbal, S. Mandal, S. Bandyopadhyay	2007
POS tagging and Chunking for Indian Languages [44]	Hindi, Bengali	CRF	H. Agrawal	2007
POS Tagging and Chunking using Decision Forests [45]	Bengali, Hindi	Decision Forests	S.C. Pammi, K. Prahallad	2007
Morphology driven Manipuri POS tagger [51]	Manipuri	Morphological Analysis with prefix, suffix and root word dictionaries	Singh & Bandyopadhyay	2008
Part of Speech Tagging in Bengali using Support Vector Machine [41]	Bengali	Support Vector Machine (SVM)	A. Ekbal, S. Bandyopadhyay	2008
Maximum Entropy Based Bengali Part of Speech Tagging [46]	Bengali	Maximum Entropy Model	A.Ekbal, R. Haque and S. Bandyopadhyay,	2008
Hidden Markov Model Based Probabilistic Part of Speech Tagging For Nepali Text [55]	Nepali	HMM	M.R. Jaishi	2009
Voted Approach for Part of Speech Tagging in Bengali [48]	Bengali	Maximum Entropy, CRF, SVM	A.Ekbal , Md. Hasanuzzaman, S. Bandyopadhyay	2009
Part of Speech Tagger for Assamese Text [39]	Assamese	HMM and Viterbi algorithm	N. Saharia, D. Das, U. Sharma, J. Kalita	2009
Part of Speech Tagging in Manipuri: A Rule-based Approach [50]	Manipuri	Rule Based	Kh. R. Singha,	2012
Design Considerations for Developing a POS tagset for Khasi [60]	Khasi	Develop Khasi tagsets based on EAGLES guidelines	M.J.Tham	2012
Part of Speech Tagging in Manipuri with Hidden Markov Model [52]	Manipuri	HMM	Kh R. Singha, B.S. Purkayastha , Kh D. Singha	2012
Part of Speech (POS) Tagger for Kokborok [58]	Kokborok	Rule Based Morphological Analyzer Driven and Supervised Methods (CRF, SVM)	B. G. Patra, K. Debbarma, D. Das, S. Bandyopadhyay	2012
POS Tagging of Assamese Language and Performance Analysis of CRF++ and TBL Approaches [40]	Assamese	CRF and TBL	A.K. Barman, J. Sarmah , S. K Sarma	2013
A Memory Based POS Tagger for Bengali [49]	Bengali	Memory Based Learning-based on similarity-based classification	K. Sarkar, A. R.Ghosh	2013

Support Vector Machines based Part of Speech Tagging for Nepali Text [54]	Nepali	SVM	T. B. Shahi, T. N. Dhamala and B. Balami,	2013
Resource Building and POS Tagging for Mizo Language [59]	Mizo	POS Tagsets developed based on PennTreeBank tagset	P. Pakray, A. Pal, G. Majumder, A. Gelbukh	2015
Hidden Markov Model Based Part of Speech Tagging for Nepali Language [53]	Nepali	HMM	A. Paul, B. S. Purkayastha, S. Sarkar	2015
Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach [57]	Nepali	HMM integrated Rule Based	P. Sinha, N. M. Veyie, B.S.Purkayastha	2015
Part of Speech Tagging Using Statistical Approach for Nepali Text [56]	Nepali	HMM and Viterbi Algorithm	Archit Yajnik	2017

D. Wordnet

Wordnet is a very crucial electronic lexical database of one or more specific language(s) which consists of an enormous collection of words and each word has Id, POS, synonyms, gloss, and examples along with a number of semantic relations among all the synsets. Therefore, the Wordnet can be considered as the combination of E-dictionary and thesaurus. The first Wordnet in the world was built at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller for the English language in 1985. In recent years, it has been edited by Christiane Fellbaum [27]. The position of Wordnet in India is still an infant. The work on Wordnet was started in India since 2000. The Wordnet of India was developed at IIT Bombay for the Hindi language [28]. There are generally two approaches used to build Wordnet, namely Merge Approach (MA) and Expansion Approach (EA) [28]. The words of every natural language may have multiple senses and different words may have the same sense. This type of problem can be easily solved using the Wordnet. The Wordnet is an important application of NLP and can be used to implement other applications of NLP like Information Retrieval, MT, WSD, E-dictionary and even automatic crossword puzzle generation. It can also be used in automatic text analysis and in the applications of Artificial Intelligence. An example of Assamese Wordnet is shown in figure 5.

