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 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 emerging 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

D S Rawat, Department of CSAIT, Amrapali Institute of Technology & Sciences, Haldwani, Nainital (Uttarakhand), India, Mobile no. 9012020089.
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 Durbari and Mahendra Kumar Pandey in year 1999 developed a MAChiNe assisted TRAnslatation 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 gazette notifications, office orders, office memorandums and circulars. L Gore and N Patil developed a system on transfer gazette notifications, office orders, office memorandums and circulars. L Gore and N Patil developed a system on transfer gazette notifications, office orders, office memorandums and circulars.

In the year 2004 and 2006 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 Anglabharti 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 Bandypadhyay 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 Telugu 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 IISe, 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)s 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.
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 pursuing Ph.D. from Kumaun University, Nainital (Uttarakhand) 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 (Uttarakhand), 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).