Indian National Translation Mission: Need for Integrating Human-Machine Translation

R. Mahesh K. Sinha
Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur
Kanpur 208016, India
rmk@iitk.ac.in

Abstract

A National Translation Mission (NTM) has recently been launched by Government of India with the primary objective of providing greater access to knowledge across the country in all Indian languages through translation of book and documents. In this paper, I present a brief account of the current Indian linguistic scenario and outline a framework for integration of human and machine translation that will provide impetus to MT researchers, developers and the translation industry in the country.

1 Introduction

The linguistic scenario in India is very complex. India is a highly multilingual country with 22 constitutionally recognized languages. Besides these, a large number of other languages that are not included in the recognized list, are used by millions. Added to this, there are ten different prominent scripts that are in use. In addition, India's languages have undergone significant mixing and cross-fertilization. Interestingly, the English language, which is understood by less than 3% of the country's population, continues to be the major link language between federal and state administrations. English is used in the country's higher-education, most of the business houses and is mandated as the authoritative language for federal laws and Supreme Court judgments.

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. The students, undergoing technical, professional courses and higher education, require translation of text books and other reference materials that are usually available in English. The minimum that they require is a rough translation that may lead to understand what is being taught in absence of the translated text-books. The question-answering through the web and filtering relevant information fall in a similar category. Tourists too 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.

The Indian translation industry is still in 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. The MT researchers and system developers are confronted with the problem of coping up with limited linguistic resources such as uni-lingual and bi-lingual corpora and other linguistic tools. This has led to development of MT systems that are primarily rule-based with some hybridization of examples and limited statistical information. Cur-

rently, these systems are being primarily used in government sector in routine official correspondence, health awareness and limited parliament documents.

Keeping the above scenario in view, the National Knowledge Commission (NKC) as constituted by Government of India (http://www.knowledgecommission.gov.in/recommendations/translation.asp) made a following observation:

"We recognize that there is an urgent need for expansion of quantity and improvement of quality of translation of different types (human, machine-aided, instant, etc.) and in different domains (literary, scientific, technical, business, etc.) that would provide greater access to knowledge across the country. The current facilities available are inadequate and less than socially required. There is latent unrecognized demand which is not being met because of incomplete and asymmetric information. Inadequacy of information compounded by the lack of coordination between potential users, also leads to market failures. Further there is inadequate dissemination of good quality translations which would provide a benchmark and create incentives for most private activity in this area. Therefore, this requires some amount of public intervention, not as a permanent feature, but as a set of measures to kick-start a process of encouraging private initiative such that the large commercially viable provision of high quality translation in different areas becomes feasible. The direct and indirect employment generation potential of translation activities is very high, and could absorb a substantial part of educated unemployed youth."

This led to a Government of India initiative leading to the formation of the National Translation Mission (NTM) (http://www.ntm.org.in/ntm). The primary goal of NTM is stated to be

"to make knowledge-based texts accessible in all Indian languages through translation. The idea of NTM stemmed from a statement of the Prime Minister of India stressing how vital is the access to translated material, for increasing access to knowledge in many critical areas."

In this paper, I present a framework aimed at integrating MT in the human translation process for enhancing translation throughput, bootstrapping MT capability, translators' training, tool deployment, employment generation, and finally providing access to knowledge in native language(s) to socially deprived masses. The framework is to provide impetus to the Indian translation industry which is in its infancy. This is motivated keeping in view the objectives of NTM and the author is a member of the project advisory committee (http://www.ntm.org.in/ntm_PACmembers.asp)

2 Indian National Translation Mission

As pointed out earlier, the National Knowledge Commission of India having recognized the critical role of translation in making knowledge available to different linguistic groups in a multi-lingual country like ours, has recommended setting up a National Translation Mission (NTM). Some of the major recommendations of the commission on NTM objectives are reproduced below from their site:

- "Provide impetus for developing translation as an industry in the country.
- Establish a store-house of information on all aspects of translation involving Indian languages, and to make this available by creating, maintaining and constantly updating information on translations published, training programs, translation tools/instruments and new initiatives, and facilities such as a 'National Register for Translators'.
- Promote printed as well as virtual publication of works on translation studies.
- Create and maintain various tools for translation.
- Provide quality training and education for translators.
- Translate pedagogic materials at all levels (including primary onwards to tertiary education) specifically in natural and social sciences.
- Project Indian languages and literatures within South Asia and outside through high-quality translation.
- Set up a national web portal on translation
- Organize annual National Conferences on translation.
- Promote book launches, festivals, fellowships and prizes etc. and encourage collaborative translation work, as well as long-term multi-translator projects, and organize workshops for translators to interact and exchange views and experiences."

