MT Embedded E-Learning in India - Challenges for NLP/AI

Ritu Nidhi, Tanya Singh, D.K. Lobiyal

Abstract -Indian languages are spoken by more than 90% of its population while most of the higher education happens in English medium. The policy makers in Indian government have realized that by introducing multilingual education electronically, they can reach out to the remotest corner of India and educate all in their mother tongue. The New Education Policy (NEP) draft just released by the government puts a heavy focus on mother tongues in education. The recent initiatives and focus on Natural Language Processing/ Artificial Intelligence (NLP/AI) in education through e-learning is not too surprising in this context. The paper presents the current initiatives in these directions by the government of India, surveys available NLP technologies particularly those for automatic translation of educational content developed by academia and industry and focuses on the Maithili language community. India's education needs are diverse and the success of e-learning depends heavily on the availability of necessary NLP tools in all languages. Almost all of major scheduled Indian languages are considered "resource-poor". While some of these languages may have the basic tools, they lack quality translation tools for delivering education in native language. The situation is more challenging in those languages where even the most basic resources and tools do not exist. Maithili - a language of Bihar and Nepal is such a language. The paper also presents an effort to develop MT resources and tools for Maithili and its application in delivering multilingual content for education.

Keyword- E-learning, NLP, MT, Indian Language, Maithili

I. INTRODUCTION

India is one of the fastest growing markets in multilingual education technology because there is a need to reach out to the remotest regions of India covering a vast multilingual and socio cultural diversity. As per 2011 census report, there are 26 % of Indian population who are illiterate. The inherent socio cultural political diversity, exclusion, complexity etc in Indian education have been nicely presented by [1]. The potential growth, and opportunities for business has been summed up by [2], [3], and [4]. India's language scene can be described as follows - a vast multilingual and multicultural country with 22 national but resource poor languages and more than 1600 other languages with less that10 % English knowing population and significant illiteracy. This makes India a land of opportunities in the area of using language technology in education. Of the 22 scheduled languages, Hindi has the dual status of National and Official Language (NOL) and English as Associate Official Language (AOL).

Revised Manuscript Received on October 15, 2019

Ritu Nidhi, scholar at Amity University, Noida, India
Dr. Tanya Singh, Professor and Dy. Director (Academics), Amity School of Engineering & Technology, Amity University, Uttar Pradesh.
Dr. DK Lobiyal, Professor, School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi.

All other 21 scheduled Indian languages enjoy national status but do not have sufficient language technology resources to compete with English. Factors like variability, language mixing, rich morphology, weak syntax, lack of technology standards have contributed to poor development of computer technologies in Indian languages. Maithili is one of the 22 schedule Indian languages spoken in the state of Bihar and neighboring Nepal and is extremely resource poor in terms of language technology resources. Lack of digital resources has resulted in a near absence of computer technologies in this language.

A.Landscape of Maithili

Maithili (ISO 639-3) is an Indo Aryan language spoken in Bihar and Nepal. It is also spoken in neighboring states of Jharkhand and West Bengal. As per the census of India (2011), we have more than 13 million Maithili speakers in India alone. Maithili is also spoken in the neighboring country Nepal where it is the second official language and the second most spoken language. A rough estimate puts the total number of Maithili speakers in the world to be around 34 million[1]. Popularly written in Devanagari, Maithili has its own script called Mithilakshara. Maithili literature dates back to 8th century AD and has at least six varieties [5].

Maithili shows typical linguistic features of an eastern Indo Aryan language. With a rich inflection system, absence of agreement at the level of verb and reasonable scrambling due to weaker syntax. Maithili has its unique features like varying layers of pronominal and verbal honorifics. Maithili has typical Indo Aryan Subject-Object-Verb (SOV) word order, however flexibility in moving syntactic constituents with relative ease poses challenges for automated linguistic analyses [5].