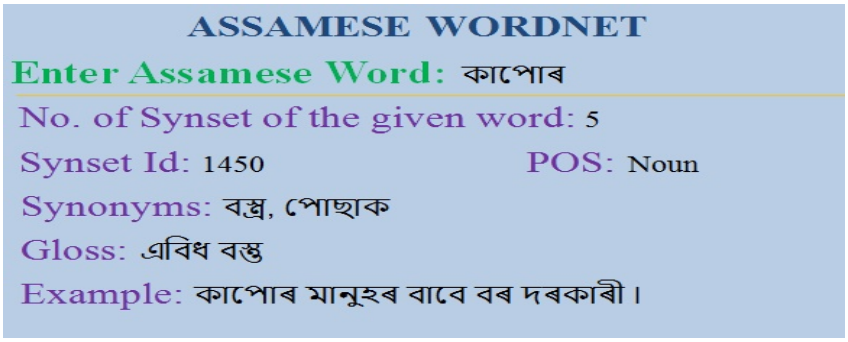

Figure 5. Example of an Assamese Wordnet.

Although India is a multilingual country, some works have been done on Wordnet for Indian languages and a very small number of works on Wordnet have been done in India for NE natural languages. Some of the existing Wordnet which has been developed for NE languages in India are shown in Table 5.

Table 5: Existing Wordnet for NE languages.

Title of the Wordnet	Used Languages	Techniques	Developers	Year
IndoWordnet: a wordnet for Indian Languages [http://www.cfilt.iitb.ac.in/indowordnet] [28]	Assamese, Bengali, Bodo, English, Hindi, Manipuri, Nepali	EA	Prof. Pushpak Bhattacharyya	2006
Development of NE Wordnet: An Integrated Wordnet for Languages of the North-East India Assamese & Bodo [25]	Assamese, Bodo	EA	U. Saikia, B. Brahma, D. Sarmah	2009
Experiences in building the Nepali WordNet - insights and challenges [22]	Nepali	EA	A.Chakrabarty, B.S. Purkayastha, A. Roy	2010
A Wordnet for Bodo Language: Structure and Development [23]	Bodo	EA	S. K. Sarma, B. Brahma, M. Gogoi, M.B. Ramchiary	2010
Foundation and Structure of Developing an Assamese Wordnet [25]	Assamese	EA	S. K. Sarma, R.Mehdi, M. Gogoi, U. Saikia	2010
Developing Bengali WordNet Affect for Analyzing Emotion	Bengali	EA	Dipankar Das, Sivaji Bandyopadhyay	2010
Development of Assamese WordNet [24]	Assamese	EA	I. Hussain, N. Saharia, U. Sharma.	2011
Experiences in building the indowordnet-a wordnet for Manipuri [26]	Manipuri	EA	Yumnam Bablu Singh, Prof. Bipul Syam Purkyastha	2011