The Government of India has approved the setting up of a National Translation Mission with an outlay of Rs. 739.7 millions for the Plan period. NTM has started registering translators, identifying books to be translated and examining copyright issues. It is in the process of making road maps for achieving the above objectives. Obviously such a gigantic complex task cannot be accomplished only by human translators. There is a need to integrate machine translation and CAT tools with the human translation. This is besides the management skills needed to execute the project. This paper is concerned only about the aspects of integration and their fallouts in the context of Indian linguistic scenario.

3 Indian scenario

- There are 22 officially (constitutionally) recognized languages. In addition, there are about 122 other living languages each of which is being used by more than a population 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.
- There are ten major scripts in use in India. Roman script is very commonly used by 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.
- Text entry and editing in Indian scripts are generally considered cumbersome. Smart user interfaces are needed for man-machine integration.
- All Indian languages have a common origin and are structurally similar to each other. They are verb-ending (SOV) and are relatively free word order. The translation methodology should be designed exploiting this.
- English is understood by less than 3% 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.
- English words, phrases and constructs are very frequently mixed in day to day communication. MT systems have to cater to such mixed language environments.
- 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 (applicable to all Indian languages) terminology in different fields applicable to Hindi and other Indian languages. 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.
- 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.
- 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.
- The language divide has significantly contributed to the digital divide. It has also contributed to widening of the social divide. The internet usage stands at a meager 2.72% of the Indian population. Besides the economic factors, one of the primary reasons for this has been a lack of contents in Indian languages on the web and the corresponding tools.
- There is a sizable urban population of educated unemployed in the country. A majority of educated married women in India are homemakers and their potential for contribution to knowledge sector remains untapped.

4 Integration of HT and MT

4.1 Aims of HT and MT integration

There are four primary aims of HT and MT integration framework proposed here:

- Reduce human efforts needed for editing the machine translated text.
- Create resources for MT researchers and developers for further research and development of CAT tools.

- Provide an automated mechanism for collecting and analyzing users' feedback for MT researchers and developers.
- Increase the overall translation throughput.

4.2 Levels of HT and MT integration

A typical translation process usually goes through a number of stages before the final delivery of the translation of the document. At each one of these stages, there is a scope of reducing human efforts using machine and has potential for man-machine integration. These are listed below at a broad level:

- Text zone extraction
- Text correction
- Subject domain identification
- Source language identification
- Foreign word identification
- Sentence isolation
- Pre-editing and simplification
- Identification of nature of the sentence
- Identification of nature of word/word-group categories
- Transliteration of names
- Dealing with Terminology
- The translation process
- Post-editing
- Target document in desired script
- Target document composition
- Translation in cyberspace and crowd sourcing

Figures 1 to 5 depict some of these stages diagrammatically with linkages to machine aids (tools), databases, human interaction and machine learning. It should be noted that a rule-based MT system is assumed here and some of the modules become irrelevant for other MT paradigms.

Learning automated text zone extraction from a document consisting of multiple columns, headings, headers, footers, images etc based on modeling of the document can be deployed. The document model is created and modified based on users' feedback. (Fig.1).

After extraction of the text zone, the text has to be corrected and marked up for foreign words and mixed language constructs. Source error modeling is used to improve the spelling correction process. A number of techniques based on heuristics, statistics or morphological analysis are employed to identify the language and foreign words. (Fig. 2).

