Survey on Machine Translation Approaches used in India

D S Rawat

Abstract— **Machine Translation is the branch of Natural Language Processing, which deals the use of software to convert from one natural language to another natural language. The work in the field of Machine Translation (MT) has been going on to understand each others languages, so that one not knowing others language can also understand his/her views with the help of this MT software. Here in this work we have surveyed the different MT approaches in India. Due to linguistic diversity in India, it is very difficult to know each others views. It is necessary to know each other views without more efforts or without learning others language. For that there is a variety of software or tools are available which are called Machine translation tools. Here in this paper, we have tried to give an overview of Machine Translation Systems in India which is built for the purpose of translation between the different Indian languages.**

Index Terms— **Example Based, Machine Translation, Natural Language Processing, Rule Based, Statistical Based.**

I. INTRODUCTION

Natural language processing (NLP) is an area concerned with the interactions between computers and human (natural) languages. Many challenges in NLP involve Automatic summarization, Discourse analysis, Information retrieval, Information extraction, Machine translation, Morphological segmentation, Natural language generation, Natural language understanding, Optical character recognition, Part-of-speech tagging, Parsing, Question answering, Relationship extraction, Sentiment analysis, Speech recognition, Word sense disambiguation and so on[9].

The term Machine Translation (MT) is a standard name for the use of computers to automate some or all the process of translating from one natural language to another. Translation, in its full generality, is a difficult, fascinating, and intensely human endeavor, as rich as any other area of human creativity [10].

MT systems can be designed either specifically for two particular languages called bilingual system, or for more than a single pair of languages called multilingual systems. Bilingual system may be either unidirectional, from one Source Language (SL) into one Target Language (TL), or bidirectional. MT methodologies are commonly categorized as Direct, Rule based, Hybrid, Example based and Statistical. The MT methodologies differ in the depth of analysis of the source language and the extent to which they attempt to reach a language independent representation of meaning or intent between the source and target languages. All over the world many attempts are being made to develop MT systems for

D S Rawat, Department of CS&IT, Amrapali Institute of Technology & Sciences, Haldwani, Nainital (Uttrakhand), India, Mobile no. 9012020089.

various languages using different above said approaches. Development of a well fledged Bilingual Machine Translation system for any two natural languages with limited electronic resources and tools is a challenging and demanding task. In order to achieve a reasonable translation quality in open source tasks, Statistical and Example based MT approaches require large amounts of parallel corpus which are not always available, especially for less resourced language pairs. On the other hand the rule based MT process is extremely time consuming, difficult and failed to analyze accurately a large corpus of unrestricted text.

II. MT DEVELOPMENT IN WORLD

The history of MT started with philosopher Leibniz and Descartes ideas of using code to relate words between languages in the seventeenth century [17]. An overview of the earlier works on MT can be seen in [17] and [18].

After the birth of computers (ENIAC-Electrical Numerical Integrator And Calculator) in 1947, research began on using computers as aids for translating natural languages [19]. Further research in this field is thrust by the demonstration of MT in the Georgetown-IBM experiment. In the year 1966 Automated Language Processing Advisory Committee (ALPAC) has submitted a report on MT progress that MT was waste of time and money [11]. This report brought MT research to halt, suspending virtually all research in the USA while some research continued in Canada, France and Germany [19]. Since after the ALPAC report MT research work was almost down from 1966-1980. In the year 1988, Georgetown-IBM experiment launched "IBM CANDIDE System", where over 60 Russian sentences were translated smoothly into English using 6 rules and a bilingual dictionary consisting of 250 Russian words, with rule-signs assigned to words with more than one meaning. Although Professor Leon Dostert cautioned that this experimental demonstration was only a scientific sample, or "a Kitty Hawk of electronic translation" [20].

