Engineering and Scientific International Journal (ESIJ)
Volume 3, Issue 1, January – March 2016
ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

A Comparative Study of Occurrence of Errors in Machine Translation in a Multilingual Environment

S.Sivakama Sundari[#1], V.Prema*[2], G.Savitha*[3]

[1] *Head, Department of Computer Science, Alpha Arts and Science College, Porur, Chennai - 116.*
siva.aasc@gmail.com
[2]*Assistant Professor, Department of Computer Applications, Alpha Arts and Science College, Porur, Chennai - 116.*
vprema78@gmail.com
[3]*Assistant Professor, Department of Computer Science, Alpha Arts and Science College, Porur, Chennai - 116.*
gsavithamca@gmail.com

Abstract— Machine translation is an inevitable field of Natural Language Processing, which includes two steps. The first step literally follows the reference method of machine translation, taking advantage of corpus of knowledge already available in the system. The second step involves the post-editing done by human intervention. This paper analyzes the errors, categorizes them, and gives a comparative study of the errors in the first and second steps. Further, this analysis is done in a multilingual environment. Every step is accompanied by the machine training on the corpus of annotations and systematic classifiers.

Keywords— Machine Translation, Machine Learning, Classifiers, Post-editing, multilingual environment.

1. Introduction

Machine Translation (MT) requires a repository of words with the corresponding semantics, syntax and classifiers. The Machine Training is performed by augmenting the most reasonable and relevant forms of synonyms for each word. Data mining and Data Warehousing are the most fundamental disciplines which are of great value at this juncture. The entire world has shrunk into a global village and there is a simultaneous knowledge explosion and technology facilitated state of art communication gadgets. Now what is to be supplemented is the readiness of the data to be translated into different languages around the world. Hence the incredible machine translation plays a vital role in the current scenario to provide a common platform for knowledge sharing worldwide.

2. Repository of Words

The repository of words are nothing but the corpus maintained by the Machine Translation system. The words are annotated with the origin, pronunciation, semantics, syntax, lexical analysis and also the classifiers according to the rule-based inference engines, which may be subject wise or parts of speech or psychological tones or reference to the corpus of knowledge or links to Wikipedia or research papers or conference papers. The depth of information is directly proportional to the effectiveness of machine learning and the enormity of the corpus of words. Both these aspects have cumulative effect by subsequent machine training and accumulated annotations.

3. Machine Learning

The back propagation algorithm is used very effectively to train the machine with the unknown data. The more irrelevance in the text, there is more possibility of the expansion of knowledge with the new lemma added to the repository. The success or failure of the machine translation largely depends on the corpus and the failure occurs resulting in the unacceptable and unreliable outputs.

When analysing the reasons of this failure, Burchardt et al. (2013) note that: "Error analysis is considerably more time consuming than anticipated. Rather than analysing a few thousands of sentences in our pilot phase, we were able to have a few hundred analyzed. While spees would improve with training and experience, detailed analysis is a labour-intensive task and large-scale annotation would require either many annotators (raising problems of inter-annotator consistency) or much time." [1]

4. Methodical Study

The original text is selected from the vocabulary list of a book of English. The list of words is fed into machine translation software and the output has been received as a result of active translation with the machine translation supported by the reference corpus. The results are analyzed by human subject expert to categorize the instances into the following categories. This process is essentially referred to as POST-EDITING [4]. This post-editing process cleanses the translated text by making it devoid of errors whatsoever.

The resultant output may reflect the following conditions.