Indo Aryan languages of north India can easily be contrasted with English which is a rigid syntax SVO language with little scope for scrambling. English also exhibits weaker morphology compared to Maithili. These and other contrasting features of the two languages create difficulties for automatic translation.

II. E-LEARNING IN INDIA

India's education is based on the colonial model established by Lord Macaulay during the British administration of India. Successive post independence governments in India have tried their best to adapt it to Indian diversity, but were not successful given the complexity and challenges therein. The question of languages, medium of instruction etc have been more difficult to implement in India than the content. Though the literacy rates have risen sharply, still there is a massive 26% population below this threshold.

The 2001 and 2011 Census Report published by Government of India present contrasting literacy figures.

[1] https://www.ethnologue.com/language/mai

Table-I: Literacy figures of Indian Population

Year	Literates (% of total population)	Illiterates (% of total population)
2001	65%	35%
2011	74%	26%

Source: Annual Report 2013-14, published by Ministry ofHRD, Government of India [6]

With one fourth of illiterate population speaking hundreds of languages, education becomes challenging. The technology needed will obviously be a range of language and speech tools that enable people to learn without having to be literate first.

A.Current statusof e-learning in India

The government of India is realizing this developmental block and has identified major initiatives in both e-education and language technology with its planning body called NITI Ayog spearheading efforts in consolidating previous work and planning new projects. Many government and non-government organization are involved in e-learning. National Programme on Technology Enhanced Learning (NPTEL), Indira Gandhi National Open University (IGNOU), Talk to a Teacher, Consortium for Educational Communication (CEC) , ePathshala, EDUSAT, Massive Open Online Course (MOOC) are some of the initiatives from the Indian government. Besides, there are a large number of private players which have taken initiatives in the e-learning area. The MHRD, Govt of India has started several key initiatives to promote anywhere anytime learning as follows [7].

e-pathshala - National Council of Educational Research and Training (NCERT) has developed e-pathshala to disseminate educational materials which includes textbooks, audio, video through web portal.

SWAYAM - The Study Web of Active Learning for Young Aspiring Minds (SWAYAM) isan online e-learning platform aninitiative of MHRD, India. Currently, MOOC offers more than 1000 courses which is listed on SWAYAM [8].

SWAYAM Prabha It provides 32 high quality educational national channels through DTH-TV. All channels curriculum based course content for secondary and senior secondary school

ShaGun portal is a web portal, which is a repository of newspaper article, photograph, videos etc. for elementary education.

National Repository of Open Educational Resources (NROER) is an open source where teachers and students can access digital educational content free.

National Digital Library (NDL) is also a repository which has features for single-window searchfacility. More than 150 lacs digital books are available through NDL.

ShalaKosh is a repository of school data. It is used for school academic administrative work.

Saransh has been developed and launched by CBSE in 2015 is a self-reviewed tool which serves as an interface for enhancing communication between schools and parents.It offers data driven analytical solution for tracking performance of students [9].

ICT in Education Curriculum for School System: NCERT has launched Information and Communication Technology (ICT) curriculum, which focuses on students to developed their capacity in ICT skills.

B. Major players in the industry

The industry has realized the opportunity that lies in digitized education in India and is investing liberally in the field.

BYJU's is one of the top e-leaning companies that offer educational technology for school students.

Educomp is an e-learning company that has 30 million learners across 65,000 schools in two decades. It is the largest education company in India.

IGNOU is the most popular public university offering distance learning in India. Run by the central government of India, it also happens to be the largest university in the world. It was established in 1985 has over 3 million active enrollment.

NIIT offers learning management, which develops e-learning platform and e-learning content.

Meritnation is an online portal for Indian school going kids. They provide very competitive live video lectures and interactive recorded videos.

Edukart offers 2000 courses in degree, diploma, certificate, entrance coaching and K12 categories. It is an online entrance coaching site that provides solutions to aspirants candidates.

