Automated Machine Translation for Regional Languages

Vijay Sita Ram Pasupuleti, KM030, 11706955, A13, Department of Computer Science and Engineering, Lovely Professional University, Punjab, India, vjsramp@gmail.com

Bharane AB, KM030, 11701916, A02, Department of Computer Science and Engineering, Lovely Professional University, Punjab, India, bharane1610@gmail.com

INTRODUCTION

Machine Translation is the automated translation process of one natural language into another. Machine Translation has tremendous potential in various fields such as health, education, information technology, business and various government agencies with regard to Indian languages. India being a multi-lingual country, languages vary by region. A total of 22 official languages are currently available in India. With the advent of information technology, many digitized documents, web pages are coming up in local languages, and the building of systems that would serve the purpose of translating in Indian languages has become indispensable. Manual translation of these documents is not only time consuming but also has a tremendous cost factor involved. As such coming up with machine translation systems that translate between the Indian languages has become very important. Moreover, translating these regional languages to English languages is also necessary. Many researchers have started working on Machine Translation systems specifically catered for the Indian languages and have gained very satisfactory results.

The research scenario in India is relatively young and machine translation gained momentum in India only from 1980 onwards with institutions like IIT Kanpur, IIT Bombay IIIT Hyderabad, University of Hyderabad, NCST Mumbai, The Technology Development in Indian Languages (TDIL) and CDAC Pune plays a major role in the systems development [17, 18]. Since then, in India, many machine translation systems have been developed, using different approaches to translating between the languages. Here we look at the major machine translation systems used for Indian language translation.

The remainder of the paper is organized as follows: Section 2 provides an idea of the different approaches to building a machine translation system. The direct approach to machine translation system building is discussed in Section 3. The Rule Based System is discussed in section 4. Section 5 provides an idea about approaches based on the corpus.

APPROACHES IN MACHINE TRANSLATION

The Machine Translation process can be broadly classified into the following Direct Machine Translation approaches, Rule Based Machine Translation, Corpus Based Machine Translation approaches.

The Rule-based approach may be further subdivided into the Transfer-based approach and the Interlingua approach, while the Corpus-based approach may be categorized into the Statistical Machine Translation and Example-based approach.

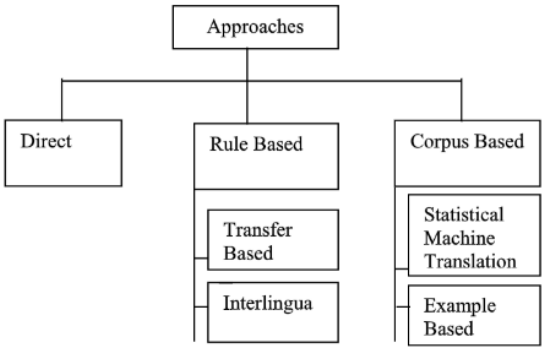

Figure 1 Machine Translation Approaches

DIRECT MACHINE TRANSLATION

Direct Machine Translation is the one of the simplest approaches to machine translation. A direct word-by-word translation of the source of input is performed in Direct Machine Translation using a bilingual dictionary and after which some syntactic arrangement is made.

In Direct Machine Translation a language is given as input, called the source language, and the output is known as the target language. The approach is typically unidirectional and only takes into account the at a time of one language pair.

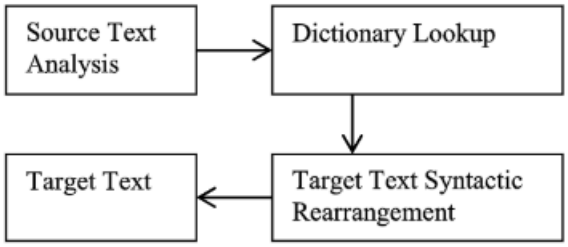

Figure 2 Direct Machine Translation Systems

The most well-known Direct Machine Translation System for Indian Languages is Anusaaraka [10], developed by IIT Kanpur, which is now being developed at IIT Hyderabad, Hindi to Punjabi Machine Translation System [5] at Punjab University, which has achieved a 95 percent accuracy.

RULE BASED MACHINE TRANSLATION

The Rule Based Machine Translation System takes into account semantic, morphological and syntactic information from a bilingual dictionary and grammar and, on the basis of these rules, generates the target output language from the source language of the input by producing an intermediate representation.

