International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1059-1071

http://www.ripublication.com

Review of Machine Translation Techniques for Idea of Hindi to English Idiom Translation

Rajesh Kumar Chakrawarti
Reader, Department of CSE,
Shri Vaishnav Institute of Technology and Science,
Indore, Madhya Pradesh, India.
rajesh_kr_chakra@yahoo.com,

Himani Mishra

PG Scholar, Department of CSE,
Shri Vaishnav Institute of Technology and Science,
Indore, Madhya Pradesh, India.
himanimishra.hm21@gmail.com,

Dr. Pratosh Bansal

Professor ,Department of IT, Institute of Engineering and Technology,
Devi Ahilya Vishwavidyalaya,
Indore, Madhya Pradesh, India.
pratosh@hotmail.com,

Abstract

In past few years, we have witnessed several significant advancements in Natural Language Processing , which has let text and speech processing to make huge gateway to world-wide information source[1] .The paper focuses on the techniques and approaches like corpus-based, rule-based, direct and hybrid approach [2] [3] used for machine translation systems together with their example systems, problems during translation which majorly includes-structural divergence, ambiguities like phase level-idiom translation and word level ambiguity, non-standard language, named entities etc., benefits and limitations of machine translation. Together with this it also shed light on problems with idiom translations as it is a very important part of any language. Many significant machine translation systems are briefly discussed.

Keywords: Machine translation; approach; idiom; language; translation.

I. INTRODUCTION

A sub-field of computational linguistics and natural language processing that uses machine to translate text or speech from one natural language to another natural language is defined as Machine translation [3]. The language which is used by humans to express themselves is called as natural language. Natural language is used for our daily communication. For instance Hindi, Punjabi, Bengali, Marathi are the natural languages used in India. The need for a translation system can be understood through an example. Sanskrit is among one of the most ancient languages. Today most people do not understand Sanskrit which contains in it a huge source of Vedas, verses, shlokas and idioms. But the people can get all these knowledge, if that information is translated into the language they understands. For gaining accession to all the information and making communication effortless, a natural language processing system is required. We have done an in-depth study of all the major approaches and the type of systems that can be implemented using these approaches, so as to find the best suited for idiom translation.

II. HISTORY

The research and work on machine translation began in 1949 with an idea of Warren Weaver proposed to use computers in natural language translation by adopting the term "computer translation". In 1952, the first conference related to it was organized at MIT which was guided by Yehoshua Bar-Hillel. In 1954, the initial automatic 'Russian to English machine translator' was devised [4].

The very first Global Conference on machine translation under the title "Languages and Applied Language Analysis of Teddington" in 1961, attended by world's linguists and computer scientists. In 1964, a committee was formed called ALPAC (Automatic Language Processing Advisory Committee) [4].

During 1970 to 1980, the project named REVERSO was initialized by some Russian Researchers (1970), another machine translation system (MTS) named SYSTRAN1 (Russian to English) by Peter Toma, was developed. A machine translation system ATLAS2 (Korean to Japanese) by FUJITSU(a Japanese firm) was developed. FUJITSU was founded on rules (1978) [4].

During 1980 to 1990, Japanese made a huge contribution and advancement in the field of machine translations (MT). In 1983, NEC developed a machine translation system using PIVOT algorithm titled as 'Honyaku Adaptor II', for Interlingua approach. Hitachi formed a Japanese to English language translation system called as HICATS (Hitachi Computer Aided Translation System) [2].

The first trilingual (three languages were included-English, German, Japanese) machine translation system was started under the project C-STAR (Consortium for Speech Translation Advanced Research) and in 1998, merchandising of machine translator 'REVERSO' was Softissimo's task[4].

During 1990 to 2000, the use of MT touched new heights as it became a part of the internet. In 2005, the initial website for automatic machine translation was set up by Google. In 2008, machine translation took a hype with 23% of Internet users exploring machine translation's features and 40 % considering doing so and in 2009, 30% of the professionals have started utilizing machine translation systems for their work, 18% perform a proofreading and 50% planned to use machine translation for translation(2010) [4].As all these research is for the language translation, it already includes ambiguities.

