2011 International Conference on Languages, Literature and Linguistics
IPEDR vol.26 (2011) © (2011) IACSIT Press, Singapore

A step towards Human–Machine unification using Translation Memory and Machine Translation System

Nishtha Jaiswal[1], Renu Balyan[1] * and Anuradha Sharma[1]

[1] Centre for Development of Advanced Computing(C-DAC), Noida, INDIA

Abstract. It is a well known fact that development of fully generalized automatic machine translation systems is still a dream that seems too far from reality to be fulfilled. A Machine Translation (MT) system with human intervention is the most appropriate way to get translation which has greater human acceptability. The main aim of this paper is to show how semi-automatic method for translation reduces the human effort, and increases efficiency of the MT system. Most of the translation work these days relates to technical material like contracts, user manual, field reports, etc. Technical translation calls for consistency (terminological, structural, phrasal, etc.) among documents and translation accuracy. Also it is time consuming and expensive to train technical translators. So there is a need for human-machine unification and this paper focuses on an approach which can help to achieve this. This paper also focuses on the need for integration of existing MT systems with tools like (Wordfast, Trados etc.) which are widely used for translation by the translation industry. Currently the approach has been applied on AnglaMT system for English to Hindi and English to Punjabi machine translation systems. This approach can however be applied to other available MT systems as well.

Keywords: MachineTranslation, Translation Memory, AnglaMT, Wordfast.

1. Introduction

AnglaMT is a rule based MT system based on AnglaBharati [1] approach. AnglaMT translates text from English to Bangla, Hindi, Malayalam, Punjabi and Urdu. An MT system can never produce translations that are as good as human translator. Although a number of AI techniques have been used for development. A rule based MT system at times fails or does not produce the required translation due to the non-conformance of grammar or lack of information needed by the system to produce translation. In such cases, the system produces only word level translations. The system does not learn from previous translations. Thus there is a need to make the system intelligent enough to automatically learn from new translations, thereby reducing human effort. Translators rarely translate completely new documents. Generally the texts that are to be translated have been seen before. So, a memory of good translations that can be accessed easily would be an ideal aid to the translator. A Translation Memory (TM) tool that can provide this feature can prove useful in such a situation. TM tool has the ability to automatically pair sentences or passages from translated documents with high accuracy. Wordfast can be used to create a TM to start with and later segments can be added based on non-availability of data in the TM using the translation output obtained from the MT system.

A typical TM System consists of two parts: translation memory, which records bilingual sentence pairs as examples; and a search engine, which searches the most similar example(s) from memory. TM provides the Target Language (TL) part of the best matched example to users for post-editing, and other matched examples could be provided to users as suggestions for translation. This paper presents a framework to include TM in existing MT system for which a GUI has been developed. The system translates previously seen inputs with the help of previously generated translations. It checks for the consistency of the translation

[*] Corresponding author Tel.: + (0120-2402551); fax: + (0120-2402569).
E-mail address: (renu17775@gmail.com).

based on previously translated data and minimizes translation effort. The human effort is being used only for post- editing and proof-reading the translations for maintaining readability and accuracy.

The rest of the paper is organized as follows: In Section 2 we describe the methodology for integrating translation memory with the machine translation system. Section 3 discusses the future work that can be further taken up and conclude the paper.

2. Integration Methodology

In order to reduce human effort for translating the same set of data again and again we propose that Translation Memory can be implemented in two ways. First by integrating TM into AnglaMT system to deal with the issue of low resources and incorporate reusability of previous translations and second by adding existing AnglaMT system as MT resource in Wordfast for local and web access. An attempt has been made to implement the first approach and it is discussed in Section 2.1. However, the second method as discussed in Section 2.2 is being proposed as a measure to improve the usability of the existing AnglaMT system and is under the development phase.

2.1. TM Integration with the AnglaMT System

The flow of the proposed approach has been shown in Fig.1.The input sentence at the AnglaMT interface passes through the Segment Matcher. The output of the Segment Matcher is a match or no match. The segment to be matched with the translation can be a complete sentence or a part of sentence. This is based on the matching criteria set in the Segment Matcher. If the source sentence matches exactly or more than 80% then the translation output is extracted from the TM and given as output. If the segment matcher gives no match or it is <80% match for source sentence then the normal machine translation process takes place and AnglaMT engine is used for translation. Multiple alternatives are produced and the best output is selected, post edited and updated in TM database if there is a need to update the TM contents.

