Workbench for Post-editing of Translations from English and Hindi to Dravidian Languages

Sobha Lalitha Devi, Pattabhi R K Rao and Vijay Sundar Ram

AU-KBC Research Centre, MIT Campus of Anna University, Chennai-600044

sobha@au-kbc.org, pattabhi@au-kbc.org, sundar@au-kbc.org

Abstract. In this paper we discuss about the translator’s workbench which provides the translator with various types of post-editing tools to facilitate the speeding up of the translation process. Post-editing is the process of correcting/changing the machine generated translation to a syntactically and semantically correct translation. This work bench is specifically customized to handle the correction of translation from English and Hindi to Tamil and Malayalam from three machine translation systems, the Google unpaid version, Google paid version and Sampark System (Indian language to Indian Language MT system, Govt. of India). The tools in the workbench include the domain dictionaries, technical term translation correction, Grammatical correction modules where verbs are checked and corrected, Appropriate word selector with morph generator and interactive translation prediction where the users edits are automatically stored and predict if there is a similar error is encountered. We conducted a field trial of our post editor and it is found that it reduced the time taken to produce the final translation. This paper reports work in progress.

Keywords: TRANSLATORS WORKBENCH, POST-EDITING, DRAVIDIAN LANGUAGES

1 Introduction

With improvement of machine-translation (MT) technology and the demand of translation industry, post-editing has become an important production process in translation. Post-editors check and correct MT output to improve the translation quality and applicability. Translation post editing and revising are the two concepts in editing the translation output from any machine translation (MT) output and translation revision is considered as a “function of professional translators in which they find features of the draft translation that fall short of what is acceptable, as determined by some concept of quality, and make any needed corrections and improvements” [1], and it has an important role in ensuring the quality of the translation. According to ISO 17100, post-editing the other concept is “editing and correcting machine-translation output”.

315

Proceedings of the International Conference HiT-IT 2023, pages 315–324,
7-9 July 2023, Naples, Italy

https://doi.org/10.26615/issn.2683-0078.2023_027
A Translator's workbench provides the user with a set of computer-based tools which help the post-editors for speeding up the translation process. Post-editing is the process where humans identify the errors in machine-generated translation to achieve an acceptable final product. Various new interactive tools are developed with new features aiming to assist the post-editors which are included in workbenches. It is found that the translators are able to use the workbench without much training given [2,3].

With successful use of Moses, Statistical machine translation, Computer Aided Translation (CAT) tools integrated with Machine translation system and post-editing the output got prominence in translation industry. [4] presented PET: a Tool for Post-editing and Assessing Machine Translation, which mainly focussed on facilitating the post-editing of translations from any MT to reach publishable quality and also to study the sentence-level information from the post-editing process, e.g.: post-editing time and detailed keystroke statistics. [5] presented the details of post-editing work using AutoDesk tool in a translation industry, where they handled translation in 13 languages. [6] presented an open source CAT tool, MateCat tool, which was integrated with translation memories, terminology bases, concordancers, and machine translation system. [7] presented INMT: Interactive neural machine translation prediction, to assist human translators with on-the-fly hints and suggestions to achieve the end-to-end translation process faster, more efficient, and creates high-quality translations.

Multi-model post editing tool (MMPE) combining traditional input modes with pen, touch, and speech modalities for PE of MT was presented by [8]. The results of this evaluation with professional translators showed that pen and touch interaction were suitable for deletion and reordering tasks, while these instruments were not used for longer correction. IntelliCAT, an interactive translation interface with neural models to streamline the post-editing process was presented by[9]. They have shown a 52.9% speedup in translation time compared to translating from scratch.

UDAAN, is an open-source post-editing tool built with experiments from English to Hindi translation task. It has machine translation system along with post-editing tools. It provides 100 in-domain dictionaries for aiding the post-editing task [10].

Though there are post-editing tools available in European and English, there are very less attempts for Indian languages. And oddly any tool specifically catering to Dravidian languages such as Malayalam, Tamil, Telugu, Kannada, which are morphologically rich and requires special processing with post-editing the translation.