III. LITERATURE SURVEY

Linguists, human translators and software developers needs a close cooperation to develop a machine translation system. Two major goals during development process are- (a) accuracy of translation and (b) speed. To obtain a single correct parse, the input text or speech should be optimized, which will finally result in a single translated output text or speech in target language. For this various methods/approaches and systems are studied below.

A. Approaches

To understand a language (let it be source language) and translate in the other language (target language) needs deep knowledge of both the language's grammar, semantics, syntax, idioms, etc., as well as the culture of its speakers. When it comes for the machine to do the translation, the major issue arises - how to make them "understand" the language as a human does [2]. For this, many MT approaches are used. Major MT approaches are:-

1. Corpus-based

a. Statistical Machine Translation (SMT),
b. Example- Based Machine Translation (EBMT) [3] [4] [5],

2. Direct Machine Translation[4],

3. Rule-Based Machine Translation

a. Transfer-Based, and

b. Interlingual based, and [4]

4. Hybrid technique [3].

1. Corpus-Based Technique:- Corpus-based method rely on the analysis of bilingual text corpora [3].It has let us discover how to exploit the statistical properties of text and speech databases. Corpus based systems are fully automatic and it does not involve much human labor as in rule-based approaches. It is further categorized as SMT and EBMT [5].

*Statistical Machine Translation technique:-*The translations generated are on the ground of a statistical framework. The parameters of this framework are driven from the study of bilingual text corpora. A huge parallel corpora is needed for training the

SMT systems [6] . CANDIDE (IBM) was the introductory statistical machine translation software. SYSTRAN was employed by Google for many years. It swapped to statistical translation method in October 2007. There are many other methods into Statistical Machine translation like- METIS II and PRESEMT [2]. Rules-based systems may take years for language pairs and significant expenditure to form, on the other hand, statistical systems can be trained to produce translations in weeks or days, thus making it little labor intensive, thus making it utilitarian for time-critical businesses, IT industries and government applications [7].

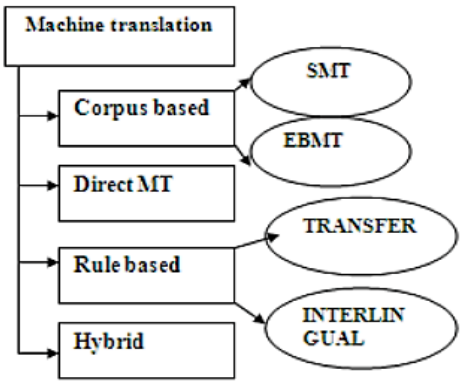

Fig. 1 Different translation approaches [2][3]

Example-Based Machine Translation -Makoto Nagao coined EBMT's concept in 1984 [2]. EBMT is basically translation by 'comparison'. This can be understood as- Suppose an EBMT system is presented with a collection of sentences in the input language and respective translations in the output language, then the system uses these examples for translation of other such similar input language sentences into output language sentences. The postulate is- if a sentence which has been translated in previous time comes again, than that previous translation may be correct this time [8] .This approach uses three steps:-Matching the fragments against the parallel corpus, adjusting the matched fragments to the target language and recombining the translated fragments roughly [6].

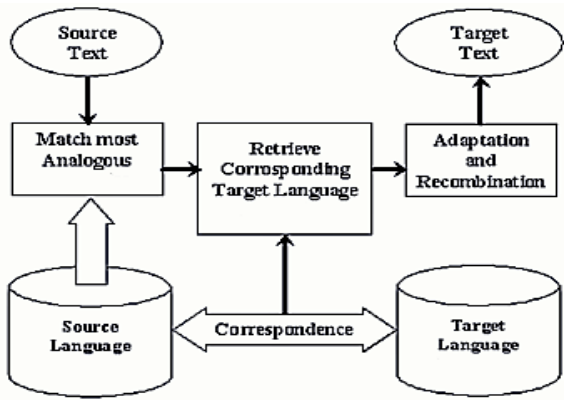

Fig. 2 Architecture of EBMT [9]

2. Direct Machine Translation:- Direct machine translation is possibly the simplest machine translation method or technique available today. In this scheme a word by word translation of the source language is done with the use of bilingual dictionary followed by few grammatical rearrangements. This approach is one directional and considers single language pair at any instant of time [6].