The process can be further divided into the following three sub-processes:

- TM Build process
- Segment Match process
- TM Update process

2.1.1. TM Build process

The first step is to build the Translation Memory. This step has not been explicitly shown in Fig.1 as this part can be skipped and one can simply assume a blank TM to start with and store translations as obtained and post-edited from the MT system. Building high quality TM is an expensive and time consuming process [2]. To start the cycle we have used Wordfast automatically to build TM. A set of 500 sentences have been translated manually for Hindi and Punjabi with the help of Wordfast to create a TM in TMX format. A Sample TM for Hindi is shown in Fig.2 and its corresponding TMX file is shown in Fig.3.

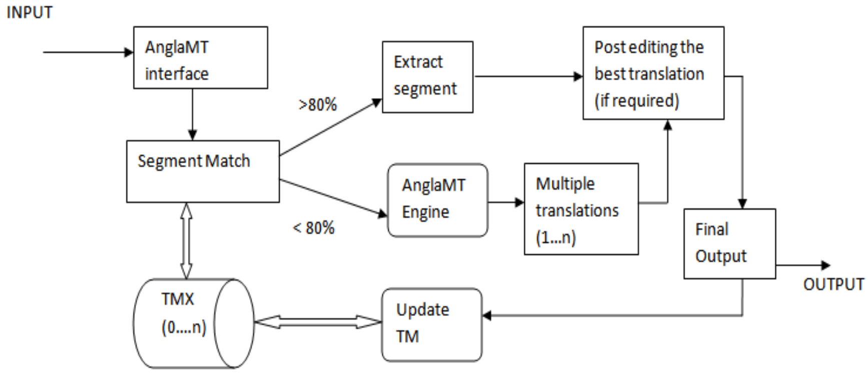
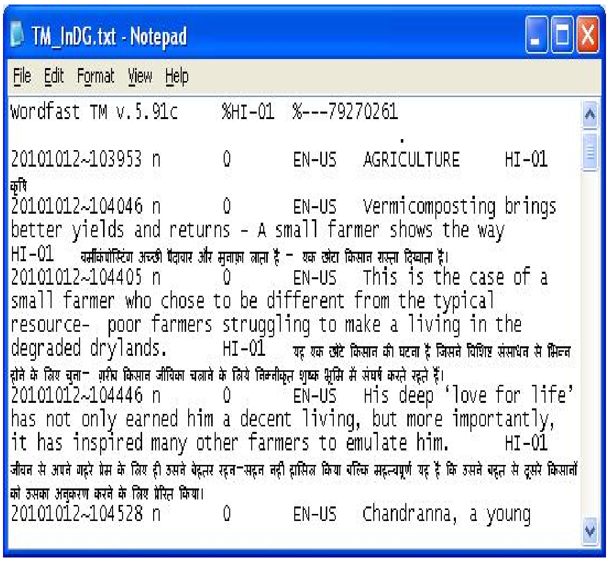


Fig. 1: System Flow.