In this paper we describe our on-going work of developing a translator’s workbench for MT output. A need of a translator’s workbench was felt while we were working on translating undergraduate and post graduate lectures from University Grant Commission of India, the courses taught for Indian college students, (https://pmevidya.education.gov.in/swayam-portal.html) under “Swayam Platform”. The courses were in 7 domains, Law, Environment, Mathematics, Chemistry, Biology, Computer Science and Public Administration. Since we had to deal with different domains and no MT system is customized for separate domains, the output we were getting from generic MT systems needed post editing at various levels. Purely manual post editing took enormous amount of time for the final product. There were 86 courses, a total 1300 hrs of lecture with a total of 3,50,000 sentences in English. The
paper is designed as follows, in the next section we discuss about the common errors from the three MT systems for English and Hindi to Dravidian Languages a) The Sampark system, b) Google free version and c) the Google paid version. In the third section we have given in details the architecture and the working of our workbench and finally the conclusion.

2 The Three Machine Translation Systems

The availability of automatic translation systems in most of the languages brought by the tech giants such as Google, Microsoft, Systran, Facebook etc has brought a paradigm shift in the translation industry. Though the translations by these automatic translation system needs correction, translators use these translation output as the base and perform post-editing of these translation to get proper translation. In this section, we give in detail the translation systems used in our workbench and discuss the errors in the MT translation output.

Workbench is designed to handle the translation output in the following language pairs, English-Hindi, English-Malayalam, English-Tamil, Hindi-Tamil, Hindi-Malayalam and Tamil-Malayalam. We have integrated three different translation systems with this workbench, namely, Sampark translation system, Google paid and free translation versions.

Sampark Translation System: Sampark translation system is built using Analysis-Transfer-Generate architecture using hybrid techniques. Modules were built using Machine Learning techniques with linguistic features and post-processed with rich set of linguistic rules. The Analysis part has the modules to perform a detailed analysis of the source sentence such as morphological analyzer, POS tagger, Chunker, Clause boundary identifier and Named Entity Recognizer. This is followed by the Transfer part, where the lexical, structural and syntactic transfers are done as required by target language using transfer grammar and lexical transfer engine. In the Generation part, target language sentence is generated using the lexical, structural and syntactic transfers performed in the transfer part. These systems were built under a funded project by Government of India. We use Hindi-Tamil, Tamil-Hindi, and Tamil-Malayalam translation systems built under this project for our workbench. They are available on http://www.tdil-dc.in/index.php?option=com_vertical&parentid=74&lang=en. We have included Sampark System to our workbench.

Google Translation: We have included both the paid and free translation services provided by Google, which are built using neural machine translation techniques, into our workbench. The free version has character limitation and limitations on number of usage. It fails very often. To overcome these issues, we have included the paid version of Google translation API to our workbench.
2.1 Errors from The Three Machine Translation:

We broadly classify the errors in the translation as simple and complex errors. Simple errors include spell errors, case transfer errors, copula errors and errors in selection of Domain /Technical terms, correct lexical item as per the context of the sentence, Translation of Named Entities. Complex errors include the sentence construction error and error in generation of complex verb phrases. We explain in detail each of the errors with example.

**Case Transfer Error:** As Indian languages are inflection and morphologically rich, the case marking with the nominals play an important role in semantic interpretation. There is no one to one mapping of case markers between languages as in locative case in Malayalam need not be locative case in Tamil. There are also one to many and many to one case mapping between Indian languages. ‘se’ “instrumental” in Hindi can be transferred into ‘aal’ instrumental, ‘ai’ accusative, ‘ilirunthu’ ablative case in Tamil. The context will determine which case should be taken while translating from Hindi to Tamil. The locative case in Malayalam changes to dative in Tamil. Consider the example. ML: avan chennaiyil (il is the locative case) poyi; TA: avan chennai (kku is the dative case) poonanan. (He went to Chennai.). Both NMT systems and Sampark systems introduce this errors in case transfer. Sampark system has specific linguistic module to handle case transfer in transfer grammar engine, so it performs better than Google translation engines.

**Copula Generation Error:** Hindi and Malayalam have copular construction and it is a necessary condition for the sentence to be syntactically correct, whereas it can be dropped in Tamil. In the translation from Tamil to Hindi and Malayalam copula generation error occurs. In Sampark translation engine, we have a copula generation engine and it is handled better in Sampark system output and these errors are more in Google translation. Copula errors often occur in Tamil to Malayalam and Tamil to Hindi translation output.