3. Rule-Based Machine Translation technique:-Rule-based MT systems examines the input (source language) and generates some sort of mediate representation generally called as intermediate representation (IR), for example, this IR can be a parse tree or it may be a abstract representation. This intermediate representation is used to produce or generate the output text (target language) [5]. Semantic, morphological, and syntactic information are considered from a bilingual dictionary and grammar which generates the target output language text from the source input language text.

Transfer based machine translation: -To get the composition of the input sentence, transfer based machine translation requires analysis of input text . It comprises of three components namely: analysis module, transfer module and generation module. The composition (structure) of source language is produced by analysis module. This source language composition is converted to a target language by the transfer module, for which it requires the sub-tree rearrangement rules. The target language text is produced by the generation module using target language structure which involves a lexical conversion of verbs, conversion of auxiliary verb for tense, transfer of gender, number and person information [5].

Interlingual-based machine translation: - The Interlingual-based machine translation method is founded on Chomsky's claim. The translation here, has two components. Those are: analysis module and synthesis module. In analysis part , "interlingua", a language-independent meaning representation is produced using input source language text. In synthesis part, target language is generated through this intermediate interlingual representation [5].

4. Hybrid Machine Translation:-Two machine translation approaches can be merged to form a hybrid approach [3]. Like a statistical-rule based approach takes strategic advantages of the capabilities of SMT and RBT machine translation approaches. Some MT companies averred a hybrid method which utilizes both rules and statistics methods among them includes- Omniscien Technologies, LinguaSys, Systran, and Polytechnic University of Valencia. These methods are different in the ways they are used or sequenced to be used:-

- *Rules post processed by statistics:* - The translation part is done using rule based approach , followed by using statistics for adjusting the output.
- *Statistics guided by rules:* - In this approach, rule-based approach is used to preprocess the data so that it can be better guided by the statistical engine [2].

B. Some Machine Translation Systems

A completely automatic Machine Translation System should significantly consists of components like- Tokenizer or segmenter, Morphological (grammar) analyzer, POS

tagger, Word sense disambiguator (to remove ambiguities like-idioms, poetry, verses, words etc), Transfer module, Dictionary and Target word generator [5]. Below we are presenting some machine translation systems which are implemented on above discussed approaches.

1. MTS using Statistical Machine Translation technique

Shakti (2003): - Developed by Bharati, R Moona, P Reddy, B Sankar, D M Sharma and R Sangal. With simple architecture, Shakti can be used to translate English text to any Indian languages. Shakti contains 69 different modules [5].

English to Indian Languages MT System (E-ILMT) (2006:)- E-ILMT was developed by C-DAC Mumbai, IISc Bangalore, IIIT Hyderabad, C-DAC Pune, IIT Mumbai et al. . It is a MTS for Tourism and Healthcare Domains.

Table 1. EXAMPLES OF SYSTEM USING SMT APPROACH [4] [5]

S.No.	Name of Translation System	Languages	Year
1.	Shakti	English- Indian language	2003
2.	English to Indian Languages MT System (EILMT)	English- to Hindi or Marathi or Bengali	2006

2. MTS using Example-Based Machine Translation(EBMT)

ANUBHARTI II (2004:)- Developed by R.M.K Sinha . A machine translation system with concepts of example-based and corpus-based techniques both with few basic morphological analysis. The conventional EBMT concept was improvised to cut down the necessity of a ample example base in ANUBHARTI.

VAASAANUBAADA (2002):- Developed by Vijayanand Kommaluri, Sirajul Islam Choudhury and Pranab Ratna. A machine translation architecture for converting the news texts was proposed, based on EMBT approach. Some preprocessing and post processing of the translated news text was required to be carried out [4] for better translation of the output.

IBM English-Hindi Machine Translation System (2006)-Developed by D. Gupta, N. Chatterjee and Raghavendra Udupa,. A machine translation based on EBMT approach was used and later on the approach was changed and switched to the Statistical method for translation. This was proposed in IBM India Research Lab[9][4].

