# Man-Machine Integration in Translation Processes: an Indian Scenario

**R. Mahesh K. Sinha**
Indian Institute of Technology, Kanpur, India
sinharmk@gmail.com

**Abstract.** Translating natural language text or speech from one language to another is a challenging task. The quality of machine translation is found to be inferior to that of translation produced by a human being. However, machines are good at providing rough translations which can be used by human translators. Thus integrating man with the machine in the translation process is one of the ways of making translation systems practical in real life. There can be a learning loop in this man-machine integration process. This paper is focused on examining these aspects with specific reference to the Indian scenario. The Indian translation scenario is complex with a multiplicity of languages and scripts. As compared to the EU scenario, besides multiplicity of scripts, Indian languages exhibit free word group order; are morphologically rich; have complex usage of predicate verbs; use various distinctive features such as replicative words, onomatopoeic combinations etc. These result in a large number of variations in semantically equivalent utterances or written forms. The translation industry is still in its infancy in terms of preparedness for technology absorption. Machine translation strategies employed are primarily rule-based.

**Keywords:** machine and human translation, Indian scenario, bootstrapping

## 1 Introduction

The task of natural language translation from one language to another is attributed to human intelligence. Besides the knowledge of the two languages, it requires an 'understanding' of the source language text. The human translation (HT) process can be visualized as transformation of the mental picture created through understanding the source language into the target language structure that in the translator's judgment is a truthful reproduction of the mental picture to the target language audience. Translation accuracy, comprehensibility and fluency are some of the major concerns in this transformation.

On the other hand, a machine can hold and process a large amount of information that may be logically structured for speedy and relevant information retrieval. However, in case of rule-based MT systems, the limitations arise due to inadequacy of the language grammar and the rules that are usually handcrafted or mined with inad-

equate corpora. A natural language is what the native speakers use for their communication among themselves. To begin with, for a natural language, no grammar formalism exists and subsequently whatever grammar rules get formulated, do not provide complete coverage. In case of corpus based statistical approach, the limitation arises primarily due to the inadequate size of the corpus used in the learning process. The corpus used must be representative, exhaustive and adequate for such systems to succeed. These limitations of machine translation invariably lead to an imperfect translation in most of the situations. Nevertheless, machine translation (MT) provides a good starting point for the human translators giving a rough understanding, generates choices with alternative translations and creates its own resource of example translations in the specific domains and/or documents. Thus the roles of man and machine can be complementary in the translation process, and most of the translation industry worldwide is geared towards facilitating such complementary roles. This is the first level or surface level of man-machine integration (HMI or MMI) in the translation process.

In this paper, I first present the Indian translation scenario and then examine some of the basic issues in HT and MT in the section 3. In section 4, the relevance of HMI in the Indian context is examined and an integration framework is presented.

## 2 The Indian scenario

### 2.1 Linguistic scenario

i. A majority of Indian languages have a common origin belonging to the Indo-Aryan family (a sub-family of Indo-European used by 74.24%) or the Dravidian family (used by 23.86%). They are structurally similar to each other with respect to verb-ending (SOV) and relatively free word-order. The Indo-Aryan family consists of north Indian languages, while the Dravidian family consists of south Indian languages, the major languages being Kannada, Malayalam, Tamil and Telugu. All of these languages have undergone intense cross-fertilization over a period of time and have had varied degree of influence. Most of the Indian languages have either been spawned or greatly influenced by Sanskrit. They share approximately 60 % of the lexicons on average. Even though the Dravidian family of languages has evolved independently of Sanskrit, Malayalam and Telugu share about 80% of lexicons from Sanskrit. Hindi-Urdu, also called Hindustani, is greatly influenced by Persian. Other major families of languages are Austro-Asiatic (1.16%) and Tibeto-Burman (0.62%), prevalent in the north east part of India. They have had quite an independent evolution and exhibit distinctive features as compared to the Indo-Aryan and Dravidian families. There are several other lesser known families of languages with even fewer speakers.

ii. According to the 1961 census of India, there are 1652 languages (mother tongues) in India. Many of these languages exist only in oral form. Some of these are on the verge of extinction. There are 22 officially (constitutionally) recognized languages. In addition, there are about 32 languages with more than one million speakers and 122 other living languages, each of which is being used by more than a popula-

tion of 10,000 people. Many of these languages are aspiring to get officially recognized. As a consequence, an interlingual MT methodology is an obvious choice.

iii. The Indian Constitution provides that Hindi in Devanagari script shall be the Official Language of the Union. The Official Language Act also lays down that "both Hindi and English shall compulsorily be used for certain specified purposes such as Resolutions, General Orders, Rules, Notifications, Administrative and other Reports, Press Communiqués; Administrative and other Reports and Official Papers to be laid before a House or the Houses of Parliament; Contracts, Agreements, Licenses, Permits, Tender Notices and Forms of Tender, etc."