III. NLP IN EDUCATION IN INDIA

NLP plays an ever bigger role in education today. It is being used to automate the tasks that were done manually earlier. Due to electronically available content, NLP in education is processing text and speech in useful manner [10].

There are several applications which come under NLP: Machine Translation (MT), Named Entity Recognition (NER), Optical Character Recognition (OCR), Parts-of-Speech Tagging (POS Tagging), Speech to Speech Translation, Text to Speech Translation etc.Cross lingual Information Access (CLIA) allows user to search query in the language he/she knows and the result will be given in different languages [11]. CLIA system has five sub systems: a. Input processing b. Search c. Processing of retrieved documents d. Output generation and e. UNL based search. Currently the CLIA systems have been developed only for a few Indian languages. A consortium of 11 institutions funded by Government of India, Ministry of Electronics & Information Technology (MEITY) was created for this purpose.

IV. ROLE OF MT AND LOCALIZATION IN INDIAN EDUCATION

MT driven localization is critical for delivering e-learning content in diverse India. There is a severe shortage of such platforms which can cater to diverse multilingual education needs in India. However, there have been some spade word in the area of MT as reported below –

A. Major developments in Indian language MT

Work done by Govt. funding

Indian government supports heavily for major scheduled languages and MT has been a prime concern considering the complex multilingual situation in India. Among major

supporters are the Ministry of Human Resource Development (MHRD) which supports projects in resource development through its nodal agencies like the Central Institute of Indian Languages (CIIL), Mysore and the University Grants Commission (UGC). The Ministry of Electronics and Information Technology (MEITY) through its flagship initiative Technology Development for Indian Languages (TDIL) supports technology development in major languages. The Department of Science & Technology (DST) of the Ministry of Science & Technologyalso funds technology development for Indian languages in application development and cognitive science related areas.

The TDIL initiative of MEITY sponsors projects to develop MT systems and related resources for Indian languages. In recent times, the following mission mode projects have been given to leading institutes in the country [12].

- English to Indian Languages Machine Translation System (E-ILMT) to CDAC, Pune
- English to Indian Languages Machine Translation System with Angla-Bharti Technology (E-ILMT-ABT) to IIT Kanpur
- Indian Language to Indian Language Machine Translation System: (ILMT) to IIIT Hyderabad
- Sanskrit-Hindi Machine Translation (SHMT) to a consortium of University of Hyderabad, JNU and five other universities/institutes

The following table presents a chronological list of various attempts at developing MT Systems for Indian language pairs including English since the late nineties by Indian academia (funded by either TDIL or other govt. agencies)

Table-II:MT Systems for Indian language pairs [13]

SN	System	Target Language	Year	Place	Method used	Domain
1	Anusaaraka [14]	Bengali, Kannada, Marathi, Punjabi, Telugu- Hindi	1995	IIT Kanpur	Direct based	General
2	Mantra [15]	English-Hindi, Gujarati, Telugu, Hindi-English, Bengali, Marathi	1997, 1999	C-DAC Pune	Transfer based	Office administration documents and Proceeding of Rajyasabaha
3	Anglabharti [16]	Hindi-English	2001	IIT Kanpur	Interlingua	General
4	Anubharti [17]	Hindi-English	1995, 2004	IIT Kanpur	Hybrid	General
5	MaTra [18]	English-Hindi	2004, 2006	CDAC, Pune	Transfer based	General
6	Anubaad [19]	English-Bangali	2000, 2004	CDAC, Kolkata	Example based	News
7	Sampark [20]	Punjabi-Hindi, Telugu-Tamil Urdu-Hindi, Telugu-Hindi	2009	IIIT, Hyderabad & consortium	Computational Paninian Grammar	General
8	Shiva and Shakti [21]	English-Hindi,Marathi, Telugu	2003	I.I.Sc. Bangalore and IIIT, Hyderabad	Transfer based	General
9	Google Translator	More than 100 languages (including foreign language)	2006	Google	Statistical based	
10	Hindi-Punjabi MT	Hindi – Punjabi	2009, 2011	University of Patiala	Direct based	
11	English-Telugu MT System	English – Telugu	2004		Rule based	General
12	Tamil-Hindi MT system	Tamil - Hindi	2009	AU-KBC, Chennai		