INTERLINGUA APPROACH

The Interlingua Approach translates words into an intermediate IL language, typically a universal language created for the system to use it as an intermediate language for translation into more than one target language. The Interlingua approach has an analyser which produces the source language's intermediate representation. The synthesizer takes over from the analyser and produces the analyser's target sentences.

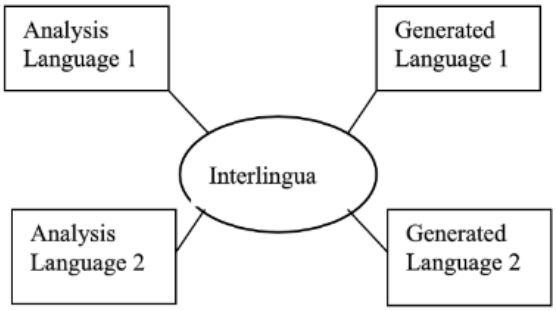

Figure 3 Interlingual Machine Translation

The major machine translation system in India based on the Interlingua approach is Anglabharti [3], developed at IIT Kanpur using a Pseudo Lingua intermediate structure for Indian Languages. AnglaHindi [6]is an extension of Anglabharti that uses rule bases, example bases and statistics to obtain translation for frequently encountered phrasal noun and verb. Work is currently under way for English to Marathi, and English to Bengali UNL.

TRANSFER BASED APPROACH

The Transfer-based approach uses translation rules to translate the input language into the output language in three phases. The approach uses a dictionary to convert source directly into target whenever a sentence matches one of the transfer rules. To this end, Source language dictionary, Target language dictionary and a bilingual dictionary are used.

The mechanism for translation is executed in three phases.

First, the source language is converted into an intermediate representation that is then converted in the second phase into target language representation. The third phase involves final target language generation.

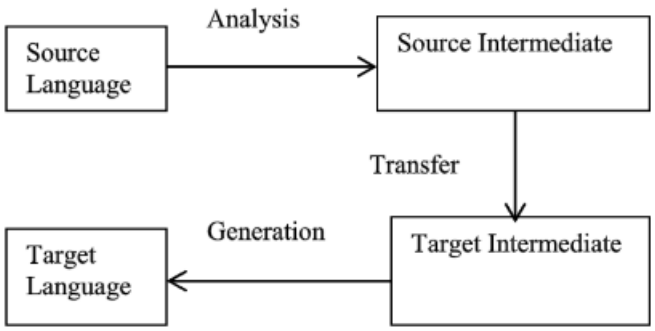

Figure 4 Transfer Based Machine Translation

Indian machine translation system based on Transfer approach includes Mantra, developed by the CDAC group of Applied Artificial Intelligence, Pune. The Matra system also developed by CDAC, Pune [11], is a further translation system.

Other systems include the Shakti Machine Translation System [13] developed by IIIT Hyderabad and IISC Bangalore that works through the combination of rule-based and statistical approach. Anubaad [12]is another translation system that is a hybrid system that uses n-gram tagging approach and works at CDAC, Kolkata's sentence level. Another notable system is the English to Kannada Machine Translation System [11], developed at the University of Hyderabad using Universal Clause Structure Grammar Formalism. Sampark [11]is another translation system based on the Paninian framework that translates Indian Languages into Indian Languages Machine developed at IIIT Hyderabad.

Recently, researchers have introduced Indian machine translation systems with distributed computing [2] and cloud computing [15, 16]. The Sampark system is deployed in a cloud environment that reduces the time of deployment [19]and the distributed system environment which increases the throughput [22].

Corpus based machine translation has evolved as preferred machine translation approach. This approach takes and trains a bilingual text corpus to get the desired output. In Statistical Machine Translation and the Example Based Machine Translation System, the corpus-based approach is mainly used.

STATISTICAL MACHINE TRANSLATION

Statistical Machine Translation is one of the most widely used approaches in modern-day machine translation. A bilingual corpus is trained in Statistical Machine Translation, and statistical parameters are derived to reach the most likely translation.

For some languages, bilingual corpus is readily available, while others have employed methods such as reputation-based social collaboration [1] to build the corporation [20] or from the web or any digitized text [21] etc.

Statistical machine translation takes place in three phases, namely language modelling, modelling and decoding of the translation.

The language model determines the likelihood of target language T which helps to achieve the fluency in the target language and to choose the right word in the language being translated. In general, it is referred to as

P(T).