(http://india.gov.in/knowindia/official_language.php)

iv. There are ten major scripts in use in India. Roman script is very commonly used by the urban population for writing e-mail and SMS etc in Indian languages. There exists a significant population who know the native language and English but not the native script. Such people prefer to use Roman script for writing the native languages.

v. Text entry and editing in Indian scripts are generally considered cumbersome. Smart user interfaces are needed for man-machine integration.

vi. English is understood by less than 7-10% of the Indian population. However, it continues to be the primary link language of the country and a major resource for new knowledge. Thus English is the major language barrier in the country for knowledge creation and dissemination.

vii. English words, phrases and constructs are very frequently mixed in day to day communication. MT systems have to cater to such mixed language environments.

viii. There is inadequate standardization of terminology for various disciplines in Indian languages. The Indian Commission of Scientific and Technical Terminology (CSTT) has evolved about 6,00,000 terms for Hindi and identified about 25,000 Pan-Indian terms in different fields applicable to Hindi and other Indian languages. Sanskrit has been used as a base for this development. However, this data is found to be inadequate for many applications and domains. As a result, the translators start using non-standard terminology that may be confusing.

ix. Indian scripts are phonetic in nature in the sense that they are written the same way as spoken which is not the case with English and other western languages. Thus the transliteration of named-entities to and from English is error prone. However, transliteration among Indian languages is somewhat straightforward.

x. Indian English is influenced by native language forms and grammar. One very often encounters errors in usage of numbers, narrative forms, interrogative forms and missing articles.

xi. The language divide has significantly contributed to the digital divide. It has also contributed to widening of the social divide. One of the primary reasons for this has been a lack of contents in Indian languages on the web and the corresponding tools.

### 2.2 Challenges as compared to other scenarios.

*Multiplicity of scripts and their phonetic nature:* There are 10 major Indian scripts in use. Many a time a text has a mix with Roman script. There are situations where we

see that people are multilingual but know only their own script. Indian scripts are highly phonetic in nature in the sense that the words are written the way they are spoken. Each Indian script has some specific consonants and vowels that ensure purity of transcription. Thus there are variations in transliterations. This directly affects collection of noise free parallel corpora and also monolingual corpora.

*Free word order:* Indian languages have relatively free word order and a group of words can move to any position in the sentence. As a consequence there is a need for normalization to take care of the equivalences. On the text generation front, certain word orders are preferred to make the generated text more natural.

*Rich morphology, Sandhi and Samaas:* Indian languages are highly inflectional and rich in morphology. Sandhi and Samaas are concepts borrowed from Sanskrit. Sandhi represents co-joining of words and multiple words may join together to form a single word. Samaas refers to formation of noun phrases with noun-noun or adjective-noun combinations. These have an impact on lexical data base creation, morphological analysis and synthesis, parsing, text generation and corpus variations.

*Replication:* Replication of words is a peculiar phenomenon encountered in all Indian languages. Words of any part of speech may get replicated and the meaning of the replicated word is different from the individual constituent word. This impacts lexical data base creation, parsing and text generation. Although, the text generation can completely avoid the phenomenon of replication by constructing equivalent form, it loses the inherent naturalness in the target text.

*Semi onomatopoeia:* Here a word has an addendum of a similar sounding onomatopoeic word. The word group together has a different meaning. The impact is similar to the word replication phenomenon. It is also common practice to use non-sensical words that represent sounds associated with certain actions.

*Predicate verbs and other verb forms:* Predicate verb phenomenon in Indian language is very complex that leads to multi-word expression formation. These need to be detected at the time of analysis and appropriately composed at the text generation stage. In addition, most of the verb forms have morphologically derived verb forms that represent causativity, transitivity etc.