After 1980 a large number of MT systems emerged from various countries while research continued on more advanced methods and techniques. Those systems mostly comprised of indirect translations or used an Interlingua (IL) as its intermediate. Statistical Machine Translation (SMT) was emerged in year 1990 and what is now known as Example Based Machine Translation (EBMT) saw the light of day [16]. At this time the focus of MT began to shift somewhat from pure research to practical application using hybrid approach. In the year 1993 the project Consortium for Speech Translation Advanced Research (C-STAR) was started. The system was trilingual project and defined for the tourism domain. In the year 2005 the Google launched a first website for automatic translation [11]. With this the new millennium, MT became more readily available to individuals via online services as well as through software for their use. In the year

2009 Bing translator by Microsoft and in June 2014 Google Translate's 37th stage was launched.

III. DEVELOPMENT IN INDIA

MT works in India reveals references of translation works in Hindi or other Indian regional languages. The earliest published work was undertaken by Chakraborty in 1966[12]. Many governmental, non governmental private sectors as well as individuals are actively involved in the development of MT system and have already generated some reasonable MT system. The main developments are as under.

In the Direct approach MT system in India first attempt was done by the Rajeev Sahgal in IIT Kanpur in the year 1995, further this is continued by IIIT Hyderabad. The purpose of this project was the MT of one Indian language to another Indian language. It uses a Paninian Grammar (PG) and exploits the close similarity of Indian languages [1][2]. In the year 2007-08 G S Josan and G S Lehal developed a system which is based on direct word-to-word MT approach from Punjabi to Hindi[13][38]. V Goyal and G S Lehal developed the extended version of Hindi to Punjabi MT System in the year 2010[44]. Again, same group developed a system that uses direct word to word translation approach for Hindi to Punjabi at Punjabi University, Patiala in 2011[14][36][43].

First Rule based MT system Mantra English to Hindi MT system was developed by Bharati in year 1997 for information preservation. The text available in one Indian language has been made accessible in another Indian language with the help of this system [37]. The system has several facilities like website translation, email translation, etc. [6]. Hemant Darbari and Mahendra Kumar Pandey in year 1999 developed a MAchiNe assisted TRAnslation tool (MANTRA)[15][37]. It has the facility of translating English text into Hindi in a specific domain of personal administration that includes gazette notifications, office orders, office memorandums and circulars. L Gore and N Patil developed a system on transfer based MT approach, which uses different grammatical rules of source and target languages and a bilingual dictionary for translation from English to Hindi in year 2002[23]. In the same year K Murthy developed MAT (Machine Assisted Translation) system for translating English texts into Kannada, which used morphological analyzer and generator for Kannada[26]. After one year in 2003 Bharati, R Moona, P Reddy, B Sankar, D M Sharma and R Sangal have developed a system named Shakti which translates English to any Indian languages with simple system architecture[7]. It combines linguistic rule-based approach with statistical approach. Next year S Bandyopadhyay developed two systems one is English-Telugu and another is Telugu-Tamil[27]. Same year S Mohanty, R C Balabantaray developed a system that translates text from English to Oriya based on grammar and semantics of the source and target language[6][40].

In the year 2004 and 2006 MaTra System came for the English to Hindi[3][4][8]. In the year 2009 English-Kannada machine-aided translation system [42][24] and Tamil-Hindi Machine-Aided Translation system [32][36][42] came into existence. Same year a consortium of 11 institutions in India have developed a multipart machine translation system for Indian Language to Indian Language Machine Translation (ILMT) funded by TDIL (Technology Development for Indian Languages) program of Department of Electronics and Information Technology, Govt. of India [33].

Interlingua Rule based MT systems are ANGLABHARTI [42], UNL(Universal Networking Language)-based[25][34][35][41]English-Hindi MT System. Both were developed in year 2001. Whereas AnglaHindi is a derivative of AnglaBharti MT System developed by R M K Sinha and A Jain for English to Indian languages in year 2003[31].

Main Hybrid MT systems are Anubharti, ANUBHARTI-II, which were developed in year 2004[34][28].