13	English-Kannada MT	English - Kannada	2009	University of Hyderabad	Transfer based	Government circulars & notices
14	Bing Translator	Around 60 languages (including foreign language)	2007	Microsoft	Statistical based	
15	Bengali-Hindi MT System	Bengali – Hindi	2009		Hybrid	
16	SaHiT (Sanskrit-Hindi Translator)	Sanskrit-Hindi	2017	JNU	Statistical (on MT Hub)	Simple Sanskrit prose
17	ESAT (English-Sanskrit Translator	English-Sanskrit	2018	JNU	Statistical (on MT Hub)	Simple Sanskrit prose

B. Indian language MT Systems by Industry

The few major players in private industries are Google translator, Microsoft Bing Translator, Linguee translation tool, SDL Trados Studio, Fluency Now, MemoQ, WordFast Pro. Google translator translates over 100 languages which includes 11 Indian and major foreign languages. [22] claims that more 500 million people use Google translator every day. A few years ago, Google translator launched app for Android and iOS which supports more than 100 languages and can translate 37 languages via photo, 32 via voice in "conversation mode", and 27 via real-time video in "augmented reality mode" [22]. It is widely used translator which supports text, speech, images and video.

Microsoft Bing Translatorwhich was launched in 2007 is based on Statistical Machine Translation (SMT). It supports 10 speech translation systems that currently feature in Skype Translator and Skype for Windows Desktop, and the Microsoft Translator Apps for iOS and Android. The Linguee translation tool combines a dictionary with a search engine, so that user can search for bilingual texts, words and expressions in different languages to check meanings and contextual translations. The SDL Trados Studio is the most recommended computer-assisted translation (CAT) tool. This software features TM, terminology, Machine Translation, and software localization. Fluency Now is a professional and a premium CAT tool and translation memory software created for individual freelancers. This tool is compatible with Windows, Linux and Mac operating systems. This tool also provides a in-built proof reading software and project and document statistics. MemoQ is a translation software designed for freelance translators and offers a number of powerful functions that enable users to reuse previous translations. MemoQ also helps in improving quality, check consistency and ensure the use of correct terminology. The Wordfast Pro is a standalone, multi-platform TM tool designed to improve the translation process for project managers and freelance translators.

V. ENGLISH-MAITHILI MT (EMMT) FOR EDUCATION

Historically, major MT efforts in India have been rule based depending heavily on hand crafted grammar and dictionaries. Some initial efforts in translating English to Indian languages used a Prolog based parser and mapped Indian language generation rules. Some of them even tried Paninian formalism for parsing the source language including English. The statistical methods have been tried recently and some systems have been developed. Most of these systems with the exceptions of those developed by Google and Microsoft are practically unusable and require significant improvements. Due to being an extremely less resourced language, digital content of Maithili is not available easily. The parallel English Maithili data was even more difficult to create. The source language (SL) English data was collected from various resources such as magazine, books and internet in different domains such as politics, cuisine, sports, Bollywood etc. This data then manually translated into Maithili (target language (TL)), aligned and validated to create a gold set using two methods - one to use Google translate to create Hindi output and then use a standalone Hindi-Maithili converter developed by the lead author to create English-Maithili data and then post edit it for quality, two to manually translate English into Maithili. The Maithili monolingual data was collected from similar domains as the English source data, cleaned and validated into the gold set.

MT Hub and Moses are both MT platforms for training SMT Systems for a pair of natural languages. The MT Hub has been developed by Microsoft to build translation systems and Moses is an open source software. Both the engines require parallel aligned corpora in the source and target language pairs along with the monolingual corpora for the target language.