*Localization, honorific and gender markings:* There is a difference in the manner in which numbers, time, seasons are used in Indian and European languages. Many of the Indian languages, while paying respect to elders and dignitaries, use plural form of the verb. Similarly the gender of the person is reflected in the verb. These have direct impact on the text generation process.

*Divergence influencing Indian English:* Indian English speakers often get influenced by their native language and make mistakes in narrations (direct/indirect speech), number, person, determiners, use of articles and modalities. The irony is that many of these mistakes are accepted norms in the community. A translation system has to take these into account.

*Mixed language forms:* In India it is very common to mix English words in the text. Many a time there may be mixture of three languages. Most of the time the foreign words undergo morphological transformation as per the native language. This makes the situation very complex for analysis, lexical data base creation and translation. *Social, political and cultural issues:* Unlike the European Union, India is a

single sovereign country with a union of multiple states with multiple languages and cultures. Our constitution guarantees equality to all of its citizens in choice of profession and place of work. The complex linguistic scenario poses several social and political issues. The choice of second or third language learning is exercised based on employment opportunities and sometimes by circumstances. This leads to mobility of a certain section of people affecting the local social set up. On one hand, use of an alien language provides a unification platform bringing in cultural affinity, language has always been used as an instrument for political advantage and pursuing 'divide and rule' policy. The voluntary and forced migrations from one linguistic zone to another generate a translation requirement ranging from personal communication to schooling of the children.

As a consequence of the above, the methodologies applicable to European languages need modifications and adaptation. The statistical techniques as employed for European languages are not likely to yield satisfactory results as a corpus collected in general may not have an adequate coverage with the variations due to distinctive features of Indian languages.

There is also a brighter side of the picture. Indian Languages are closer among themselves. There is a rich tradition of unifying grammatical studies and logical deduction, Paninian framework and 'Navya Nyaya' are a few examples. Further, there has been intense cross fertilization for a much longer span of time as compared with EU languages. Thus the interlingua approach works much better for the translation among the languages of the sub-family. Further, a number of techniques and methodologies applicable to a language, are also equally applicable to each language of the sub-family. As a consequence, a methodology dealing with a foreign language to and from any individual member of the Indian language sub-family is applicable to all other members of the family. The need for translation from and to a foreign language (other than English) is primarily in tourism, import/export business and in addressing the national/international security issues. This is besides the general translation requirement of scientific, technical, patent and literary documents.

### 2.3 Translation Needs in India.

1. The proceedings of the parliament are translated into all the 22 languages by the interpreters and the source language can be any one of the Indian languages. Thus interpreters have to be employed for every pair of languages. Any member of the parliament can raise a question/answer in any regional language, which needs to be circulated to all members in translated form. All the reports of the parliament have to be prepared in English and Hindi.

2. All Government websites have to be provided in English and Hindi. A three language formula (English, Hindi and state language) is promoted by many of the states.

3. All communications within a state take place in the regional language. The communication with the Union is in English and/or Hindi. The medium of instruction for primary and secondary education is in the regional language. For the higher education and most of the professional courses, English is the medium of instruction. Thus there is a language mismatch when one moves to higher education. This is a typical scen-

ario for the rural and middle income group population in each state of the union. However, the urban higher income people send their children to English medium school right from an early stage.

4. The language used in judiciary at the lower court level is mostly in the regional language whereas the language used at the state High courts and the Supreme Court is English. For the fine interpretation of the laws, acts and statutes, English version is considered as authentic and not its translated version. As a consequence, in case of an appeal to the high Court or the Supreme Court, the judgment and proceedings of the lower courts have to be translated into English.

5. Needless to state that all financial and business sectors require their forms, brochures, manuals translated into the regional languages for promotion of their business. It is mandatory for the Government sector industries to provide annual reports in English and Hindi. The pharmaceutical and drug industry have to provide their drug information in all regional languages. Indian railways pass through multiple linguistic and script zones. All information pertaining to travel and transport have to be multilingual and reservation charts and other name lists have to be transliterated.

6. The process, manufacturing and maintenance industries are usually manned by technicians and laborers whose knowledge of English is very poor or know only the regional language. For example, all railway carriage and wagon maintenance staff members have to be trained and any communication gap may result in compromising the safety of the system. This is so for all plant operations and process flow control where human operators are in the loop. Usually the manufacturers of the equipment and machinery do not provide translation into regional languages. For all imported equipment translation of their manuals into regional languages are required. The defence sector has a special requirement of maintaining confidentiality. The aircrafts, warships and other combat machinery are imported from multiple countries, and the manuals have to be translated without delay before being deployed.