E. Word Sense Disambiguation

Word Sense Disambiguation is one of the challenging problems in the field of NLP. It is also known as Lexical Ambiguity Resolution**.** The WSD is the identification of sense or meaning of a polysemy word occurring in a text or discourse. It is an intermediate step towards many NLP applications such as Information Retrieval, Information Extraction, Content Analysis, Semantic web, Machine Translation, Question Answering Systems, Text Mining, etc. There are mainly three approaches of WSD, namely Knowledge Based (KB) approach, Machine Learning Based (MLB) approach and Hybrid approach. The WSD using selectional preferences is one of the KB approaches which relies on knowledge bases such as WordNet, CyC, Aquilex, etc. While Overlap Based KB approach relies on machine readable dictionary. It includes techniques as Mickrokosmos approach to WSD, Lesk's Algorithm, Walkers Algorithm, WSD using conceptual density, WSD using Random Walk Algorithm. The MLB approach makes use of corpus to build a classification model. It employs two stages-learning and classification. The MLB approach can be divided into supervised, semi-supervised and unsupervised method. Naive Bayes Algorithm, Decision List Algorithm, Exemplar based, SVM, HMM, Boosting Algorithm are some algorithms belonging to supervised method. Semi-supervised method includes Yarowsky's supervised Algorithm, Decision Lists, Decision Tree, Bootstrapping Algorithm. Unsupervised method comprises of Hyperlex, Lin's Approach, WSD using parallel corpora etc. The Hybrid approach is a combination of KB Approach and MLB Approach. SenseLearner, Structured Semantic Interconnections (SSI) and Iterative Approach to WSD are few models of Hybrid Approach.

The WSD even though one of the toughest applications of NLP and is a very important step in computational linguistics. Let us consider an example, "There is not a single plant in our village". The word "plant" has two meaning-Tree and Factory. In this context, the word "plant" is a doubtful word in English language. Similarly, in Bodo language, the word "लाइफां (Plant)" means बिफां (Tree) and कारखाना (Factory). Therefore, लाइफां is a doubtful word in Bodo language. Almost all of the NE languages are tone languages, this feature increases the number of polysemy words. In this aspect, WSD for NE languages is a must, but only a few works on WSD has been done till date. Some of the existing WSD works which have been developed for NE languages in India are mentioned in Table 6.

Table 6: Existing WSD for NE languages.

Title of the WSD	Used Languages	Techniques	Developers	Year
Dictionary Containing Example based Nepali WSD [38]	Nepali	Lesk Algorithm approach for nouns	N. Shrestha, P. Hall, S. K. Bista	2008
Word Sense Disambiguation in Bengali applied to Bengali-Hindi Machine Translation [33]	Bengali	Unsupervised Graph-based approach	A. Das, S. Sarkar	2013
Assamese Word Sense Disambiguation using Supervised Learning [29]	Assamese	Naive Bayes Classifier Approach	P. P. Borah, G. Talukdar, A. Baruah	2014
A Decision Tree Based Word Sense Disambiguation System In Manipuri[34]	Manipuri	Decision Tree Based	R. L. Singh, K. Ghosh, K.Nongmeikapam, S. Bandyopadhyay	2014
Knowledge based Approaches to Nepali Word Sense Disambiguation [35]	Nepali	Knowledge Based	A. Roy,S. Sarkar,B. S. Purakayastha	2014
Word sense disambiguation in Nepali language [36]	Nepali	Modified Adapted Lesk algorithm	U.R.Dhungana, S. Shakya	2014
Implementation of Walker Algorithm in Word Sense Disambiguation for Assamese language [31]	Assamese	Walker based algorithm	P. Kalita, A.K. Barman	2015
Automatic classification of Bengali sentences based on sense definitions present in Bengali Wordnet [32]	Bengali	Naive Bayes probabilistic model	A. R. Paul, D. Saha, N. S. Das	2015
Word Sense Disambiguation using WSD specific WordNet of polysemy words [37]	Nepali	Based on clue words based Wordnet	U. R. Dhungana, S. Sakhya, K. Baral, B. Sharma	2015
Decision Tree based Supervised Word Sense Disambiguation for Assamese [30]	Assamese	Supervised Decision Tree Based	J. Sarmah, S. K. Sarma	2016