S Bandyopadhyay developed an MT system which translates news headlines from English to Bengali using Example based Machine Translation approach in year 2000 and 2004[37][39]. In the year 2002 K Vijayanand, S I Choudhury and P Ratna developed an Automatic Machine Translation system for Bengali-Assamese News Texts with using the same above approach named VAASAANUBAADA [21]. MT system Shiva is designed using an Example-based and Shakti is designed using the combination of rule based and statistical based approaches. The Shakti system is working for three target languages like Hindi, Marathi and Telgu and can produce machine translation systems for new languages rapidly. Shiva & Shakti are the two Machine Translation systems from English to Hindi developed jointly by CMU, IIIT, Hyderabad and IISc, Bangalore. The system is used for translating English sentences into an appropriate target Indian language. In the year 2004 ANGLABHARTI-II and Hinglish MT System were developed in the same category [30][34][42].The MATREX(MT using Example)is developed by Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar and Andy Way using the marker based chunking in year 2008[5][45].

Statistical MT system Shakti was developed by Bharati, R Moona, P Reddy, B Sankar, D M Sharma and R Sangal in year 2003, which translates English text to any Indian language with simple system architecture[34][42]. English to Indian Languages MT System (E-ILMT) is a MT System for English to Indian Languages in Tourism and Healthcare fields. It is developed by a collective efforts of Nine institutions namely C-DAC Mumbai, IISc Bangalore, IIIT Hyderabad, C-DAC Pune, IIT Mumbai, Jadavpur University Kolkata, IIIT Allahabad, Utkal University Bangalore, Amrita University Coimbatore and Banasthali Vidyapeeth Banasthali[29]. In the year 2014 Kunal Sachdeva, Rishabh Srivastava, Sambhav Jain and Dipti Misra Sharma of IIIT Hyderabad have given a idea of Hindi to English MT system by training a regression Model in the statistical based Machine Translation [22].

IV. CONCLUSION

This paper tells the development done in the field of Machine translation world-wide and especially with context to the Indian languages. Also we have given the various standardized approaches for machine translation. This paper will be useful for new researchers to understand the development done in the field of the Machine Translation, so that they can enhance the methods and do the more useful to take the all mankind close to each other.

REFERENCES

[1] Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, (1997) "ANUSAARAKA:Machine Translation in stages", *Vivek, a quarterly in Artificial Intelligence*, Vol. 10, No. 3, NCST Mumbai, pp. 22-25

[2] Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, (2001) "ANUSAARAKA: Overcoming the language barrier in India", *published in Anuvad: approaches to Translation.*

[3] Ananthakrishnan R, Kavitha M, Jayprasad J Hegde, Chandra Shekhar, Ritesh Shah, Sawani Bade & Sasikumar M., (2006) "MaTra: A Practical Approach to Fully- Automatic Indicative English- Hindi Machine Translation", *In the proceedings of MSPIL-06.*

[4] Ananthakrishnan R, Kavitha M, Jayprasad J Hegde, Chandra Shekhar, Ritesh Shah, Sawani Bade & Sasikumar M, (2006) "MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation", *in proceedings of the first national symposium on Modelling and shallow parsing of Indian languages (MSPIL-06* .

[5] Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar & Andy Way, (2008) "The MATREX (Machine Translation using Example): The DCU Machine Translation System for ICON 2008", *in Proceedings of ICON-2008: 6th International Conference on Natural Language Processing*, Macmillan Publishers, India.

[6] Antony P. J., (2013) "Machine Translation Approaches and Survey for Indian Languages", *International journal of Computational Linguistics and Chinese Language Processing* Vol. 18, No. 1, pp. 47-78.

[7] Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal, (2003) "Machine Translation: The Shakti Approach", *Pre-Conference Tutorial, ICON-2003.*

[8] CDAC Mumbai, (2008) "MaTra: an English to Hindi Machine Translation System", a report by CDAC Mumbai formerly NCST.

[9] Christopher D. Manning and Hinrich Schutze, (1999). "Foundations of Statistical Natural Language Processing", MIT Press.

[10] Danial Jurafsky & James H. Martin, (2005) "Speech and Language processing", Pearson Education.