7. The India defence and paramilitary services have personnel coming from all linguistic zones of the country and they may be deployed in any other linguistic zone. Their recruitment, evaluation and training have to be provided in regional languages.

8. The police and criminal investigation sector presents a scenario wherein police stations work in the regional language using the regional script. However, the criminals move very quickly from one linguistic zone to another. Thus the criminal records have to be maintained or at least provided on demand in English or another regional language.

9. The competitive examinations held at the national level like those conducted by the Union Public Service Commission, IITs, IIMs and several others, the question papers have to be set in any one of the 22 languages as per choice of the candidate and the answers need to be evaluated in the regional language on demand. Here the translation process has to ensure both high level of accuracy and security.

10. With the high level of penetration of mobile and handheld devices in the country, multilingual communication for information retrieval, advertising, etc. both in textual and speech forms are needed. The rural literacy rate in the country being low, there is more a need for multilingual speech processing and translation. Currently,

only SMS can be received in regional languages with no translation. The speech to speech translation is still a distant reality.

11. The regional language radio and television broadcasts and vernacular newspapers are popular. However, the provision of choice of language is mostly limited to a few channels such as Discovery, Cartoons etc. The subtitling in the entertainment industry is mostly in English or in Hindi. There is a huge translation requirement here which has great commercial potential. The vernacular newspapers are a good source of corpus provided copyright issues are addressed. The national news reported in these newspapers can be used as a source for parallel text mining.

Thus it is seen that the translation requirements in India range from a simple rough translation to a formal translation that requires a high quality translation. The volume of the requirement is huge and requires a huge investment for clearing the backlog. Some of these are mandatory but still cannot be complied with due to practical constraints. The students undergoing technical, professional courses and higher education require translation of text books and other reference materials that are usually available mostly in English. The minimum that they require is a rough translation (in the absence of the translated text-books) that may lead to understand what is being taught. The question-answering through the web and filtering relevant information fall in a similar category. Tourists also have this kind of a requirement. On the other hand all official communications, bulletins, gazettes, manuals, legal documents etc. have to be translated with a high degree of accuracy, The socially relevant topics such as environment, health, agriculture, vocational training, marketing all have varying translation quality requirements.

### 2.4 MT R & D and translation industry scenario

The work on machine translation in the country started in the early eighties [1,4]. However, even today, the MT researchers and system developers in India are confronted with the problem of coping with limited linguistic resources in terms of corpora. This has led to development of MT systems that are primarily rule-based with some hybridization of examples and limited statistical information. Currently, these systems are being primarily used in the government sector in routine official correspondence, health awareness and limited parliament documents.

The major centres where machine translation research & development is being carried out are IIT Kanpur, IIT Mumbai, IIIT Hyderabad and CDAC Pune. AnglaBharati methodology [6] for translation from English to all Indian languages based on pseudo-interlingua & RBMT and AnuBharati methodology [6] using hybrid EBMT for translation among and from Indian languages have been developed at IIT Kanpur. AnglaBharati technology is being used by some CDAC centres for implementation for different Indian languages. At IIT Mumbai a universal networking language (UNL) based interlingua MT system for English to Hindi [2] has been developed. The group also created a *factored* English-Hindi SMT system with reordering of English sentences [3]. The group at IIIT Hyderabad developed a system called Anusarak [1] which the authors called ‘Machine translation in stages’ for translation among Indian languages. The group also developed a English-Hindi MT system called ‘Shakti’ us-

ing a transfer approach. It had also led a consortium mode project on development of IL-IL multi-part RBMT system named Sampark (http://sampark.org.in). The CDAC Pune group developed a TAG (tree adjunct grammar) rule based English-Hindi MT system called MANTRA (www.cdac.in/html/aai/mantra.asp). The group led a consortium mode project on English-IL multi-engine (RBMT, EBMT,SMT) paradigm based MT system whose performance is dominated only by RBMT system.