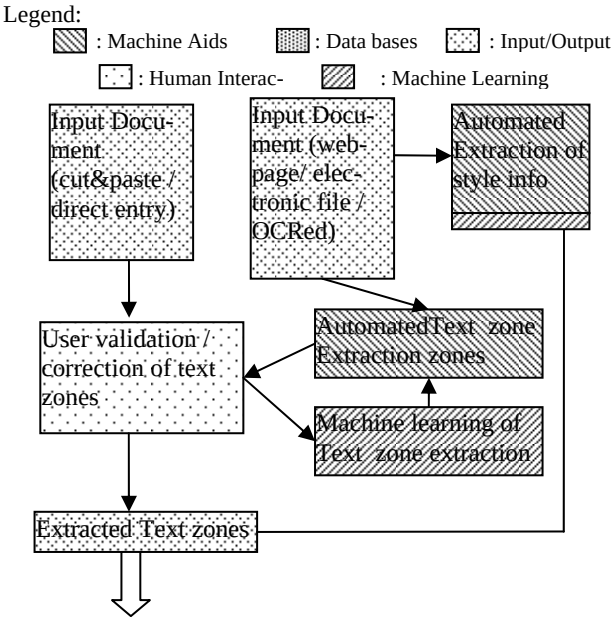

Fig.1: Text zone extraction

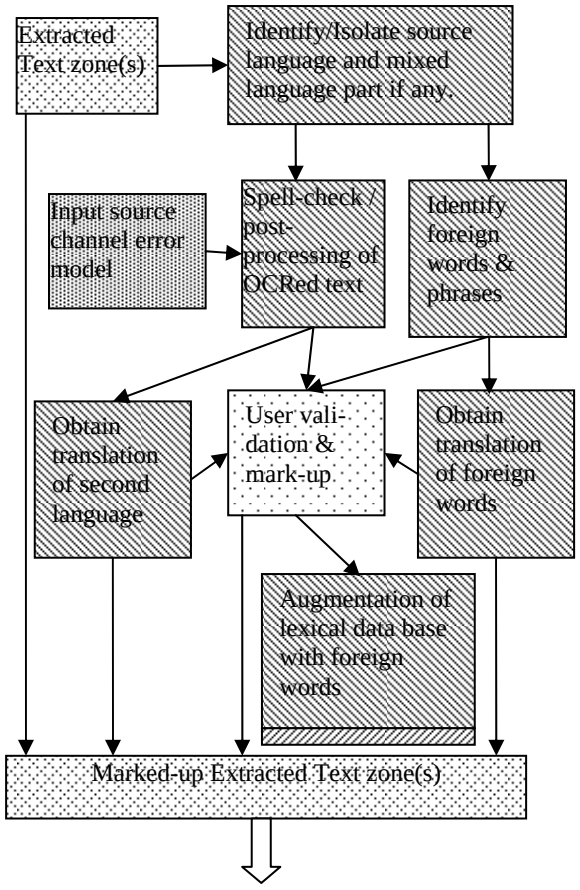

Fig.2: Text correction and foreign word & language identification

The marked up text now has to be split into components and the component types are identified. Keeping in view the limitations of MT system, the long and complex sentences are attempted to be simplified. There is a lot of scope of learning and automating these processes. To the least, the machine may prompt the user for probable ambiguity where pre-editing may be helpful. A log of the original and the simplified sentences is a very useful resource for automating the simplification process. (Fig. 3).

The next stage is to mark up each component of the text that are identified as different translation units for which a separate mechanisms for translation may be required to be invoked. Such units include identification of headings, named entities, acronyms etc. The rules for identification of these units are framed and learnt based on heuristics and modified with users' feedback. Transliteration of named entities, terminology and acronyms is another task where machine learning on created data could be employed. (Fig. 4).

Now the input text is ready to be translated. Fig. 5 depicts the overall translation process. It is proposed that the machine translation systems will be a dedicated network of MT servers catering to specific groups of languages. The professional and apprentice translators can get registered with the sites and can use all the resources at the site. Each professional translator may have a number of apprentice translators. The machine translated text is first examined and post-edited by an apprentice translator. This is further examined by the professional translator and a pre-final version of the translated text is generated. It is called pre-final because the terminologies are not yet substituted. Due to inadequate standardization of terminology or user preferences of keeping the English terminology as such, the terminology substitution is applied as per users' specification. The final translation is formatted as per the original text. This may be delivered in the desired script using machine transliteration.

The hierarchy of apprentice and professional translators provides a mechanism for automatic gradation of apprentice translators by comparing the post-edited outputs. For the apprentice translator this data is useful for self training.