2.1.2. Segment Match Process

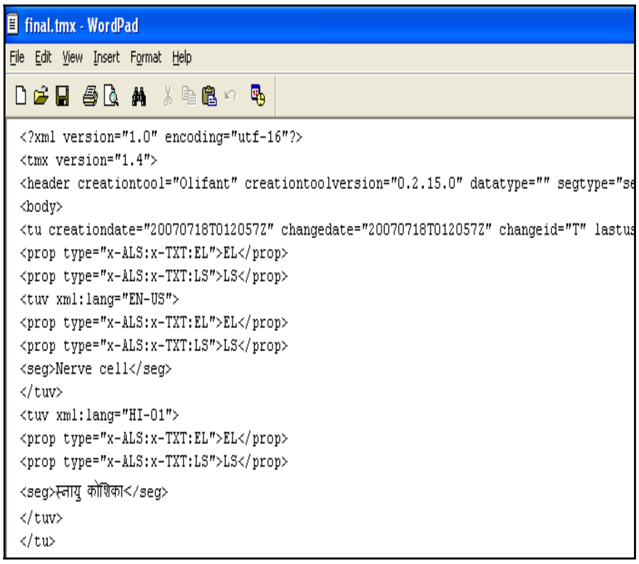
Once the TM database is created then it can be used for searching translations of sentences if they are present in it. Once a sentence is given to the MT system for translation, this module matches the input given through the AnglaMT interface with the contents of the TM database. The sentence or segments of the source language (English) that closely match the English sentences in the TM are found. The closest match (>80%) is found using word edit distance based algorithm [3]. The translation unit of the segment/sentence in the TM that matches with the input sentence is extracted from the TM database and given as machine output. If translation is not perfect, post editing is done. The process has been shown with two examples Fig.4 and Fig. 5. We indicate in Fig. 4 how the translation for the sentence “What is heart attack?” takes place. The exact match for the whole sentence is not found in the TM but a match for the segment “heart attack” is found from the TM and the translation unit is extracted. The rest of the translation is taken care of by the AnglaMT system. Similar case is shown for another sentence in Fig. 5. However the translation is not as perfect as for Example 1 shown in Fig. 4 and hence once the translation is received it is post-edited and then perfect translation for the input sentence is obtained. The post-edited segment which was missing from the TM database is added to the TM for updating. In the next cycle of translation a better translation will be extracted from the TM for a similar sentence. This would reduce the human effort needed for post-editing in the later cycles of translation.



```
Wordfast TM v.5.91c     %HI-01  %---79270261

20101012~103953 n       0       EN-US   AGRICULTURE     HI-01
कृषि
20101012~104046 n       0       EN-US   Vermicomposting brings
better yields and returns - A small farmer shows the way
HI-01    वर्मीकंपोस्टिंग अच्छी पैदावार और मुनाफ़ा लाता है – एक छोटा किसान रास्ता दिखाता है।
20101012~104405 n       0       EN-US   This is the case of a
small farmer who chose to be different from the typical
resource-  poor farmers struggling to make a living in the
degraded drylands.      HI-01    यह एक छोटे किसान की घटना है जिसने विशिष्ट संसाधन से भिन्न
होने के लिए चुना– ग़रीब किसान जीविका चलाने के लिये निम्नीकृत शुष्क भूमि में संघर्ष करते रहते हैं।
20101012~104446 n       0       EN-US   His deep ‘love for life’
has not only earned him a decent living, but more importantly,
it has inspired many other farmers to emulate him.      HI-01
जीवन से अपने गहरे प्रेम के लिए ही उसने बेहतर रहन–सहन नही हासिल किया बल्कि महत्वपूर्ण यह है कि उसने बहुत से दूसरे किसानों
को उसका अनुकरण करने के लिए प्रेरित किया।
20101012~104528 n       0       EN-US   Chandranna, a young
```

Fig. 2: Sample TM for Hindi

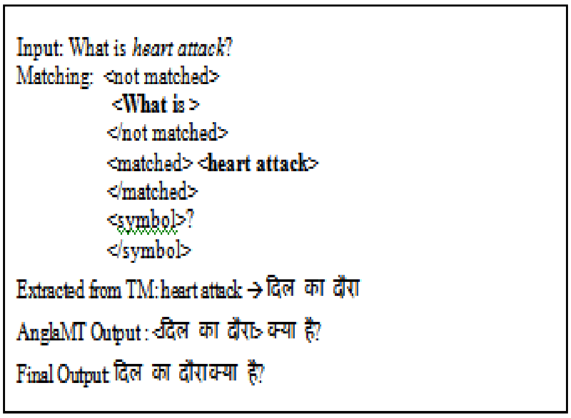


```
<?xml version="1.0" encoding="utf-16"?>
<tmx version="1.4">
<header creationtool="Olifant" creationtoolversion="0.2.15.0" datatype="" segtype="se
<body>
<tu creationdate="20070718T012057Z" changedate="20070718T012057Z" changeid="T" lastus
<prop type="x-ALS:x-TXT:EL">EL</prop>
<prop type="x-ALS:x-TXT:LS">LS</prop>
<tuv xml:lang="EN-US">
<prop type="x-ALS:x-TXT:EL">EL</prop>
<prop type="x-ALS:x-TXT:LS">LS</prop>
<seg>Nerve cell</seg>
</tuv>
<tuv xml:lang="HI-01">
<prop type="x-ALS:x-TXT:EL">EL</prop>
<prop type="x-ALS:x-TXT:LS">LS</prop>
<seg>स्नायु कोशिका</seg>
</tuv>
</tu>
```