As far as research & development on speech-to-speech translation is concerned, the major bottleneck is automatic speech recognition (ASR). Real-time translation with almost instantaneous response is another bottleneck. Indian ASR technology is still very "fragile", and it is sensitive to noise, mismatch in train and test conditions, speaker variability, accents, dialects, etc. There is even a big difference in performance between "read speech" and "spontaneous speech". There is wide geographical variation in speech even in Indian English and Hindi. For example, Hindi as spoken in Delhi is not the same as in Hyderabad. Besides this, in accent there is an influence of mother tongue speech. We do not have good transcribed data for Indian languages with speech collected in natural mode with accent/dialectical variations. Most of the systems developed are lab systems rather than practical systems. ASR would give decent output for translation, if we constrain the task and actually collect speech data for that task and build systems that exploit domain context. The speech-to-speech translation in such limited domain can start only when ASR attains a certain level of maturity.

As we see from the above description, our major hurdle in MT R & D is non availability of appropriate linguistic resources and tools. There is an initiative to have an Indian LDC, however this is a gigantic task involving both manpower and money. High performance tools for POS tagging, morphological analysis, parsing, named entity recognition, verb frames, wordnet are needed. Lack of standardization inhibits sharing of resources among different research groups. Further, there is an acute shortage of trained manpower to work on language technology projects. The universities and colleges are not having appropriate curricula to support this. There are less than 5% of CS and IT graduates who undertake thesis work in this area in the country. Besides this, the IT professionals find other jobs more lucrative. Although there exists very sound traditional linguistic knowledge in terms of grammar formalism and logic, there is a wide gap in exploiting them in computational framework. The computer science and technology professionals are not exposed to this valuable heritage of the country and the scholars possessing this knowledge are not familiar with computational formalism needs. There is very limited funding available for research and most of the funding available (primarily from the Government sector) has the pressure of delivering time bound product oriented technology. There is hardly any participation from private industry.

The Indian translation industry is still in its infancy and is largely dependent on the part-time free-lance translators. While some of them use CAT tools, their usage is limited due to their cost effectiveness and applicability to Indian languages. The industry is also confronted with lack of standardization in usage of terminology, difficulty in dealing with non-Roman scripts and inadequate training. None of them are actually using a machine translation system. This is primarily because the MT system

options available to them have limited performance, and they do not find them attractive or cost effective. Many of them still use paper and pencil for preparing initial draft and for editing. There is also a lack of standardization in the work flow.

## 3 Issues in HT and MT and integration processes

Whether a text can be translated by humans or machines, or in a HMI process, is driven by several factors, some of which are as under:

*Critical vs. casual contents:* MT is acceptable for casual translation for personal use as it is fast and involves no intermediary. HT or HMI is invariably used for critical documents such as books, reports, legal, formal communications, etc.

*Volume vs. quality:* There cannot be quality assurance with MT. In case of a large volume quality translation, a large backlog is a usually observed phenomenon as it involves HT and sometimes with multiple iterations. MT can handle large volumes but may not be acceptable for many applications.

*Secure vs. unsecure MT:* If one uses an on-line or a server based MT such as Google translate, the user's data is unsecure as it becomes available to the developers for further development and may be available to other competitors. The situation is similar when using crowd sourcing or social networking for obtaining or evaluating translations. The translation industries do not share translation memory or other data to maintain non-disclosure to their clients.

*Data driven vs. work-flow driven:* MT is primarily data driven, whereas HMI is highly work-flow driven. Due to lack of standardization in HMI work-flow, pipelining or sharing of the task becomes difficult.

*Cost effectiveness and User friendliness:* The investments in MT systems have to be cost-effective and should reach the break-even point within a short span of time. The degree of ease of operation, training requirements and the translation workbench environment dictate the acceptability of an MT system. Their integration in the work-flow and awareness are important.

*HT and MT integration processes:* Use of controlled language through pre-editing, automating human pre-editing/post-editing through machine learning using the data at different stages of the translation process, creating machine-aids based on understanding of the human translation process where he/she has to pay more attention are some of the ways in which the HT-MT integration can take place. Each of these when automated provide improvisation to MT engine performance and reduce HT efforts. The entire process can be bootstrapped.