**Lexical Selection:** Correct selection of context specific lexical item is a challenge and this is not handled by any of the three translation engines. It is comparatively less in Google paid version, as it is trained with large data. The lexical selections are of three types: 1. Domain/Technical Term Selection, 2. Context depended term selection and 3. Named Entity selection. The Domain terms and context terms are not properly selected by all three systems but Named entity is handled by Google paid version and Sampark systems.

Another issue in term selection is regarding equations in Mathematics and Chemistry. They are not handled by all three systems.

**Errors in Complex Construction:** Translation of Multi-clause (sentences with multiple embedding) and long sentences using Google brought in error in the target language syntactic construction and require editing at a higher level. Sampark system on the other hand, handles these sentences better than Google as it analyses the source language at the syntactic level.
Verb Transfer: Sampark engine having word generation using linguistic information generates more precise verb forms than Google translation. In Google translation we get incorrect verb forms instead of the exact verb form. Both Malayalam and Tamil, being morphologically rich languages, have complex verb generation.

Unknown Word Handling: Unknown words are the potential challenge in Translation output. Sampark systems fail to handle the unknown words. It will just transliterate those words. It will not even generate the word with proper inflections. This affects the translation of the sentence. In the case of Google translation unknown words are handled using different techniques and presents it as either transliteration word or generate into another word using sub-word techniques. In both the cases, validating these words is required.

Degree of Translation Error: Sampark and Google systems produce perfect translation at the syntactic level for simple generic sentences whereas they bring in errors if the sentences are complex. The lexical errors are common in both the constructions. Google paid version is better than the free version, but both the engines fail in translation of complex sentences with multiple embedding.

3 Workbench and Its Architecture

The working of our translator’s workbench is as follows. The user can upload the input data for translation in any of the format (doc, docx, pdf, txt) and can choose the target language. After choosing the target language the user can choose the translation engine among the three translations systems (Google translation (paid and free service), and Sampark translation) for English- Hindi, English-Tamil, English-Malayalam, Hindi-Tamil, Hindi-Malayalam and Tamil-Malayalam translations. The architecture of the workbench is presented in figure 1(Fig is attached at the end of the paper after reference). The output in the target language from the translation systems and the input source language are shown to the Post editor. The editor can see the target sentence and mark it as correct or not correct by clicking yes or no. If it is yes it will change the colour to green and can move to the next sentence. If it is not correct then it will show all the tools available on either side of the tool box in a hierarchical order, while correcting at each hierarchy, the next level will be shown. The workbench itself will show certain level of errors through its intelligent Error identifier which is equipped with automatic NE identifier, Domain Term Identifier and Technical, mathematical names.

3.1 Translation Error Handling

The tool identifies errors in translation using a NER, Domain Identifier and Terminology Identifier. It also identifies partially translated sentences and certain improper translation and alerts the user. The partial translation is identified by comparing the number of characters and words in the input sentence and the translated output sentence. When there is drastic difference, then an alert message is given to translator
while he corrects that sentence. Consider the following example of English to Malayalam translation.

**Input:** For instance, if individuals with capital 'T' are more successful in reproduction than the individuals with small 't', the frequency of the former will be higher.

**Translation Output:**

Mal: ഉദാഹരണമായി, പ്രത്യേകതകളുടെയും പരിസ്ഥിതികളുടെയും ചില വിശദീകരണം സാധാരണയായി അദ്ദേഹം നല്ല സഫലതയിൽ യാത്ര ചെയ്തിരുന്നു. 

(For example, capital(N)+ulla(psp) individual(N)+plural+copula+conj)

Considering the above example, the input sentence has 29 words and 179 characters and translated sentence has 4 words and 49 characters. While comparing English and Malayalam we expect the number of words in Malayalam to be more than 40% of words in English. If the translation has less than 40% of words then it is flagged as partial translation. This value 40% is obtained by the empirical analysis of the English-Malayalam parallel data.

**Improper Translation:** Improper translation is identified using pattern analysis. Consider the following example (English to Malayalam translation).

**Input:** The second website is www.theagricos.com.

**Translation Output:** രണ്ടാമത്തെ വെബ്സൈറ്റ് www.ആഗ്രികോസ്.com.

The input sentence has an url and in the translation it should remain the same. But the part of the url is translated. This will be identified as improper translation and user will be alerted.

3.2 Domain Based Translation Correction

Workbench has automatic domain classifier and it will classify the domain of the input and choose the appropriate dictionaries for the target language. The tool will highlight the Technical and Domain Terms in source language and the editor can use that to correct the output. We have domain and technical term dictionary, mathematical dictionary, Chemical name dictionary and botanical name dictionary. Domain and technical term glossary for look up is included in the workbench and also has domain and technical term dictionaries used for identifying the terms and highlighting it.