IV. CONCLUSION

The research work on Natural Language Processing is growing in the world as well as in India. Many applications of NLP have been developed in India for Indian languages. The E-Dictionary, MT system, POST, Wordnet, WSD, Information Retrieval and Spelling Checker are the important applications of NLP. The E-dictionary and MT system are prominently language learning tool and are frequently used by human in daily life. The MT and Wordnet are very challenging research problems in the field of Computational Linguistics and NLP. The E-dictionary and Wordnet can be used to implement the other applications of NLP. In this paper, we have briefly discussed the importance, techniques and features of the five most important application of NLP, namely E-Dictionary, MT, POST, Wordnet, and WSD. We have also discussed the existing works of these applications which have been developed for NE natural languages in India. It has been found that only a few works have been developed in India for NE natural languages. North-East is a multilingual and one of the popular regions of India. There are many official languages and different communities use different languages in NE India. In this regard, these applications will be helpful for communication amongst the people of different communities. Therefore, through this paper, we are trying to explore the importance, techniques, features and the existing works on the above mentioned NLP applications for NE natural languages. It is expected that this paper would be helpful for students, research scholars and the public as a whole.

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] Yumnam Bablu Singh, "Corpus and wordnet based Multilingual Manipuri Dictionary", academia.edu, 2014.
[4] S. P. Meitei, S. Ningombam, H.M. Devi, and B.S. Purkayastha, "A Manipuri-English Bilingual Electronic Dictionary: Design and Implementation", International Journal of Engineering and Innovative Technology, Vol.2, No.1, 2012.
[5] Saiful Islam, "An English to Assamese, Bengali and Hindi Multilingual E-Dictionary", International Journal of Current Engineering and Scientific Research, Vol.3, No. 9, 2016.
[6] S. Islam and B. S. Purkayastha, "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] S. Islam and B. S. Purkayastha, "Development of Multilingual Assamese Electronic Dictionary", International Journal of Computer Science and Information Technologies, Vol. 6, No. 6, pp. 5446-5452 , 2015.
[8] P. Sarkar and B. S. Purkayastha, "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] S. Kar and A. Chakrabarty, "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] S. K. Sarma, D. Sarmah, B. Brahma, M. Mahanta, H. Bharali and U. Saikia, "Building Multilingual Lexical Resources Using Wordnets: Structure, Design and Implementation", Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 161–170, COLING, Mumbai, 2012.
[11] P. Das and K. K. Baruah, "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] N.J. Kalita and B. Islam, "Bengali to Assamese Statistical Machine Translation using Moses (Corpus Based)", Proceedings of the International Conference on Cognitive Computing and Information Processing, 2015.
[13] A. Godase and S. Govilkar, "Machine Translation Development for Indian Languages and its Approaches", International Journal on Natural Language Computing, Vol. 4, No. 2, pp-55-74, 2015.
[14] P. J. Antony, "Machine Translation Approaches and Survey for Indian Languages", Computational Linguistics and Chinese Language Processing, Vol. 18, No. 1, pp. 47-78, 2013.
[15] M. T. Singh, R. Borgohain and S. Gohain, "English-Assamese Machine Translation System", International Journal of Computer Applications, Vol. 93, No. 4, pp-1-6 , 2014.
