INDIAN MACHINE TRANSLATION SYSTEMS

[1]Dr. Sushil Kumar, [2]Ms. Parvin Akhter
[1]Department of Electrical & Electronics, Pragati College Of Engineering & Management,Raipur
[2]Department of Electronics & Communication, Ram Krishnan Dharmath Foundation,Bhopal
Email-[1]sklbit@rediffmail.com,[2]parvin.akhter90@gmail.com

Abstract—This paper gives a survey of the work done on various Indian machine translation systems either developed or under the development. Some systems are of general domain, but most of the systems have their own particular domains like parliamentary documents translation, news readings, children stories, web based information retrieval etc.

Index Terms— Machine translation, computational linguistics, language processing

I. INTRODUCTION

Indian is the largest democratic country in the world and there are more than 30 languages and approximately 2000 dialects used for the communication by the Indian peoples and out of these languages Hindi and English are taken as language for of cial work and there are 22scheduled languages used by the different states for their administrative work and communication purposes. These 22 languages includes Assames, Bengali, Bodo, Dogri, Gujrati, Hindi, Malayalam, Manipuri, Marathi, Nepali,Oriya, Punjabi, Sanskrit, Kannada, Kashmiri, Konkani, Maithili, Santali, Sindhi, Tamil, Telugu, Urdu. Because of different culture and multilingual environment in India there is a big requirement for inter-language translation for the transfer of information and sharing of the ideas. Peoples of different states perform their work in their respective regional languages whereas the work at the Union Government of ces is performed in English language which is assumed to be one of the most speaking languages in the world or Hindi Language. So to synchronize between state government and the central / Union government there is a need for translation from regional languages to English language and vice versa. In India because of different culture there is different news papers published locally as well as globally. From the above discussion it is clear that there is large scope of translation of text from English to Indian Languages and vice versa. The initial work on Indian Machine Translation (in the beginning of 90's) was performed at various locations by different persons like IIT Kanpur,

Computer and Information Science department of Hyderabad, NCST Mumbai, CDAC Pune, department of IT, Ministry of Communication and IT Government of India. In the mid 90's and late 90's some more machine translation projects also started at IIT Bombay, IIT Hyderabad, department of computer science and Engineering Jadavpur University, Kolkata, JNU New Delhi etc. The next part of the paper gives a brief introduction of the various machine translation works done so far although there are some advancement is going on some projects so the latest information may be taken from the respective websites. The next section is divided into ve sub-sections A to E for the different languages.

II. MACHINE TRANSLATION SYSTEM

The development of the Indian Machine Translation system can be divided into different categories. The scope of this paper is restricted to Hindi, Punjabi, Sanskrit, Bengali and English Language as a source language.

A. Anusaaraka

A project named "ANUSAARAKAA" for machine translation from one Indian Language to

another Language in 1995. It has been used for translation from Telugu, Kannada, Bengali, Punjabi and Marathi to Hindi language translation and vice versa. The ANUSAARAKA system is a language accessor cum Machine Translation system that works on principles of Paninian Grammar (PG)". The system provides both the robustness incase of failure and no loss of information while translating the text. The output of the system follows the grammar of the source language. The approach for the translation in this system is divided in two parts: 1) The ANUSAARAKA system which is based on language knowledge 2) the domain specific knowledge based on world knowledge, statistical knowledge etc. It was started at IIT Kanpur and now shifted to IIIT Hyderabad Currently ANUSAARAKA is working for Telugu, Kannada, Marathi, Bengali, and Punjabi to Hindi language translation and in near future reverse translation will also be feasible.

B. The Mantra(Machine assisted Translation Tool)

A machine translating system named "MANTRA" which translates the text from English to Hindi language with a precise domain in Office order, administrative work texts etc.in1999. The basis of this system was the Tree Adjoining Grammar(TAG) formalism from the University of Pennsylvania. It uses Lexicalized Tree Adjoining Grammar (LTAG) for representing the English and the Hindi Language. It uses the TAG for parsing as well as Generation purposes. Now this system is also used in the nance, agriculture, health care, information technology, education and the general purpose activities of the government domains. The system named 'MANTRA-RAJYSABHA' is developed for the RAJYASABHA purposes. Currently the work for the language pairs English-Bengali, English-Telugu, English-Gujarati, Hindi-English, Hindi-Marathi, Hindi-Bengali is also going on.