Fig. 3: TMX format

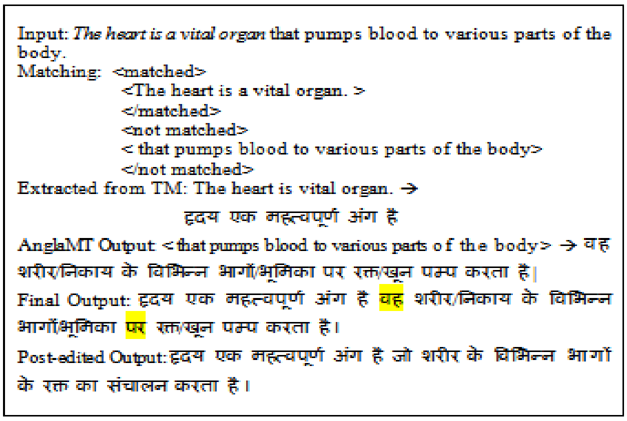
2.1.3. TM Update Process

After each input sentence is translated the TM is updated if the sentence is not present in the TM or if there is a need to update it. The translators make multiple passes over the text thereby incrementally improving the translation accuracy due to dynamic TM updates. Such update mechanism is maintained by incrementally updating the TM database by performing sentence alignment (or partition) between the source text and the translation.



Input: What is *heart attack?*
Matching: <not matched>
<What is >
</not matched>
<matched> <heart attack>
</matched>
<symbol>?
</symbol>
Extracted from TM:heart attack → दिल का दौरा
AnglaMT Output : <दिल का दौरा> क्या है?
Final Output दिल का दौराक्या है?

Fig. 4: Example 1



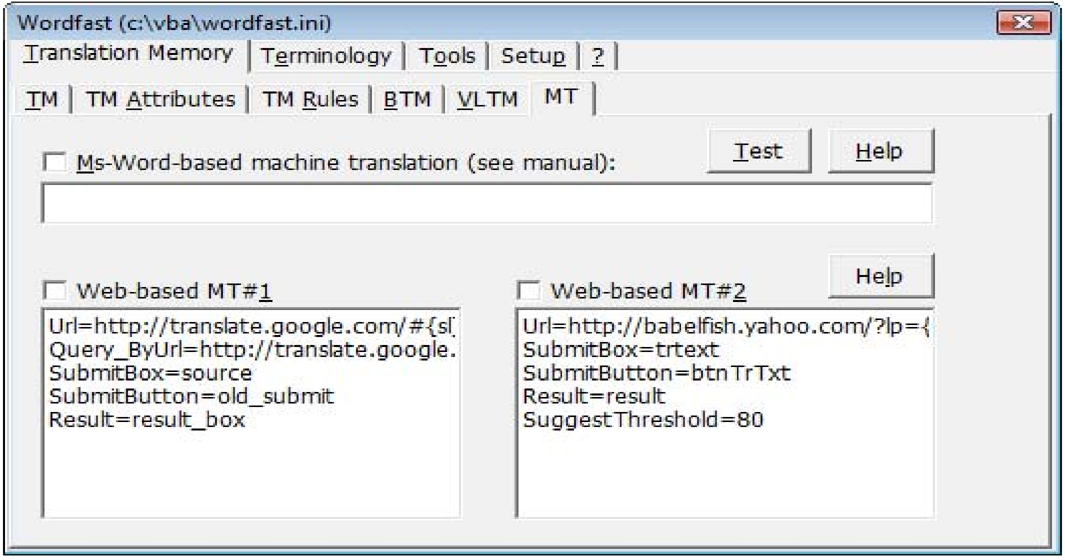
Input: *The heart is a vital organ* that pumps blood to various parts of the body.
Matching: <matched>
<The heart is a vital organ. >
</matched>
<not matched>
< that pumps blood to various parts of the body>
</not matched>
Extracted from TM: The heart is vital organ. →
हृदय एक महत्वपूर्ण अंग है
AnglaMT Output <that pumps blood to various parts of the body> → वह शरीर/निकाय के विभिन्न भागों/भूमिका पर रक्त/खून पम्प करता है|
Final Output: हृदय एक महत्वपूर्ण अंग है वह शरीर/निकाय के विभिन्न भागों/भूमिका पर रक्त/खून पम्प करता है।
Post-edited Output:हृदय एक महत्वपूर्ण अंग है जो शरीर के विभिन्न भागों के रक्त का संचालन करता है।