[16] S. Sanyal and R. Borgohain, "Machine Translation system in India", Annals of Faculty Engineering Hunedoara – International Journal of Engineering, pp-137-142, 2013.
[17] P. Das, K. K. Baruah, A. Hannan and S. K. Sarma, "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] S. Islam and B.S. Purkayastha, "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.
[19] Ela Kumar, "Natural Language Processing (Book)", 2011.
[20] Robert Lew, "Online dictionary skills", Adam Mickiewicz University, 2013.
[21] T. D. Singh and S. Bandyopadhyay, "Manipuri-English Example Based Machine Translation System", International Journal of Computational Linguistics and Applications. VOL. 1, NO. 1-2, PP. 201-216, 2010.
[22] A. Chakrabarty, B. S. Purkayastha and A. Roy, "Experiences in building the Nepali WordNet - insights and challenges", Proceedings of the 5th Global Wordnet Conference (GWC), New Delhi, pp. 192-197, 2010.
[23] S. K. Sarma, B. Brahm, M. Gogoi and M. B. Ramchiary, "A Wordnet for Bodo Language: Structure and Development", Proceeding in global wordnet conference (GWC10), Mumbai, india, 2010.
[24] I. Hussain, N. Saharia and U. Sharma, "Development of Assamese WordNet", Machine Intclligence:Recent Advances, Narosa Publishing House, Editors. B. Nath. U. Sharma and D. K. Bhattacharyya, ISBN-978-81-8487-140-1, 2011.
[25] .S. K. Sarma, M. Gogoi, U. Saikia and R. Medhi, "Foundation and structure of Developing Assamese WordNet", Proceedings of the 5th International Conference of the Global WordNet Association (GWC), 2010.
[26] Y. B. Singh and B. S. Purkyastha, "Experiences in building the indo wordnet- a wordnet for Manipuri", International Journal of Engineering Science and Technology, ISSN : 0975-5462, pp. 3997-4002, Vol. 3, No. 5, 2011.
[27] C. Fellbaum, "WordNet: An Electronic Lexical Database", Cambridge, MIT Press, 1998.
[28] P. Bhattacharyya, "Indowordnet", In Language Resources and Evaluation Conference (LREC), Malta, 2010.
[29] P. P. Borah, G. Talukdar and A. Baruah, "Assamese Word Sense Disambiguation using Supervised Learning", International Conference on Contemporary Computing and Informatics (IC3I), 2014.
[30] J. Sarmah and S. K. Sarma, "Decision Tree based Supervised Word Sense Disambiguation for Assamese", International Journal of Computer Applications Foundation of Computer Science (FCS), NY, USA, 2016.
[31] P. Kalita and A.K. Barman, "Implementation of Walker Algorithm in Word Sense Disambiguation for Assamese language", International Symposium on Advanced Computing and Communication (ISACC), 2015.
[32] A. R. Paul, D. Saha and N. S. Das, "Automatic classification of Bengali sentences based on sense definitions present in Bengali wordnet", International Journal of Control Theory and Computer Modeling (IJCTCM), Vol.5, No.1, 2015.
[33] A. Das and S. Sarkar, "Word Sense Disambiguation in Bengali applied to Bengali-Hindi Machine Translation", International Conference On Natural Language Processing (ICON), Noida, 2013.
[34] R. L. Singh, K. Ghosh, K. Nongmeikapam, and S. Bandyopadhy, "Knowledge based Approaches to Nepali Word Sense Disambiguation", International Journal (ACIJ), Vol.5, No.4, 2014.
[35] A. Roy, S. Sarkar and B. S. Purakayastha, "Word sense disambiguation in Nepali language", International Journal on Natural Language Computing, Vol.3, No.3, 2014.
[36] U. R. Dhungana and S. Shakya, "Word sense disambiguation in Nepali language", Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP), 2014.
[37] U. R. Dhungana, S. Sakhya, K. Baral and B. Sharma, "Word Sense Disambiguation using WSD specific WordNet of polysemy words", Proceedings of the IEEE 9th International Conference on Semantic Computing (IEEE ICSC), 2015.
[38] N. Shrestha, P. Hall and S. K. Bista, "Dictionary Containing Example based Nepali WSD", International Conference On Natural Language Processing (ICON), Pune, India, 2008.