C. Anubharti-II Technology

A system with an approach for machine aided translation having the combination of example-based and corpus based approaches and some elementary grammatical analysis. In ANUBHARTI the traditional EBMT approach has been modified to reduce the requirement of a large example base. ANUBHARTI-II in 2004 uses Hindi as a source language for translation to other Indian language [8].

D. Hindi Generation from Interlingua

Prof. Pushpak Bhattacharyya and Prof. Om. P. Damani reported a work on Hindi generation (Hindi Deconverter) from UNL graphs with the satisfactory results. The linguistic concerns have been clearly separated from the computational tasks which results in the possibility of the generation of the other languages also [31].

E. Punjabi to Hindi Machine Translation System

In 2007 a system based on direct word-word translation for machine translation between Punjabi as source language and Hindi as target Language was proposed .The system has reported 92.8% accuracy [24].

F. Sanskrit-Hindi Anusaarka

In 2009 a language accessor cum machine translation system for Sanskrit -Hindi language pair by following the Anusaarkaapproach was proposed and it allows the user to access the source language text and give the rough output in the target language. The translation mechanism was transparent to the end user [27].

G. Constrained Based Parser for Sanskrit Languaga

In 2010, Dr. Amba Kulkarni, SheetalPokar, DevanandShukl designed A Constrained Based Parser for Sanskrit Language at University of Hyderabad in the department of Sanskrit Studies. Based on the designing principles obtained from the generative grammars the parser was modeled for nding the directed tress, from the graph with the nodes as words and edges showing the relations between the words. To rule out the non-solutions they used Mimamsa constraints of Akanksa and Sannidhi to give the priority the solutions. The current system allows only the limited and simple sentences to be parsed [32].
ANGLABHARTI: A machine translation system at IIT Kanpur in 1991 was developed which translates from English to Indian Languages. The concept of pseudo- Interlingua has been used for the development of the machine aided translation system named" ANGLABHARTI "in 1991. This concept exploits the commonality in the Indian Languages. ANLGABHARTI is based on Rule Based Translation System (RBTS) with context free grammar structure of the English language as a

source language and produces a pseudo Interlingua code which is applicable to a Group of Indian Languages. The movement rules are obtained by the corpus analysis and the target constituents are obtained from this analysis. The aim of the ANGLABHARTI was to provide a translation system in which 90% work will be done by the machine and the 10% post editing work will be done by the human [1].

H. AnglaBharti Technology

The AnglaBharti project was launched by Sinha et al. (2001) at the Indian Institute of Technology; Kanpur in 1991 for Machine aided Translation from English to Indian languages. Professor Sinha et al. (2001) has pioneered Machine Translation research in India. The approach and lexicon of the system is general-purpose with provision for domaincustomization. A machine aided translation systemspecifically designed for translating English to Indianlanguages. English is a SVO language while Indianlanguages are SOV and are relatively offered word-order. Instead of designing translators for English toeach Indian language, Angla Bharti uses a (Daveetal., 2001) pseudo-interlingua approach. It analyses English only once and creates an intermediate structure called Pseudo Lingua for Indian Languages.

In AnglaBharti they use rule based system with context free grammar like structure for English, A set of rules obtained through corpus analysis which is used to distinguish conceivable constituents. Overall, the attempts to generalizing the constituents and replacing them with abstracted form from the raw examples. The abstraction integrate example-based approach with rule-based and human engineered post-editing.

AnglaBharti is a pattern directed rule based system with context free grammar (Sinha and Jain, 2003) like structure for English (source language) which generates a 'pseudo-target' (PLIL) applicable to a group of Indian languages (target languages). A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the PLIL is constructed. The idea of using PLIL is primarily to exploit structural similarity to obtain advantages similar to that of using Interlingua approach. It also uses some example-base to identify noun and verb phrasal's and resolve their ambiguities.