## 4 HMI in an Indian Scenario

An increase in the translation throughput and improvisation through learning is one of the major outcomes of the HMI process. Another major outcome is production of clean aligned corpora that find application in deriving translation memories and phrase level ready translation pairs. This data is very valuable for effective implementation of SMT and EBMT paradigms. These are general outcomes of the HMI cycle, which is essentially universally applicable to all the language pairs. It is the

available resources (both language & technology) that make the difference in its effectiveness. In fact, given the Indian scenario HMI appears to be the only way to meet the translation demands in India. The following paragraphs summarize, why and how HMI is more suited specifically to the Indian context:

i. Since the concept of spelling in Indian language is somewhat loose, there is a large number of variations in writing the same word and in transliteration of named entities by different translators. For any word alignment task, this becomes sparser. Through HMI preferences and equivalences could be recorded and the machine can be used to normalize this.

ii. Entry of texts in Indian scripts is comparatively a more complex task and sometimes error prone. In HMI, the copy and paste mechanism reduces this effort.

iii. The Indian language texts may not be stored in Unicode / UTF-8 coding form. HMI environment offers a way by which code converters can provide normalization.

iv. Many a time the same text needs to be translated into multiple languages. Since Indian languages are much closer to each other both in structure and lexicons, the text translated into say Hindi may be used to translate into another Indian language by a translator knowing Hindi. HMI offers a way to use such cross lingual information in post-editing.

iv. There is inadequate standardization in usage of terminology. Translators tend to use transliterated terminology or use non-standard substitutes. Through HMI, terminology could be substituted appropriately based on preference and usage. Transliteration variations due to the manner they get pronounced in different regions is also tackled using HMI.

v. In Indian language texts, it is very common to mix English and other language which may appear in un-transliterated form. This usually happens when the meaning of the source language word is not known to the translator. This is a very common phenomenon observed on Google Translate. Interestingly, very often the unknown words get morphologically transformed. As an example, ‘blessed’ in Hindi may be translated as ‘bless kiyaa’. Through HMI, meanings get gathered and lexical database augmented.

vi. There is a shortage of competent manpower in the translation industry. The human pre-editing, MT and post-editing outputs in the HMI cycle can be used both for training and for machine learning. Multi-tier post-editing performed by novice to expert translators provides a mechanism for acquiring skills.

vii. The common errors in Indian English and missing articles handled through HMI cycle provide valuable information for machine learning.

viii. Many a time people know the language but not the script. In all such cases, they tend to use Romanized script. HMI offers a way in which it can be converted to script of choice before pre-editing.

Keeping the Indian scenario in view, a HMI translation process framework [5] is proposed and is shown in Figure 1. The translation system is assumed to be equipped with a centralized MT server or a cloud computing environment. It has a distributed nation-wide network of professional translators and apprentice-translators. Crowdsourcing over this network will provide a broad platform even for training, certification, standardization as well as employment opportunity to educated unemployed. The

different components/modules of the system with generated databases, human interfaces, machine aids and machine learning have been marked.