TABLE 2. EXAMPLES OF SYSTEM USING EBMT APPROACH [4] [5] [9]

S.No.	Name of Translation System	Languages	Year
1.	Anubharti	Hindi- Indian Language	1995
2.	Vaasaanubaada	Bengali-Assami	2002
3.	Anubharti-II	Hindi-Indian Language	2004
4.	shiva	Hindi- Indian Language	2004
5.	IBM-MTS	English-Hindi Language	2006

3. MTS using Direct Machine Translation

Anusaaraka systems among Indian Languages (1995): -This system was given by Rajeev Sangal (was initialized at IIT Kanpur and currently at IIIT Hyderabad). The translation among Indian languages was the main objective of this system. The accepted input languages includes (Telugu, Kannada, Bengali, Punjabi and Marathi).

Punjabi to Hindi MT System (2007, 2008:)- Developed by G S Josan and G S Lehal. Direct word-to-word machine translation method is applied and includes components like- preprocessing, word-to-word conversion through Punjabi-Hindi lexicon, morphological analysis, word sense analysis (phase level and word level both) disambiguation (removing the ambiguity like the word sense) and some after translation tasks [4].

Web-based Hindi-to-Punjabi MT System (2010):- Goyal V and Lehal G S. As most of the websites are written in Hindi as compared to Punjabi. The machine translation for Hindi to Punjabi translation was extended to the internet. It included many features like website translation, email translation, etc. [4].

Table 3. EXAMPLES OF SYSTEM USING DIRECT APPROACH [4] [5]

S.No.	Name of Translation System	Languages	Year
1.	Anusaaraka systems	Indian language.	1995
2.	Punjabi to Hindi MT System	Punjabi-Hindi	2007,2008
3.	Web-based Hindi-to-Punjabi MT System	Hindi-Punjabi	2010

4. MTS using Hybrid technique

Bengali to Hindi MT System (2009):- Developed by Chatterji S, Roy D, Sarkar S. A hybrid MTS (combination of corpus-based (SMT) with a rule-based lexical TBMT) i.e. multi-engine Machine Translation concept was proposed.

*Lattice-Based Lexical Transfer in Bengali Hindi MT Framework (2011):-*Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar, and Anupam Basu . An approach for accurate linguistic translation in Bengali- Hindi translation architecture [4] was given.

Table 4. EXAMPLES OF SYSTEM USING HYBRID APPROACH [4]

S.No.	Name of Translation System	Languages	Year
1.	ANUBHARTI-II	Hindi-Indian languages	2004
2.	Bengali to Hindi MT System	Bengali-Hindi	2009
3.	Lexical Transfer in Bengali Hindi MT Architecture	Bengali-Hindi	2011

C. Various Existing Online/Offline Translation System for HINDI to ENGLISH

We can understand any concept or any process if we understand the language in which it is available. Therefore major problem in the progress of any individual is language. To communicate with other countries/states either they have to learn that language or use translation. In India there are near about 400 Million Hindi language speakers often they need Hindi to English Translation software [3]. Here we are tabulating some Machine translation software available that translates Hindi into English or vice Verse.

Table 5. LIST OF ONLINE/OFFLINE TRANSLATION SYSTEMS

S. No.	Name of System	Characteristics	Link
1.	India Typing	To make it easy to work with the English language [10].	http://indiatyping.com/index.php/translations/Hindi-to-English-translation
2.	Soft112	Here the text to be converted can be taken from any document or file and can be used for conversion [11].	http://English-to-Hindi-and-Hindi-to-English-converter-software.soft112.com/
3.	SAMPARK	Developed by IIT-Hyderabad, IIT-Kharagpur,IIT-Bombay etc. and this program is funded by TDIL for providing	http://ilmt.tdil-dc.gov.in/sampark/web/index.php/content

		an easily understandable website for language translation [12].	
4.	Hindi to English translator	This translator along with simple Hindi to English translation also provides multiple services like—translate and listen, translate and compare,web translation,webmaster tools TTS, Imtranslator [13].	http://translation2.paralink.com/Hindi-English-Translator
5.	Google translator	Google provides a common user unlimited services among which google translator is one. It also allows us to type in any language(like typing Hindi in English) and then converting it into the language of our choice [3] [14].	https://translate.google.co.in/
6.	Dictionay.com	This website provides some daily used sentences and some example sentences generally used in many different environments like at the airport, at the hotels, in a car, in the kitchen etc. for quick review and learning purpose [15].	http://translate.reference.com/
7.	Lexicool	It provides websites own translator together with the links to other translator websites. Now the user can search on multiple sites by clicking single website [16].	http://www.lexicool.com/Hindi-dictionary-translation.asp
8.	Babylon	It provides the translation of texts, phrases.Babylon software has 10 years of experience and it also has 1600 dictionaries to refer from [17].	http://translation.babylon-software.com/English/to-Hindi/