The EMMT system training was done for 10,000 parallel aligned corpora for the English-Maithili pair and 10,000 sentence corpora for the target language.

The dataset used in these experiments were identical. The following section documents the results of both trainings -

Moses

BLEU = **6.18,** 30.2/6.6/2.2/0.9

Table-III: Training data

Brevity of penalty(BP)	Hypothetical length (MT output)	Reference data length	Ratio (hyp_len/ re_len)	BLEU Score	Duration
0.869	8815	10053	0.877	**6.18**	50minutes

The four figures "30.2/6.6/2.2/0.9" are the precisions of 1-grams, 2-grams, 3-grams, and 4-grams, respectively and the global scores (BLEU 6.18) is computed as the geometric mean of these precisions multiplied by the Brevity Penalty (BP) calculated according to the length ratio (ration).

MTHub

Details of the training are given below –

Table-IV: Training data

Type of the training data file	Extracted Sentence Count	Aligned Sentence Count	Used Sentence Count (training set)
Monolingual (Maithili)	13,048	-	13,023
Parallel (English-Maithili)	10,334	7,562	6,427
Parallel (Maithili target)	10,758	7,562	6,427

The results obtained may be tabulated as follows

Table-V: Training statistics

Extracted sentences	Aligned sentences	Used sentences	BLEU	Duration
23,382	7,562	19,450	**11.07**	1 hour 31 minutes

BLEU Score: 11.07

Language Pair: English to Maithili

VI. CHALLENGES

Despite a phenomenal growth in internet and mobile usage, the challenge of educating in native languages remains a big issue in India. The lack of resources and tools in Indian languages adds to this challenge because any e-learning application must integrate language technology applications to deliver content to non-English population of India (>90%). Therefore, for each of the major schedule languages of India, the priority should be to assess the educational environment and need and do sincere documentation of tools needed and tools available. The efforts should be (a) to evaluate existing tools and integrated in e-learning platforms. and (b) to develop missing key tools and technologies. As the case is, most of the Indian languages do not have the necessary resources and tools, so the primary focus of the government and industry should be to develop such resources on a priority basis.

VII. CONCLUSION AND FUTURE WORK

The paper has presented the current status of NLP applications in India's e-learning horizon and the challenges and opportunities in a socio-cultural, linguistically diverse nation with substantial illiteracy. The paper has also briefly surveyed current government and industry initiatives in e-learning and documented the status of MT research and applications in the country and their current and potential applications in e-education. The status of Maithili language in the technology space and the progress in the development of English-Maithili MT system for potential use in education have also been presented. The authors intend to work towards an integrated localization platform for Maithili with EMMT engine for use in adapting English content for education in Maithili.