Since Malayalam and Tamil are morphological rich languages, while replacing the lexical items, suffix attached to the lexical items has to be handled. Consider the following example, which has English to Tamil translation.

**Input:** In addition to the nuclear genome plants also contain genes in the organisms like chloroplast or plastids and mitochondria.

**Translation Output:**

**Corrected Translation:**
In the above example, the lexical item ‘plants’, is translated as ‘நவீன இயற்கையியல்’ (for industries), which means ‘for the industries’, but the sentence is from biological domain and the lexical substitution has to be ‘நவீன’ (for plant). In this example ‘நவீன’ has plural marker ‘கல்’ (kaL) and dative case marker ‘க்கு’ (kku) suffixed to the root noun ‘நவீன’. While replacing ‘நவீன’ with ‘இயற்கையியல்’, the lexical item has to be generated for these two suffixes as ‘இயற்கையியல் கல்’. Here we accomplish by the identification of the suffixes and generating the new word with the required morphological suffixes by using a light weight morphological analyser and morphological generator built using finite state automata and paradigm dictionaries. In the present system we have 15 different domain dictionaries.

3.3 Spelling Correction and Alterations

After the translation corrections are done, we ensure the spelling correction in the translation. We have in-house built spell checkers for Malayalam and Tamil using n-gram and finite state automata techniques. If the spell checker identifies the error word, it will be highlighted and the possible suggestion words will be provided to the user as a suggestion list. Here we also take care of the named entities, as the spelling of the NEs has to be consistent throughout the document. Using in-house developed Named Entity Recognizer (NER), we identify the NEs in the source sentences and spelling for those NEs in the translated sentences are maintained consistently.

For Tamil text, we have sandhi-validation module. In Tamil we have to maintain the lengthening of sound between the words by marking the end of the previous words with vowel. Consider the following example, உள்ளிட்டு பள்ளி (school education) Between the two words ‘இ’ a vowel is inserted as the sandhi (morphotactic rule). These sandhi validations are done using a set of linguistic rules.

Evaluation and Conclusion

We have evaluated our workbench using the data from Swayam courses. The total number of English sentences we have used through the workbench is 3,50,000 sentences, Number of simple and complex sentences are 1,15,500 and 2,34,500 respectively. Other statistics are given in the table 1.
### Table 1. Sentence Wrongly Translated by Systems

<table>
<thead>
<tr>
<th>S.No</th>
<th>Details</th>
<th>Google paid (English to Tamil)</th>
<th>Count in Sampark (Tamil to Malaylam)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Completely Wrong Translation in Mathematical Domain</td>
<td>97500</td>
<td>87500</td>
</tr>
<tr>
<td>2</td>
<td>Completely Wrong Translation in other Domain</td>
<td>15000</td>
<td>11450</td>
</tr>
<tr>
<td>3</td>
<td>Number of Sentences where NE/Domain Term has to be corrected</td>
<td>8000</td>
<td>9500</td>
</tr>
<tr>
<td>4</td>
<td>Number of Sentences where Mathematical Equations has to be corrected</td>
<td>114000</td>
<td>10900</td>
</tr>
</tbody>
</table>

We have employed 20 Tamil and 20 Malayalam editors to correct the machine output. The time taken for correction by editors is as given in Table 2

### Table 2. Time Taken by Editors

<table>
<thead>
<tr>
<th>S.No</th>
<th>Sentence Type</th>
<th>Time Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simple Sentences</td>
<td>30 sec to 1.5 mins</td>
</tr>
<tr>
<td>2</td>
<td>Complex Sentences</td>
<td>3 to 5 mins</td>
</tr>
<tr>
<td>3</td>
<td>Sentences with complex Mathematical Equations</td>
<td>5 to 15 mins</td>
</tr>
</tbody>
</table>

We have discussed about our translators workbench specifically customized to handle the correction of translation from English and Hindi to Malayalam and Tamil using three translation systems, namely, Sampark system, Google paid version and Google free version translation service. We have described the errors in the translation output of these systems. Domain dictionaries, technical term translation correction and grammatical correction modules are included in the workbench. These features help in speeding up the post-editing task.

### References

Fig. 1. Architecture of the Workbench