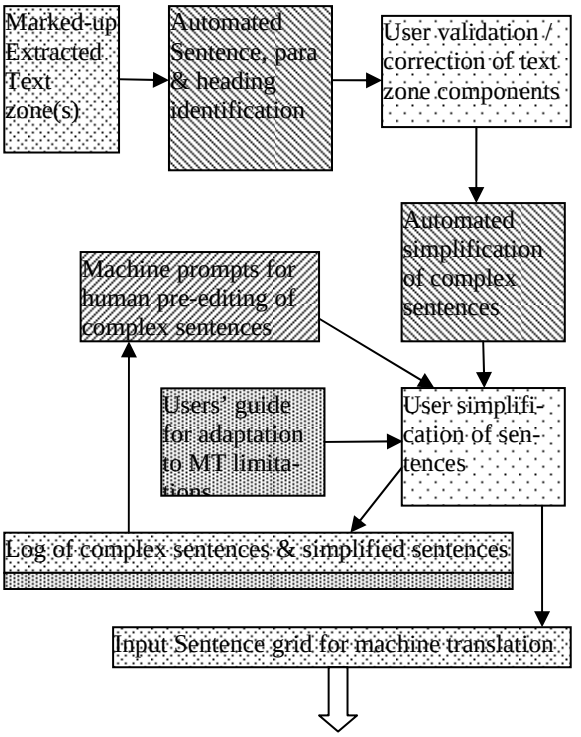

Fig.3: Sentence isolation, simplification and grid creation

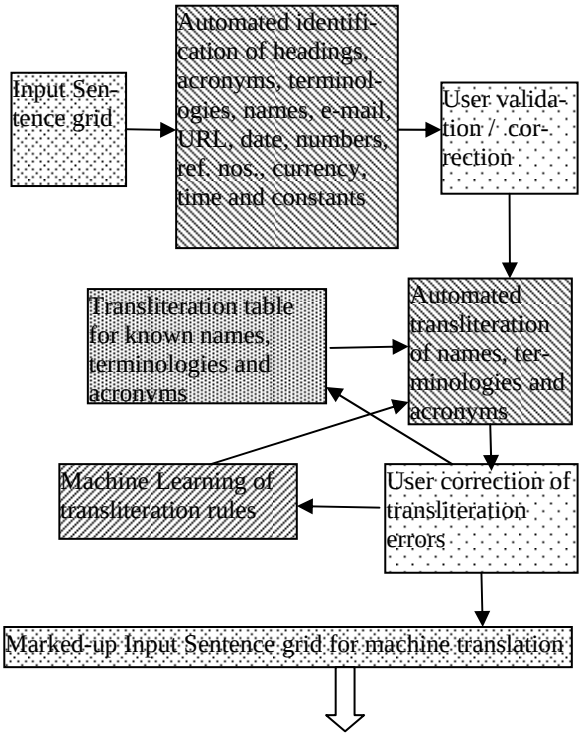

Fig.4: Identification of sentence/word categories and transliteration of name

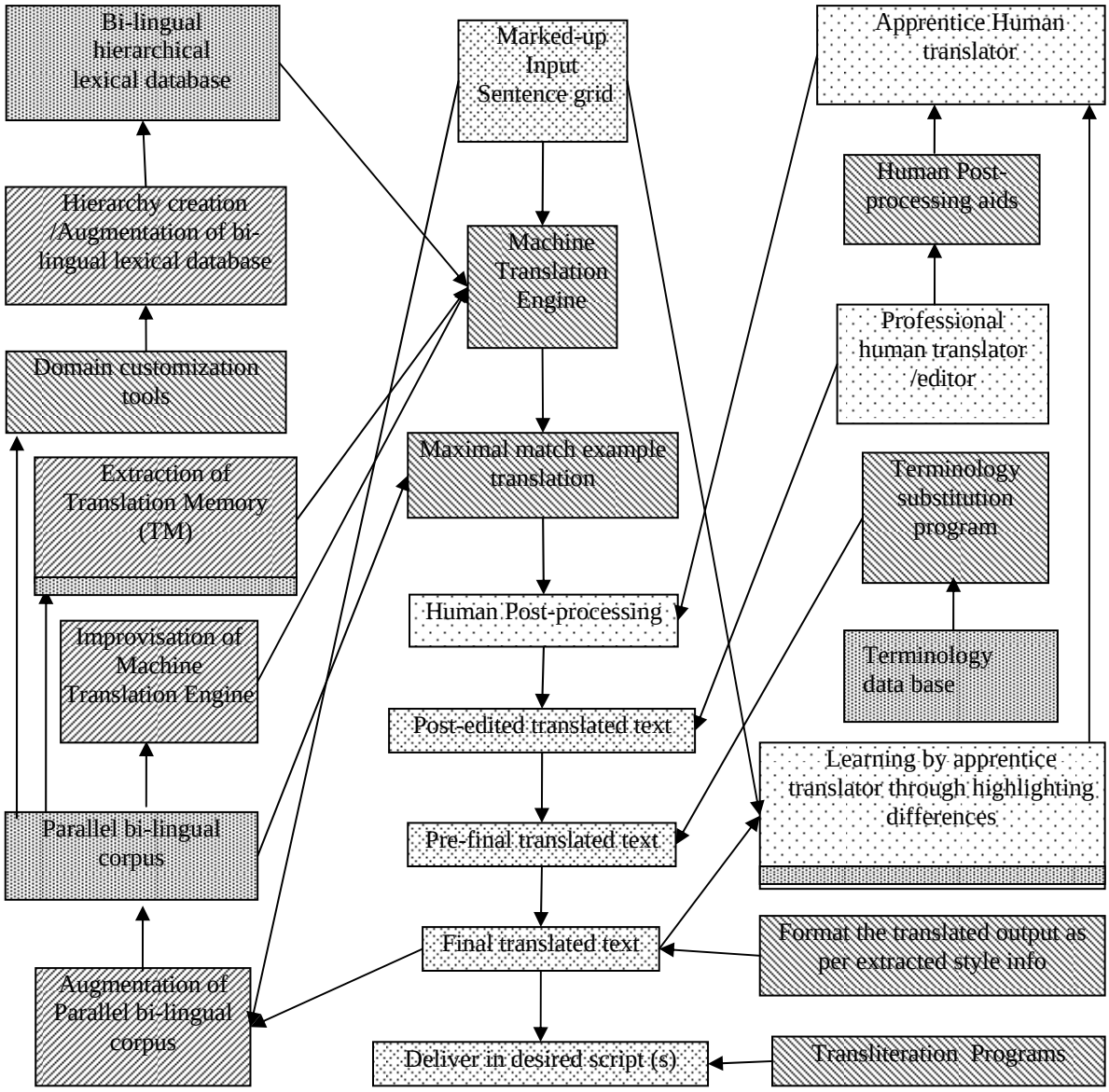

Fig.5: The overall translation process

The entire process generates a domain specific bi-lingual corpus. This is used for enriching domain-specific lexical database and extracting translation memory. It is also used in building 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.

Crowd sourcing is another way by which limited parallel corpus is generated and at the same time data highlighting the limitations of the machine translation engine gets collected.

Here, it is worth mentioning that the corpora generated through this process are much more relevant and noise free. There have been several projects on parallel corpora generation and our experience has not been very good. The primary reason for this is the errors that creep in data entry

both at the source and target languages. These errors in source and target languages have no correlation with each other. The translation agencies are not ready to share these data. It is proposed to build a national reservoir of all the relevant data generated out of the translation process with the seed financial support of the government.

Further, 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 (Fig.1) is only that of scanning and image handling with no linguistic expertise needed. The tasks of marking up the extracted text zones (Fig.2), isolating sentences/ simplification/pre-editing (Fig.3) and identifying text components (Fig.4) primarily need the knowledge of the source language. No translation expertise is really needed. 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 bi-lingual person but need not be a translation expert to begin with. At the post-editing stage, the editor need not be a translator but only a target language expert.

4.3 Pre-editing: User adaptation to machine limitations

The general users have a very high level of expectation from a MT system without understanding its limitations. We deploying our MT systems, we observed that the professional translators were not enthusiastic about it. They felt that unless the MT system yields a human competing performance, it is not useful to them. On the other hand, the Hindi officers at the Government departments were receptive to learning adaptation to MT limitations and were able to improve their translation throughput for annual reports by about 30%. It is worth mentioning here that the task of pre-editing is almost like re-writing the text such that it could be easily understood by a person with medium level language skill. The pre-editing task need not be done by the translators. A language expert can do this once (s)he is made aware of the target readers. As a by-product of this process, we get the book/manual/document produced in a much simplified language. Given below are some guidelines on how one could get better results from the machine when the source language is English:

1. For human translators, ungrammatical constructions to a certain degree are tolerable because the humans can derive the right interpretation based on their world-knowledge. For machines, it is quite problematic especially when it is a rule-based system. Make sure that input sentences are grammatical. Use punctuation marks wherever needed grammatically.

2. Coordinate conjunctions are another source of a number of ambiguities that are hard to resolve by a MT system. Care should be taken in their usage by explicitly giving their scope wherever possible.