On the other hand, the translation model helps to calculate the target language's conditional probability T given the source language S generally referred to as P (T).

Lastly, the maximum product probability of both the language model and the translation model is calculated in the decoding phase which gives the most statistically probable sentence in the target language.

P (S, T) = argmax P(T) P(S|T)

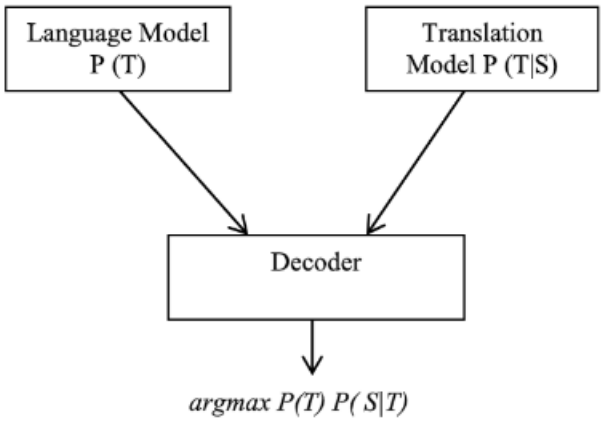

Figure 5 Statistical Machine Translation

The notable systems in the Indian perspective that use the Statistical Machine Translation approach are the English Hindi Machine Translation System [14] developed by IIIT Hyderabad that combines the Rule Based Machine Translation approach and the phrase-based Statistical Machine Translation method.

Cochin University of Science and Technology developed another translation system for translating English into Malayalam [9] using the Statistical Machine translation approach.

EXAMPLE BASED MACHINE TRANSLATION SYSTEM

Example Based Machine Translation System usually uses previous examples of translation to translate the language from source to target. Example Based Machine Translation System's basic idea is to get examples of existing translation in its example-base and provide the new translation based on that example. It generally takes place in

three phases – matching, alignment, and recombination. In the matching phase, the system searches for examples similar to the input from the example-base.

The part of the example to be used is identified and aligned in the Alignment phase, compared with other examples. The reusable parts identified during the alignment phase are assembled in the final phase, and the target language is produced.

Machine translation system in India using the Example-based approach includes Vaasaanubaada [4] which is primarily targeted for bilingual news text translation in Assamese Bengali. Anubharti [8] developed at IIT Kanpur is a hybrid Example based system combining pattern-based approach with example-based approach. The Shiva Machine Translation System [14] developed by IIIT Hyderabad and Carnegie Mellon University, USA, is another notable system which uses the example-based approach.

MACHINE TRANSLATION SYSTEMS

The various Machine Translation systems for Indian languages with their language pair and features are given in Table 1:

Table 1 Machine Translation Systems

TRANSLATION SYSTEM	APPROACH	LANGUAGE PAIR	FEATURES
Anusaaraka	Direct	Bengali, Kannada, Marathi, Punjabi, Telegu to Hindi	Uses Paninian grammar and matches local words between source and target language.
Hindi to Punjabi MTS	Direct	Hindi to Punjabi	Direct word to word translation. Morphological analysis, word sense disambiguation, post processing and transliteration
Mantra	Transfer Based	English to Hindi, Gujrati, Telegu. Hindi to English, Bengali, Marathi	Uses Tree Adjoining Grammar Formalism.
Matra	Transfer Based	English to Hindi	Human assisted translation project uses rule bases and heuristics.
Shakti	Transfer Based	English to Hindi, Marathi, Telegu	Works by combining rule based and statistical approach.
Anubaad	Transfer Based	English to Bengali	Hybrid system which uses n-gram approach for POS tagging. Works at sentence level
English Kannada MTS	Transfer Based	English to Kannada	Uses Universal Clause Structure Grammar Formalism
Sampark	Transfer Based	Punjabi-Hindi, Telegu to Tamil, Hindi-Urdu, Hindi to Telegu	Uses Computational Paninian Grammar (CPG) approach for analyzing language and combines it with machine learning
Anglabharti	Interlingual	English to Hindi, Tamil	Uses intermediate structure Pseudo Lingua for Indian Languages.
AnglaHindi	Interlingual	English to Hindi	Uses rule-bases, example-base and statistics to obtain translation for frequently encountered noun and verb phrasal.
UNL English Hindi MTS	Interlingual	English to Hindi	Uses Universal Natural Language as interlingua
English Hindi MTS	Statistical Machine Translation	English to Hindi	Combines Rule Based Machine Translation and phrase based Statistical Machine Translation
English Malyalam MTS	Statistical Machine Translation	English to Malyalam	Uses SMT by using monolingual Malayalam corpus and a bilingual English/Malayalam corpus in the training phase
Vaasaanubaada	Example Based Machine Translation	Bilingual Bengali Assamese	Preprocessing and post processing task, longer sentences fragmented at punctuation, backtracking for unmatched results.
Anubharti	Example Based Machine Translation	English-Hindi	Hybrid Example based system which combines pattern based and example based approach.
Shiva	Example Based Machine Translation	English- Hindi	Uses linguistic rules and statistical approach to infer linguistic information.