IV. PROBLEM IDENTIFICATION

During machine translation, various problems arise especially with idiomatic sentences. The reasons for these issues are:-

1. *Structural Divergences:* - Talking about Hindi-English translations, English has Subject-Verb-Object (SVO) structure with a inadequate morphology whereas in case of Hindi, it has Subject-Object-Verb (SOV) structure and is a morphologically rich language. There structural and morphological differences are responsible for the trouble in translation using some approaches [18].
2. *Approach used:* - Sometimes the method used to develop a translator comes with some pros and cons which become the advantages and disadvantages of our translator itself. For e.g. - transfer-based approach is better than Corpus-Based MTS because Corpus-Based MTS require a large amount of word aligned data for translation that is not available for many languages [19].
3. *Ambiguity:*-Similarly to make a machine understand the sense of a word in another language is also a hard nut to crack [20]. This issue arises due to

presence of more than one rendition of words or sentence. This issue comprises of following points:

a) *Phrase level ambiguity:* - When a group of words can be understood in multiple ways it is termed as 'phase level ambiguity'.These phrases refers to a totally different meaning than the words used in it indicate. For instance, the expression 'spill the beans' can be rendered as- to the beans that are disgorged or idiomatically the phrase can be decoded as "to leak out secret information".

b) *Word level ambiguity:* -The word ambiguity can be defined as- multiple interpretations of words framing the phrase. For example to bear the lion in his den. Here 'bear' have multiple renditions like- "a carnivore animal -bhalu", "to have", "to tolerate", "to suffer" [21].

4. *Cultural Problems:*- Culture has always put a great impact on the language of the persons following it. Many issues arises during cross-cultural translation. As the difference between the given source culture and target culture increases, the more severe trouble would be to translate especially idioms, poetry, phrases etc. [21].

5. *Named entities:*- Named entities is concerned to named entity identification in information derivation. Name entities are concerned with the realistic or abstract entities in the real world like- person's name, organizations, companies, places and a name in an idiom etc. The initial difficulty that arises is to identify them. If this cannot be identified by the machine translation system during translation, it would change the text's meaning [2].

V. BENEFITS

Using the machine for translation over human translator has many benefits (if the machine translation system is properly implemented). Some of the benefits of MTS are described below:-

1. *Too much to be translated:*-In the 21st century, there is a huge amount of data to be translated due to globalization, education and industrialization [9].
2. *Terminology's consistently:*-There may arise a situation where a human translator is incapable or unaware of the terminologies as language is a world in itself.
3. *Increase speed and throughput:*-Time is a crucial element today. MTS are always much faster and less error prone than human translators.
4. *Reduced cost:*-A MTS developed once can be used multiple times.
5. *Boring for human translators:*- It is a well-known fact that human translator will get bored after some time of this job which will affect the quality of translation and the time required in it [9].

6. *As a teaching tool*- Some significant advantages of applying machine translation in the classroom are found by Dr. Ana Nino of the University of Manchester has. An example of such teaching approach is using "MT as a Bad Model"[2].

VI. APPLICATIONS

Some of the applications of machine translation systems are described below:-

1. Its very common to contact call centers that use speech-understanding systems for solutions of various issues (for instance, when accessing travel information) [1].
2. There is a special module in web architecture, called- *Request Interpreter in their architecture.* This module is responsible to translate native language of the user of website and connect it to a schematic structure, accessible by service generation engine [22].
3. An article written by Amy Isard, Jon Oberlander, Ion Androutsopoulos and Colin Matheson titled as "Speaking the Users' Languages", depicts a system which produces descriptions of unseen objects in the text of various degrees of complexity. It modifies these descriptions to the user's expertise—for e.g., for adults, children, or expert [1].
4. MTS nowadays are very frequently used by students, faculties, professionals and common people for understanding any information in their native language.