[11] D.V. Sindhu, B.M. Sagar, S Rajashekar Murthy, (2014) "Survey on Machine Translation and its Approaches", *International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014.*

[12] Grace Noone, (2003) "Machine Translation -A Transfer Approach, A project report", www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/nooneg0203 .pdf,.

[13] G. S. Josan & G. S. Lehal, (2008) "A Punjabi to Hindi Machine Translation System", *in proceedings of COLING-2008*: Companion volume: Posters and Demonstrations, Manchester, UK, pp. 157-160.

[14] Gurpreet Singh Josan & Jagroop Kaur, (2011) "Punjabi To Hindi Statistical Machine Transliteration", *International Journal of Information Technology and Knowledge Management*, Volume 4, No. 2, pp. 459-463.

[15] Hemant Darabari, (1999) "Computer Assisted Translation System- An Indian Perspective", *in proceedings of MT Summit VII*, Thailand.

[16] http://www-03.ibm.com/ibm/history/exhibits/701/701-translator.html

[17] Hutchins J, (1993) "The first MT patents," MT News International, pp.14-15.

[18] Hutchins, W. J. and Lovtskii, E., (2000) "Petr Petrovich Troyanskii (1854-1950): A forgotten pioneer of mechanical translation", published in Machine translation, vol.15, no.3, pp.187-221.

[19] Hutchins J, (2005)"The history of machine translation in a nutshell," http://www.hutchinsweb.me.uk/Nutshell-2005.pdf,.

[20] John Hutchins, "The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954", www.hutchinsweb.me.uk/GU-IBM-2005.pdf.

[21] Kommaluri Vijayanand, Sirajul Islam Choudhury & Pranab Ratna, (2002) "VAASAANUBAADA - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts", *in proceedings of Language Engineering Conference-2002*, Hyderabad, India © IEEE Computer Society.

[22] Kunal Sachdeva, Rishabh Srivastava, Sambhav Jain, Dipti Misra Sharma,(2014)"Hindi to English machine translation: Using effective selection in multi-model SMT", *9th International Conference on Language Resources and Evaluation, LREC 2014,*

[23] Lata Gore & Nishigandha Patil, (2002) "English to Hindi - Translation System", *In proceedings of Symposium on Translation Support Systems*. IIT Kanpur. pp. 178-184.

[24] Latha R. Nair & David Peter S., (2012) "Machine Translation Systems for Indian Languages", *International Journal of Computer Applications* (0975 – 8887) Volume 39– No.1.

[25] Manoj Jain & Om P. Damani, (2009) "English to UNL (Interlingua) Enconversion", *in proceedings of 4th Language and Translation Conference (LTC-*09).

[26] Murthy. K, (2002) "MAT: A Machine Assisted Translation System", *In Proceedings of Symposium on Translation Support System(STRANS-2002)*, IIT Kanpur. pp. 134-139.

[27] Parameswari K, Sreenivasulu N.V., Uma Maheshwar Rao G & Christopher M, (2012) "Development of Telugu-Tamil Bidirectional Machine Translation System: A special focus on case divergence", *in proceedings of 11th International Tamil Internet conference*, pp 180-191.

[28]Projects.uptuwatch.com/cs-it/anubharti-an-hybrid-example-based-approach-for-machine-aidedtrapnslation/

[29] R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya, Ritesh Shah & M. Sasikumar, (2008) "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation", *in proceedings of International Joint Conference on NLP (IJCNLP08),* Hyderabad, India.

[30] R. Mahesh K. Sinha & Anil Thakur, (2005) "Machine Translation of Bi-lingual Hindi-English (Hinglish) Text", *in proceedings of 10th Machine Translation Summit* organized by Asia-Pacific Association for Machine Translation (AAMT), Phuket, Thailand.