Sr. No.	Name of the system	Languages for Translation	Approaches Used	Domain	Year
1	ANGLABHARTI-1 (IIT K)	ENG-IL	Pseudo-interlingua	General	1991
2	ANUSAARAKA (U H)	IL-IL	PG	General	1195
3	MANTRA (C-DAC- P)	ENG-IL4 HINDI-EMB	TAG	Administration, of ce orders	1991
4	VAASAANUBAADA (A U)	BENGALI-ASSAMESE	EBMT	NEWS	2002
5	ANGLABHARTI-II (IIT-K)	ENGLISH-IL	GEBMT	General	2004
6	ANUBHARTI-II (IIT-K)	HINDI-IL	GEBMT	General	2004
7	MATRA CDAC-M	ENGLISH-HINDI	Transfer based	General	2004
8	SHIVA & SHAKTI (IIIT-H, IIS-B)	ENG-IL3	EBMT & RBMT	General	2004
9	UNL MTS (IIT-B)	ENG-HINDI	Interlingua	General	2003
10	ANUBAD (J U)	ENG-BENGALI	RBMT AND SMT	NEWS	2004
11	HINGLISH (IIT-K)	HINDI-ENG	Pseudo interlingua	General	2004
12	ANUVAADAK (SUPER INFOSOFT)	ENG-IL	Not-Available	Not-Available	
13	PUNJABI-HINDI (P U)	PUNJABI-HINDI	Direct word to word	General	2007
14	SAMPARK (IIIT-H, CDAC-N ,IIT-KGP, ANNA-U	IL-IL IL5	CPG	Not-Available	2009
15	IBM MTS	ENG-HINDI	EBMT & SMT	Not-Available	2006

IL4 Hindi, Bengali, Telugu, Gujarati
IIT-B Indian Institute of Technology Bombay
IIT-KGP Indian Institute of Technology Kharagpur
IIT-K Indian Institute of Technology Kanpur
UH University of Hyderabad
PU Punjabi University
CDAC-N CDAC Noida
IL IndianLanguage
JU Jadavpur University
ANNA-U Anna University
IIS-B Indian Institute of Sciences Bangalore
EBMT Example Based Machine Translation
ENG English
SMT Statistical Machine Translation
IL3 Hindi, Marathi, Telugu
IL5 Punjabi, Urdu, Tamil, Marathi, Hindi
CPG Computational Paninian Grammar
EMB ENG, MARATHI, BENGALI
CDAC-P CDAC Pune

III. CONCLUSION

Machine translation is relatively new in India- about two decades of research and development efforts. the goal of TDIL project and the various resource centres under the TDIL project works on developing machine translation systems for Indian languages.
There are governmental as well as voluntary efforts under way to develop common lexical resources and tools for Indian languages like pos tagger, semantically rich lexicons and word nets.

REFERENCES

[1] Sitender, SeemaBawa, “Survey of Indian Machine Translation System”. Dept of Computer science and Engineering.IJCST Vol. 3 ISSUE 1, JAN- MARCH 2012.
[2] Sanjay Kumar Dwivedi, PramodPremdassukhadeve “Machine Translation system in Indian Perspective”. Dept of Computer Science and engineering. ISN 1549-3636 ,2010
[3] Sudhir K Mishra “Sanskrit karaka analyzer for Machine Translation” a Ph.D thesis , SCS JNU New Delhi, 2007.
[4] G.S.Josam, G. S. lehal “A Punjabi to Hindi Machine Translation system”, Coling 2008. Companiono volume Posters an Demonstrations, Manchester, UK , pp 15160, 2008
[5] Tejinder Singh Saini, Gurpreetsingh , lehal “Shahmukhi to Gumukhi Transliteration System”, A Corpus based Approach , Advanced in natural Languages Processing Application Research in computing Science 33, pp.151 162,2008.
[6] Akashar Bharti, Amba Kulkarni,"Anusaarka: An Accessor cum Machine Translator", Department of Sanskrit Studies, University of Hyderabad, Hyderabad, 2009.
[7] G.S. Josan G.S. Lehal “A Punjabi to Hindi Machine Translation System ”, Coling 2008: Companion volume: Postersand Demonstrations, Manchester , UK pp. 157-160, 2008.
[8] SahaGautamKumar,"The EB-ANUBAD translator- A Hybride Scheme", Journal of Zhejjang University Science,pp. 1047-1050, 2005.
[9] Vishal Goyal, Gurpreet Singh Lehal,"Web Based Hindi to Punjabi Machine Translation System", journal of emerging technologies in web intelligence, Vol. 2, May 2010.
[10] EnConverterSpeci cation Version 2.1UNU/IAS/UNL Centre, Tokyo 150-8304, Japan, 2000