VII. LIMITATION

There are some factors which limit the machine translation systems to be used to the fullest of its potential. Together with those discussed in problem identification some are as under:

1. *Dictionary used*:-Translation will be greatly affected by the depth and richness of dictionary used [2].
2. As it is a machine, failure of the machine can't be predicted. Like all other system, it may crash down at any instance.
3. Idioms are difficult to interpret as they point to some other meaning than the words used [21].

VIII. CONCLUSION AND FUTURE SCOPE

In this survey paper, we studied various MT approaches, techniques, and many machine translation systems together with their benefits and limitations in a longitudinal and latitudinal way. Many research have been done on various

approaches and ideas which have been even implemented in many systems globally. The study reveals that an approach comes with its own pros and cons. So hybrid technique was also introduced for maximizing the benefits. A lot of work has been done. But then too, the MT systems developed till date, have defects in its rule set, dictionary, translation technology and approaches applied and is evident from the research and studies that encouragement and more research is required in the field of MT to build intelligible translation systems for upcoming future of technologies. Some better and easier methods are awaited.

REFERENCES

[1] Ciravegna F, Harabagiu S (2003) Recent Advances in Natural Language Processing. IEEE magazine. computer.org/intelligent.

[2] Machine translation. https://en.m.wikipedia.org/wiki/Machine_translation

[3] Nair J, Amrutha K K, Deetha R (2016) An Efficient English to Hindi Machine Translation System Using Hybrid Mechanism. Conference on Advances in Computing, Communications and Informatics (ICACCI).

[4] Garje GV, Kharate GK (2013) SURVEY OF MACHINE TRANSLATION SYSTEMS IN INDIA. International Journal on Natural Language Computing (IJNLC).Vol. 2. No.4

[5] Nair LR , David PS. (2012) Machine Translation Systems for Indian Languages. International Journal of Computer Applications (0975 – 8887) . Volume 39– No.1

[6] Wagadiya N, Ravarta P English-Hindi Translation system with Scarce resources. International journal of innovative research and development.

[7] Geer D (2005) Statistical Machine Translation Gains Respect. IEEE Computer Society.

[8] Code project (2010) Develop your own translation system . http://www.codeproject.com/Articles/100126/DevelopYourOwnLanguageTranslationSystem

[9] Sinhal RA, Gupta KO (2014) A Pure EBMT Approach for English to Hindi Sentence Translation System. I.J. Modern Education and Computer Science. http://www.mecs-press.org/

[10] India Typing . Hindi to English Translator. http://indiatyping.com/index.php/translations/Hindi-to-English-translation

[11] Soft112. Hindi to English and English to Hindi Converter Software. http://English-to-Hindi-and-Hindi-to-English-converter-software.soft112.com/

[12] SAMPARK. Indian Language Technology Proliferation and Deployment center. (TDIL). http://ilmt.tdil-dc.gov.in/sampark/web/index.php/content

[13] Hindi to English translator. http://translation2.paralink.com/Hindi-English-Translator

[14] Google translator. https://translate.google.co.in/

[15] Dictionary.com. http://translate.reference.com/

[16] Lexicool. Online English<>Hindi Translations, Dictionaries and Resources. http://www.lexicool.com/Hindi-dictionary-translation.asp

[17] Babylon. English to Hindi Translation. http://translation.babylon

[18] Dungarwal P, Chatterjee R et al. (2014) The IIT Bombay Hindi, English Translation System at WMT 2014.

[19] Gehlot A, Sharma V et al. (2015) Hindi to English Transfer Based Machine Translation System. International Journal of Advanced Computer Research. Volume-5

[20] Kumar P, Srivastava S et al. Syntax directed translator for English to Hindi language. IEEE transaction. http://ieeexplore.ieee.org/document/7434282/.

[21] Gaule M, Josan GS Machine Translation of Idioms from English to Hindi . International Journal Of Computational Engineering Research. Vol. 2. Issue. 6

[22] Bosca A, Ferrato A et al. Composing Web Services on the Basis of Natural Language Requests. Proceedings of the IEEE International Conference on Web Services (ICWS'05), 0-7695-2409-5/05