[31] R.M.K. Sinha & A. Jain, (2002) "AnglaHindi: An English to Hindi Machine-Aided Translation System", *International Conference AMTA(Association of Machine Translation in the Americas).*

[32] Salil Badodekar, (2004) "Translation Resources, Services and Tools for Indian Languages", *a report of Centre for Indian Language Technology, IITB*, http://www.cfilt.iitb.ac.in/Translationsurvey/survey.pdf

[33] Sampark: Machine Translation System among Indian languages (2009) http://tdildc. in/index.php?option=com_vertical&parentid=74, http://sampark.iiit.ac.in/

[34] Sanjay Kumar Dwivedi & Pramod Premdas Sukhadeve, (2010) "Machine Translation System in Indian Perspectives", *Journal of Computer Science* 6 (10): 1082-1087, ISSN 1549-3636, © 2010 Science

[35] Shachi Dave, Jignashu Parikh & Pushpak Bhattacharyya, (2002) "Interlingua-based English-Hindi Machine Translation and Language Divergence", *Journal of Machine Translation*, pp. 251-304.

[36] Sitender & Seema Bawa, (2012) "Survey of Indian Machine Translation Systems", *International Journal Computer Science and Technolgy*, Vol. 3, Issue 1, pp. 286-290, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

[37] Sudip Naskar & Shivaji Bandyopadhyay, (2005) "Use of Machine Translation in India: Current status" *AAMT Journal*, pp. 25-31.

[38] Sugata Sanyal & Rajdeep Borgohain, (2013) "Machine Translation Systems in India", *Cornel University Library*, arxiv.org/ftp/arxiv/papers/1304/1304.7728.pdf

[39] S. Bandyopadhyay, (2004) "ANUBAAD - The Translator from English to Indian Languages", *in proceedings of the VIIth State Science and Technology Congress*. Calcutta. India. pp. 43-51

[40] S. Mohanty & R. C. Balabantaray, (2004) "English to Oriya Translation System (OMTrans)" cs.pitt.edu/chang/cpol/c087.pdf

[41] Smriti Singh, Mrugank Dalal, Vishal Vachhani, Pushpak Bhattacharyya & Om P. Damani, (2007) "Hindi Generation from Interlingua (UNL)", *in proceedings of MT Summit, 2007*

[42] Vishal Goyal & Gurpreet Singh Lehal, (2009) "Advances in Machine Translation Systems", *National Open Access Journal,* Volume 9, ISSN 1930-2940

[43] Vishal Goyal & Gurpreet Singh Lehal, (2011) "Hindi to Punjabi Machine Translation System", *in proceedings of the ACL-HLT 2011 System Demonstrations*, pages 1–6, Portland, Oregon, USA, 21 June 2011.

[44] Vishal Goyal & Gurpreet Singh Lehal, (2010) "Web Based Hindi to Punjabi Machine Translation System", *International Journal of Emerging Technologies in Web Intelligence*, Vol. 2, no. 2, pp.148-151, ACADEMY PUBLISHER

[45] Yanjun Ma, John Tinsley, Hany Hassan, Jinhua Du & Andy Way, (2008) "Exploiting Alignment Techniques in MATREX: the DCU Machine Translation System for IWSLT 2008', *in proceedings of IWSLT 2008*, Hawaii, USA

D.S.Rawat was born in Uttrakhand, India, in 1972. He received the A.M.I.E. in Computer engineering from Institution of Engineers(India), 8 Gokhle Road, Kolkata in 2004, and the M.E.(Master of Engineering) in Computer Science & Engineering from Faculty of Engineering, M.B.M. Engineering College, Jai Narayan Vyas University, Jodhpur(Rajasthan) and presently perusing Ph.D. form Kumaon University, Nainital(Uttrakhand) in Information Technology.

In 2011 after leaving Indian Armed forces, he joined the Department of Computer Science & Information Technology, Amrapali Institute of Technology & Sciences, Haldwani, Nainital (Uttrakhand), as an Assistant Professor.

His current research interests include Natural Language Processing, ANN & Fuzzy logic, Artificial Intelligence, Theory of Automata & Formal Languages, Wireless Networks, Semantic Web, Discrete Structure and Data structure. Mr. Rawat is an Associate Member of Institution of Engineers (India).