GOOGLE'S LANGUAGE CODES

Language	Codes
Afrikaans	af
Arabic	ar
Azerbaijani	az
Belarusian	be
Bulgarian	bg
Bengali	bn
Bosnian	bs
Catalan	ca
Cebuano	ceb
Czech	cs
Welsh	cy
Danish	da
German	de
Greek	el
English	en
Esperanto	eo
Spanish	es
Estonian	et
Finnish	fi
French	fr
Irish	ga
Galician	gl
Gujarati	gu
Hausa	ha
Hind	hi
Hmong	hmn
Croatian	hr
Haitian Creole	ht
Hungarian	hu
Armenian	hy
Indonesian	id
Igbo	ig
Icelandic	is
Italian	it
Hebrew	iw
Japanese	ja

Javanese	jw
Georgian	ka
Kazakh	kk
Khmer	km
Kannada	kn
Korean	ko
Latin	la
Lao	lo
Lithuanian	lt
Latvian	lv
Punjabi	ma
Malagasy	mg
Maori	mi
Macedonian	mk
Malayalam	ml
Mongolian	mn
Marathi	mr
Malay	ms
Maltese	mt
Myanmar (Burmese)	my
Nepali	ne
Dutch	nl
Norwegian	no
Chichewa	ny
Polish	pl
Portuguese	pt
Romanian	ro
Russian	ru
Sinhala	si
Slovak	sk
Slovenian	sl
Somali	so
Albanian	sq
Serbian	sr
Sesotho	st
Sudanese	su
Swedish	sv
Swahili	sw
Tamil	ta
Telugu	te
Tajik	tg
Filipino	tl
Turkish	tr
Ukrainian	uk
Urdu	ur
Uzbek	uz
Vietnamese	vi
Yiddish	yi
Yoruba	yo
Chinese Simplified	zh - CN
Chinese Traditional	zh - TW
Zulu	zu
Persian	fa
Thai	th

CONCLUSION

In this paper, along with their features, we discussed and examined the various Machine Translation Systems in India. We also discussed the different approaches that are used to construct a machine translation system. Many researchers and research groups have developed different systems of translation which apply different approaches. Although many languages are yet to be covered under Machine Translation in India, significant research and work has gone on to include many of the languages. Many languages and it is believed that in a few years 'time most of the major Indian languages will be covered under the Machine Translation Ambit.

REFERENCES

[1] Animesh Kr Trivedi, Rishi Kapoor, Rajan Arora, Sudip Sanyal and Sugata Sanyal, RISM - Reputation Based Intrusion Detection System for Mobile Ad hoc Networks, Third International Conference on Computers and Devices for Communications, CODEC-06, pp. 234-237. Institute of Radio Physics and Electronics, University of Calcutta, December 18-20, 2006, Kolkata, India.

[2] Sugata Sanyal, Ajith Abraham, Dhaval Gada, Rajat Gogri, Punit Rathod, Zalak Dedhia, Nirali Mody: Security Scheme for Distributed DoSinMobile Ad Hoc Networks, 6th International Workshop on Distributed Computing (IWDC-2004), A. Sen et al (Eds.). Springer Verlag, Germany, Lecture Notes in Computer Science, Vol.3326. ISBN: 3-540-24076-4, pp.541-542, 2004.

[3] R.M.K. Sinha et al., ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi, 1995 IEEE International Conference on Systems, Mannd Cybernetics, Vancouver, Canada, 1995, pp 1609-1614.

[4] Vijayanand K., Choudhury S.I., P., Ratna P., VAASAANUBAADA -Automatic Machine Translation of Bilingual Bengali-Assamese News Texts, Language Engineering Conference,2002, Hyderabad, India, pp.183-188.