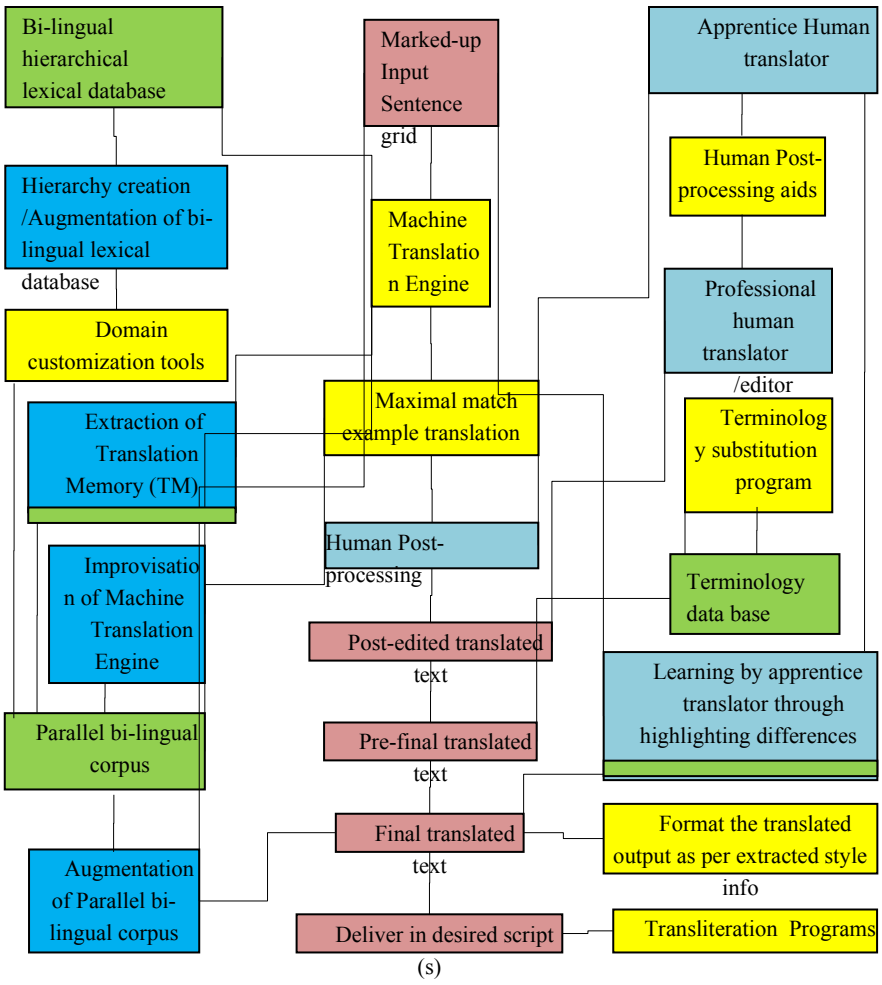


**Fig. 1.** The overall translation process with man-machine integration

There is a tendency to use the English terminology words as they are, in transliterated form. The mappings of terminology including their transliteration have a lot of regional influence. Keeping these in view, the methodology allows liberty at the regional level. Ultimately, as and when the standardization is formalized, these can get substituted.

The entire process generates a domain specific bilingual corpus. The corpora generated through this process are much more relevant and noise free. This is used for enriching domain-specific lexical database and extracting translation memory. It is also used in building a translation model for use in SMT and for EBMT. The target corpus

generated is used in language modeling in the specific domain. This language model is useful in building tools for automated post-editing. This in turn improves the MT system performance and results in reduction of HT effort in the next HMI cycle.

The entire translation process in this framework is pipelined with participation of varying skills at different stages. The skill needed at the stage of extracting text zones is only that of scanning and image handling with no linguistic expertise needed. The tasks of marking up the extracted text zones, isolating sentences/simplification/pre-editing and identifying text components primarily need the knowledge of the source language. A pool of the marked-up text can be generated and made available for experimentation to different machine translation paradigm. The apprentice translator has to be a bilingual person but need not be a translation expert to begin with. At the post-editing stage, the editor need not be a translator.

## 5 Conclusions

The machine translation research & development in India is faced with many challenges, and the Indian translation industry is in its infancy. Man-machine integration in the translation process is very much required both to meet the translation demand and to provide impetus to MT research in the country. This is the only way to take the MT systems from lab to users. Keeping the Indian constraints and state of art in view, a multilevel framework for man-machine integration is presented with a bootstrapping mechanism. In order to make this successful, cooperation from different sectors, private as well as Government, and their partnership in revenue modeling will be crucial. The experiences of the European Union community with a similar multilingual scenario are very much relevant in this context. Similarly Indian experiences in dealing with the constraints and exploiting homogeneity within a language family are relevant to the European Union researchers and developers.

## References


1. Bharati A., Sangal R., Sharma D. and Kulkarni A.: Machine translation activities in India: A survey, Published in the Proceedings of workshop on survey on Research and Development of Machine Translation in Asian Countries, Thailand, May 13-14, (2002).
2. Dave S., Parikh J. and Bhattacharyya P.: Interlingua Based English Hindi Machine Translation and Language Divergence, Journal of Machine Translation (17), September (2002).
3. Ramanathan A., Choudhary H., Ghosh A. and Bhattacharyya A.: Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT, ACL-IJCNLP 2009, Singapore, August, (2009).
4. Sinha R.M.K.: A Journey from Indian Scripts Processing to Indian Language Processing. IEEE Annals of the History of Computing, Jan-March: 8-31,(2009).
5. Sinha R.M.K.: Indian National Translation Mission: Need for Integrating Human- Machine Translation, (http://www.mt-archive.info/MTS-2009-Sinha-1.pdf), Proceedings MT Summit XII, Aug.26-30, (2009), Ottawa, Canada. `
6. Sinha R.M.K.: An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, Invited Paper, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004), November 17-19, (2004), Tata Mc Graw Hill, New Delhi.