3. The English words ending with -ing is another source of ambiguity in MT. A word ending with -ing can be either a (gerund) noun or an adjective or a participial verb. It is advised that in case –ing words are used as nouns, use an article before it or transform it to an infinitive form. For example, the sentence ‘Flying plane can be dangerous’ should be re-written as ‘To fly a plane can be dangerous’ or as ‘The plane which is flying can be dangerous’ depending upon the actual meaning.

The –ing words that play the role of an adjective may be rewritten as a relative clause or a suitable preposition may be added. For example, rewrite ‘machine using oil’ as ‘machine by using oil’ or as ‘machine that use oil’ based on what is actually meant.

4. The relative pronouns such as *that, which, who,* etc should be written explicitly.

5. Avoid use of personal pronouns as far as possible.

6. The complementizer *that* should be written explicitly.

7. Do not use long noun phrases as far as possible.

8. Substitute infinitive *to* with *in order to* in a purpose clause instead of just *to*.

9. Avoid using phrasal verbs (verb+particle) wherever possible by substituting with one-word verbs. For example, ‘went on talking’ be rewritten as ‘continued to talk’.

10. Simplify long sentences wherever possible and expand short sentences to include implied references. Make sure that each segment can stand alone syntactically.

11. Avoid passive constructions wherever possible.

12. Do not use ellipsis.

13. Avoid using / as in and/or and user/system.

14. Replace the polysemous preposition by appropriate unambiguous preposition, wherever possible. For example, ‘Advisory Council to the Prime Minister’ be replaced by ‘Advisory Council of the Prime Minister’.

15. If a reference to a noun phrase is made through the head noun of the noun phrase, replace the noun with an appropriate pronoun.

16. Use hyphenation in ambiguous collocation/ multiple word expressions. Alternatively, write these two words as a single word.

17. Change to-infinitive clauses after the verb to prepositional phrases according to their adverbial function. For example, rewrite ‘went to buy’ as ‘went for buying’.

18. Use honorific markers overtly. In most of the Indian languages, the subject in singular number with honorific semantic tag triggers plural inflection on the verb. To identify the honorificity of the subject nominal is not always an easy task.

19. Replace polysemous words/phrases with monosemous words/phrases, wherever possible. For example, rewrite ‘pen such a poem’ as ‘compose such a poem’.

20. Use simpler counterparts for the unusual use of have/had construction, if possible. For example, rewrite ‘has had a successful run’ as ‘has been a successful run’.

21. Avoid using constructions with stranded preposition. For example, ‘Where are you coming from?’ be rewritten as ‘From where are you coming?’.

22. Avoid using dummy there- and it- constructions, wherever possible.

23. Avoid using all the upper case, italic, underlined lexical elements for the purpose of emphasis or style markers. These cause problem in identification of acronyms and foreign words.

24. Use simpler counterpart for have/had construction, if possible. For example, rewrite ‘had an accident’ as ‘met with an accident’; ‘had breakfast’ as ‘ate breakfast’.

25. Do not use ‘s’ to make plural of any arbitrary construct.

4.4 Bootstrapping machine translation system performance

Initially the human effort required in the HT-MT integration framework as outlined above is large. However, this effort gets reduced as the resources generated in the translation process lead to improvement in performance of CAT tools used and development new CAT tools. With the availability of domain specific parallel corpora, it will be possible to develop EBMT, SMT, hybrid and multi-engine paradigm. Any shortcuts, such as contractual translation of books or documents, will not only be expensive but also not be able to provide inputs for further research and development.

The translation systems distributed nation-wide in the cyberspace with a network of professional translators, apprentice-translators and crowd-sourcing will provide a broad platform for training, certification, standardization as well as employment opportunity to educated unemployed.

5 Conclusions

Creation of knowledge in native languages is vital for the economic growth, employment, prosperity, good governance, bridging the societal gaps and peace within a country. Internationally, it contributes to a better mutual understanding, international trade and peace. The Indian linguistic scenario presents a unique challenge to information technologists and the Indian NTM initiative will hopefully provide valuable insights both to the technologists and the social scientists.

Acknowledgments

I am thankful to Mr. V.N. Shukla, Director Special Applications, CDAC Noida for providing inputs and users’ feedback on deployment of our machine translation system.

References

R. Mahesh K. Sinha. 2009. A Journey from Indian Scripts Processing to Indian Language Processing. *IEEE Annals of the History of Computing*, Jan-March: 8-31. [Readers will find a number of references here on Indian language computing]