Fig. 5: Example 2

2.2. AnglaMT as a Resource for Wordfast

Wordfast is a Computer-Aided Translation (CAT) program designed as a Microsoft Word add-on [4]. The primary purpose of Wordfast is to help translate Ms-Word documents; it can also be used with PowerPoint and Excel documents, and tagged files. It is a format often used in the localization industry. Wordfast uses segmentation and Translation Memory (TM) to produce translations. Wordfast also offers advanced terminology functions, three simultaneous glossaries, concordance search in unlimited numbers of TMs, reference search in unformatted documents, links to external, third-party dictionaries or web-based terminology databases etc.

Currently Wordfast is using two popular MT systems which are Google and Babelfish. In case a translation segment is not found in the translation memory the sentence is automatically redirected to the selected MT system. The translation given by the MT system is produced by Wordfast as translation. This translation can be further edited if needed. Fig.6 shows the existing approach being used by Wordfast. On the similar pattern we propose to integrate our existing AnglaMT system with Wordfast. The integration can be done in two ways, using local source and the remote source. This paper thus proposes the idea to integrate AnglaMT either as a local or a remote resource. The work for implementation of this approach is in progress.



□ Ms-Word-based machine translation (see manual):

□ Web-based MT#1

```
Url=http://translate.google.com/#{sl
Query_ByUrl=http://translate.google.
SubmitBox=source
SubmitButton=old_submit
Result=result_box
```

□ Web-based MT#2

```
Url=http://babelfish.yahoo.com/?lp={
SubmitBox=trtext
SubmitButton=btnTrTxt
Result=result
SuggestThreshold=80
```

Fig. 6: Local/Remote source in Wordfast

3. Future Work and Conclusion

The approaches described in the paper show initial investigations for combining translation memory in RBMT framework as an enhanced support like an example base. The integration of TM with AnglaMT has improved the translation quality and the probability of getting accurate translations has increased. This inference is based on evaluation of 100 sentences. As the TM database grows the results become more acceptable and accurate. However based on the observations, we feel more sophisticated matching algorithm needs to be built. In this paper we have explained how translation memory has been integrated with the MT system that serves as a tool for a translator to reuse and maintain his daily translation work. This improves efficiency of the system and reduces human effort for post-editing. Currently we are working towards implementing the AnglaMT system as a local as well as a web resource in place of the currently used MT systems namely Google and Babelfish.

We are also working towards analysis of how the translation software influences translators' cognitive process. In view of recent advances within MT combined with an increasingly competitive market, professional translation will soon be carried out as HAMT [5, 6]. Such automation of translation software will include human translators' work and the impact on their cognitive processes will increase by a considerable amount. Therefore, in future to help translators to digitize their work, we shall need more empirica studies of how they interact with TM and other translation technologies.

4. Acknowledgements

We express our profound sense of gratitude and regard to Prof.R.M.K Sinha and Mr. V.N.Shukla for guiding us in the right direction. We would like to thank Dr. George Varkey for his support and cooperation. We are grateful to our other colleagues for their support and help. This work is being taken up as an exercise for further development and improvement of AnglaMT (Phase II) project and is being funded by the Department of Information and Technology, New Delhi, India. We also thank the Indian government for funding this project.

5. References

[1] R.M.K.Sinha, K.Sivaraman, A.Agrawal, R.Jain, R.Srivastava, A.Jain, "ANGLABHARTI: A Multilingual Machine Aided Translation Project on Translation from English to Indian Languages", *IEEE Conf. on Systems, Man and Cybernetics*, Vancouver, Canada, Oct 22-25, 1995, pp. 1609-1614.

[2] D. Marcu 2001. Towards a unified approach to memory and statistical based machine translation. *Proc. of the 39th Annual Meeting on Association for computational Linguistics* (ACL 2001), Toulouse, France, pp.386-393.

[3] Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proc. of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 223–231.

[4] Wordfast Technical Reference manual - Version 5.90x.

[5] Jin-Xia Huang, Wei Wang and Ming Zhou. 2002. A Unified Statistical model for Generalized Translation Memory System. *Proc. of the 19th International Conference on Computational Linguistics*.

[6] Fiederer, R., O‟Brien, S.: Quality and Machine Translation: A realistic objective? *The Journal of Specialized Translation 11*, 52-74 (2009)