[39] N. Saharia, D. Das, U. Sharma, and J. Kalita, “Part of Speech Tagger for Assamese Text”, Proceedings of the ACL-IJCNLP Conference Short Papers, pp.33–36, Singapore, 2009.
[40] A.K. Barman, J. Sarmah and S. K Sarma, “POS Tagging of Assamese Language and Performance Analysis of CRF++ and TBL Approaches”, UKSim 15th International Conference on Computer Modelling and Simulation, 2013.
[41] A. Ekbal and S. Bandyopadhyay, “Part of Speech Tagging in Bengali Using Support Vector Machine”, International Conference on Information Technology, IEEE, 2008.
[42] A. Ekbal, R. Haque and S. Bandyopadhyay, “Bengali part of speech tagging using conditional random field”, Proceedings of Seventh International Symposium, 2007.
[43] A. Ekbal, S. Mandal and S. Bandyopadhyay, “POS Tagging Using HMM and Rule-based Chunking”, Proceedings of the IJCAI, 2007.
[44] H. Agrawal, “POS tagging and Chunking for Indian Languages”, Proceedings of the IJCAI, 2007.
[45] S. C. Pammi and K. Prahallad, “POS Tagging and Chunking using Decision Forests”, Proceedings of the IJCAI , 2007.
[46] A.Ekbal, R. Haque and S. Bandyopadhyay, “Maximum Entropy Based Bengali Part of Speech Tagging”, Advances in Natural Language Processing and Applications Research in Computing Science, pp. 67-78, 2008.
[47] S. Dandapat, S. Sarkar and A. Basu, “A Hybrid Model for Part-of- Speech Tagging and its Application to Bengali", International Journal of Information Technology, Vol. I, No. 4, 2004.
[48] A.Ekbal, M..Hasanuzzaman and S. Bandyopadhyay, “Voted Approach for Part of Speech Tagging in Bengali”, Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, pp-120–129, 2009.
[49] K. Sarkar and A.R. Ghosh, “A Memory Based POS Tagger for Bengal”, Proceedings of the 1st Indian Workshop on Machine Learning, IIT Kanpur, India, 2013.
[50] K. R. Singha, B.S. Purkayastha and K. D. Singha, “Part of Speech Tagging in Manipuri: A Rule-based Approach”, International Journal of Computer Applications, ISSN-0975 – 8887, Vol. 51. No.14, 2012.
[51] T. D. Singh and S. Bandyopadhyay, “Morphology driven Manipuri POS tagger”, Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 91-98, Hyderabad, India, 2008.
[52] K. R. Singha, B.S. Purkayastha and K. D. Singha, “Part of Speech Tagging in Manipuri with Hidden Markov Model”, International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, 2012.
[53] A. Paul, B.S.Purkayastha and S. Sarkar, “Hidden Markov Model Based Part of Speech Tagging for Nepali Language”, International Symposium on Advanced Computing and Communication (lSACC), India, 2015.
[54] T. B. Shahi, T. N. Dhamala and B. Balami, "Support Vector Machines based Part of Speech Tagging for Nepali Text", Central Department of Computer Science and IT, Tribhuvan University, Nepal, 2013.
[55] M.R. Jaishi, "Hidden Markov Model Based Probabilistic PartOf Speech Tagging For Nepali Text", Masters Dissertation, Central Department of Computer Science and IT , Tribhuvan University, Nepal, 2009.
[56] Archit Yajnik, “Part of Speech Tagging Using Statistical Approach for Nepali Text”, World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol.11, No.1, 2017.
[57] P. Sinha, N. M. Veyie and B.S. Purkayastha, “Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid Approach”, International Journal of Emerging Technology and Advanced Engineering, ISSN- 2250-2459, Vol. 5, No.5, 2015.
[58] B. G. Patra, K. Debbarma, D. Das and S. Bandyopadhyay, “Part of Speech (POS) Tagger for Kokborok”, International Journal of Computational Linguistics and Natural Language Processing, 2012.
[59] P. Pakray, A. Pal, G. Majumder and A. Gelbukh, “Resource Building and POS Tagging for Mizo Language”, Proceedings of the 14th Mexican International Conference on Artificial Intelligence (MICAI), IEEE, Mexico, 2015.
[60] M.J. Tham, “Design considerations for developing a parts-of-speech tagset for Khasi”, Proceedings of the IEEE 3rd National Conference on Emerging Trends and Applications in Computer Science, 2012.
[61] U. Schöneberg and W. Sperber, “POS Tagging and its Applications for Mathematics-Text analysis in mathematics”, Proceedings of International Conference of Intelligent Computer Mathematics, pp 213-223, Coimbra, Portugal, 2014.