[5] Vishal Goyal, Gurpreet Singh Lehal, “Hindi to Punjabi Machine Translation System”, In the Proceedings International Conference for Information Systems for Indian Languages, Patiala, Department of Computer Science, Punjabi

University, Patiala, March 9-11, 2011, pp. 236-241, Springer CCIS139, Germany (2011)

[6] Sinha R.M.K, Jain A., AnglaHindi: an English to Hindi machine-aided translation system, MT Summit IX, New Orleans, USA, 23-27 September2003, pp.494-497.

[7] Shachi Dave, Jignashu Parikh, Pushpak Bhattacharyya., "Interlingua-based English–Hindi Machine Translation and Language Divergence", Machine Translation, 2001, Volume 16, Issue 4, pp 251-304

[8] Sinha R.M.K, An Engineering Perspective of Machine Translation: AnglaBharti - II and AnuBharti -II Architectures, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS-2004), November 17-19, 2004.

[9] Sebastian M.P., Kurien S.K., Kumar G.S., English to Malayalam translation –a statistical approach, A2CWiC-10, First Amrita ACM –W celebration in women in Computing in India, 2010, article 64, pp.1-5, DOA 10.1145/1858378-1858442.

[10] Akshar Bharati, Vineet Chaitanya, Amba P Kulkarni and RajeevSangal; Anusaaraka: Machine Translation in Stages, Vivek, A Quarterly in Artificial Intelligence, 10, 3, National Centre of Software Technology, Bombay (renamed as CDAC, Mumbai), July 1997 pp.22-25.

[11] Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve, "Machine Translation System in Indian Perspective", Journal of Computer Science vol. 6(10), pp. 1082-1087, 2010 Science Publications, 2010.

[12] S. Bandyopadhyay, ANUBAAD -The Translator from English to Indian Languages, 7th State Science and Technology Congress, Calcutta, India, 2000.

[13] Bharati, A., R. Moona, P. Reddy, B. Sankar and D.M. Sharma et al., 2003. Machine translation: The Shakti approach. Proceeding of the 19thInternational Conference on Natural Language Processing, Dec. 2003, MT-Archive, India, pp: 1-7.

[14] Arafat Ahsan, Prasanth Kolachina, Sudheer Kolachina, Dipti Mishra Sharma, Rajdeev Sanghal, "Coupling Statistical Machine Translation with Rule-based Transfer and Generation", AMTA-The Ninth Conference of the Association for Machine Translation in the Americas. Denver, Colorado, 2010.

[15] R Bhadauria, R Chaki, N Chaki, S Sanyal; "A Survey on Security Issues in Cloud Computing"-arXiv preprint arXiv:1109.5388, 2011 -arxiv.org

[16] Shantanu Pal, Sunirmal Khatua, Nabendu Chaki, Sugata Sanyal; "A New Trusted and Collaborative Agent Based Approach for Ensuring Cloud Security"; Annals of Faculty Engineering Hunedoara International Journal of Engineering; Vol. 10, Issue 1, February, 2012. pp. 71-78. ISSN: 1584-2665.

[17] Murthy, B. K., Deshpande W. R., Language technology in India: past, present and future, 1998, http://www.cicc.or.jp/english/hyoujyunka/mlit3/7-12.html [Dec 11,2011]

[18] V Goyal, G S Lehal. "Advances in Machine Translation Systems". Language in India, Vol. 9, No. 11, 2009, pp. 138-150.

[19] Pawan Kumar, B. D. Chaudhary, Rashid Ahmad, and Rajeev Sangal. "Machine Translation System as Virtual Appliance: For Scalable Service Deployment on Cloud." The 7th IEEE International Symposium on Service-Oriented System Engineering (IEEESOSE 2013),2013.

[20] Shinsuke Goto, YoheiMurakami, and Toru Ishida. "Reputation-based selection of language services." InServices Computing (SCC), 2011 IEEE International Conference on, pp. 330-337. IEEE, 2011.

[21] Adam Kilgarriff, Gregory Grefenstette. "Introduction to the special issue on the web as corpus. "Computational linguistics29, no. 3 (2003): 333-347.

[22] Rashid Ahmad, Pawan Kumar, B. Rambabu, Phani Sajja, Mukul K. Sinha, and Rajeev Sangal. "Enhancing Throughput of a Machine Translation System using MapReduce Framework: An Engineering Approach.", ICON-2011: 9th International Conference on Natural Language Processing (ICON-2011)2011.