REFERENCES

1. Susan Lynn, Gabel, Scot, Danforth, (2008), "Disability & the Politics of Education: An International Reader", Peter Lang.
2. Chatterjee, Shivaji (2014), "Why e-learning has a promising future in India", In: Satellite Evolution Asia, Nov-Dec 2014 (https://www.satellite-evolution.com/)
3. Pande1, D., Wadhai, V. M., Thakre, V.M. Thakre (2016), "Current trends of E-learning in India", International Research Journal of Engineering and Technology, Volume: 03 Issue: 01
4. Kandhari, M. M. (2018), "E-Learning Is Transforming the Face of Education In India", Business World
5. Nidhi, R., Singh, T. (2019), "Machine Translation and Divergence Study for English-Maithili",Advances in Intelligent Systems and Computing, Springer, book series (AISC, volume 933)
6. Government of India. (2011). Census Report, 223.
7. http://pib.nic.in/
8. https://swayam.gov.in/about
9. https://en.wikipedia.org/wiki/Saransh
10. Litman, Diane,(2016),"Natural Language Processing for Enhancing Teaching and Learning", In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)
11. Cross Lingual Information Access System for Indian Languages CLIA Consortium,(2006) (http://www.mt-archive.info/IJCNLP-2008-CLIA.pdf)
12. Sinha Prachi, Bhowmick Priti, Jha, Girish N, (2009) Linguistic pedagogy and e-learning: the case for Sanskrit, in "Contemporary Themes and Issues in Language Pedagogy" vol2.
13. Nidhi, R., Singh, T. (2019), "SMT algorithms for Indian languages - A case study of Moses and MT Hub for English-Maithili language pair", presented a paper in International Conference On Emerging Trends in Information Technology (ICETIT-2019)
14. Bharati, A., Chaitanya, V., Kulkarni, A.P., Sangal, R., (1997), Anusaaraka: Machine translation in stages, VIVEK-BOMBAY"-, vol. 10, pp. 22–25
15. Darbari, H. (1999),"Computer-assisted translation system–an Indian perspective", In: Machine Translation Summit VII, 13th-17th September, pp. 80–85
16. Sinha, R., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., Jain, A.(1995),"Anglabharti: a multilingual machine aided translation project on translation from english to indian languages, in Systems, Manand Cybernetics, Intelligent Systems for the 21st Century", In: IEEE International Conference on, vol. 2. IEEE, pp. 1609–1614
17. Jain, R., Sinha, R., Jain, A.(2001), "Anubharti-using hybrid example-based approach for machine translation", STRANS-2001, IIT Kanpur, pp. 20–32
18. Ananthakrishnan, R., Kavitha, M., Jayprasad, J.H., Chandra Shekhar, R.S., Sawani Bade, S.M. (2006), "Matra: A practical approach to fully-automatic indicative english-hindi machine translation in Symposium on Modeling and Shallow Parsing of Indian Languages"
19. Bandyopadhyay, S. (2000),"Anubaad-the translator from English to Indian languages", In:Proceedings of the 7th State Science and Technology Congress, Calcutta, India, pp. 1–9
20. Ahmad, R., Kumar, P., Rambabu, B., Sajja, P., Sinha, M.K., Sangal, R.(2011), "Enhancing throughput of a machine translation system using mapreduce framework: An engineering approach", In: International Conference on

Natural Language Processing(ICON)
21. Sangal, R. (2004),"Architecture of shakti machine translation system", IIIT Hyderabad
22. https://en.wikipedia.org/wiki/Google_Translate

AUTHORS PROFILE

Ritu Nidhi is a research scholar at Amity University, Noida, India doing her doctorate in the Amity Institute of Information Technology. Her research interests include Machine Translation, localization, Indian language content creation, web development and standards. Technology development for resource poor languages of India is her passion.

Dr. Tanya Singh is a Professor and Dy. Director (Academics), Amity School of Engineering & Technology, Amity University, Uttar Pradesh. She has experience of 20yrs in the field of Teaching, Research, Planning and Development in Education and Operational role outs. She has emerged as Technical Evangelist for Networking, Cyber Security. She is a member of Bureau of Indian Standards (BIS), IEEE, IEC, IETE, ISOC. She has two patents and published three books. She has publications in National and International journals and also has several national and international conference paper. She has developed audio video material for teaching in Amity University. She has been awarded by CISCO, ACM and IETE.

Dr. DK Lobiyal is a Professor, School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi. His research areas include Wireless Sensor Networks, Mobile ad-hoc Networks and Natural Language Processing. He has published more than 90 publications in reputed national/international journals. He has written a book chapter in Advances in Experimental and Medical Biology, PubMed, Springer Verlag. He has more than 70 conference papers in India and abroad. He is a member of Computer Society of India (CSI) and Institution of Electrical, Electronics Engineers (IEEE) and Fellow of Institution of Electronic and Telecommunication Engineers (IETE), India. More details on Prof Lobiyal can be found at http://www.jnu.ac.in/Faculty/dkl/cv.pdf