- No Error – Text matter is acceptable
- Error Categories[3]
  - Semantic Error – Conveys wrong Meaning

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

- Syntax Error – Grammatical mistake
- Lexical Error – Improper lexical division
- Morphological Error – Structure of sentence wrong
- Formatting Error – Disobeys the rules of formatting
- Unclassified Error – Either belongs to more than one category mentioned above or ambiguous error

5. Reasoning on the Occurrence of Errors

The different categories of errors are categorized and their occurrence can be reasoned as follows. In a multilingual environment there is difference and dissimilarity of the languages in question in

a. Syntax
b. Semantics
c. Morphology
d. Format
e. Lexical Analysis
f. Contextual meaning
g. Practical usage of words
h. Culture and geographical disposition of the population of the native speakers [7]
i. Economical status of the native speakers of the language in question.
j. Code-switching and cross border linguistic abilities.[2]

All the above mentioned attributes can be looked upon as barriers for the generation of the acceptable and error free output. The final refinement process requires the human intervener to possess an exhaustive knowledge of all these attributes, in addition to the subject knowledge of the topic which is being translated. In case of lack of such inference engines created by human experts being implemented in the process, the active translation output, could be predominantly a mishap and misleading garbage of textual content.

6. Error Classifier- Statistics of the Study

Table1. Error Classifier

Category	**Number**
No Error	72
Syntax Errors	11
Semantic Errors	12
Morphological Errors	34
Lexical Errors	14
Format Errors	8
Uncategorized Errors	10
Total Number of Test Instances	**161**

7. Assumptions and Inferences

Morphological errors are more prevalent than other category of errors in the given text. It might have been due to the drastic disparity between the structure of the two languages in question namely English and Tamil. The next step of refinement almost results in 'NIL ERROR' status, notwithstanding the fact that the human expert is forced to refer other modes of language repository and choose the exact words and sentence pattern, to satisfy the requirements of the language concerned.

8. Multilingualism

The earlier research arose the interest of the investigator, and hence an extension this approach is done with the involvement of three languages namely, English, Tamil and Hindi, leading to a typical multilingual environment. The original text content was selected in English and with the help of machine translation software; Tamil version was obtained as the output.

This is considered to be the active translation and the result of machine translation (MT). This output was verified in the light of classifiers as done above [8]. The effort was aiming to produce a 'NIL ERROR' status. The errors were cleansed up on one hand, and on the hand we are left with the statistics on the category and number of instances of a

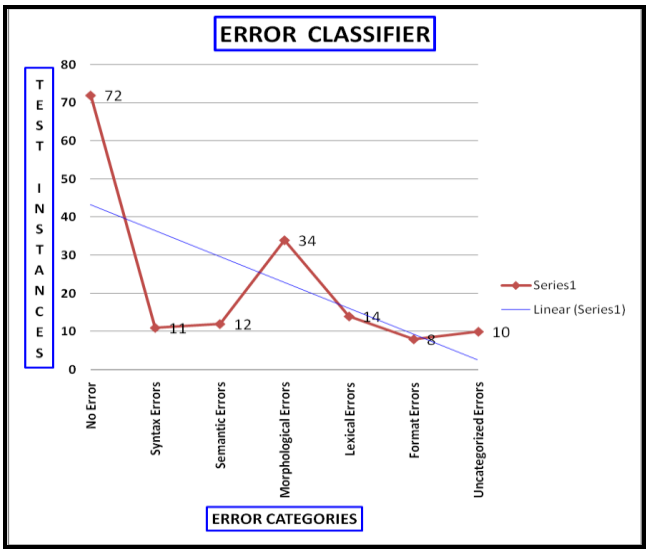

Fig. 1: Comparative statistics of Errors

specific error category. The third dimension to this operation is added, when the same process is done with another language, Hindi. English version is translated with the aid of machine translation software and the resultant output was analysed on the same lines of research. Now we have received a comparative study of two different conversion processes, involving essentially three languages in the scenario [6].

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

9. Error Classifier Involving Multilingualism Statistics of the Study

Table 2. English to Tamil Translation

English to Tamil Translation	
Category	**Number**
No Error	325
Syntax Errors	35
Semantic Errors	32
Morphological Errors	45
Lexical Errors	24
Format Errors	21
Uncategorized Errors	26
Total Number of Test Instances	**508**

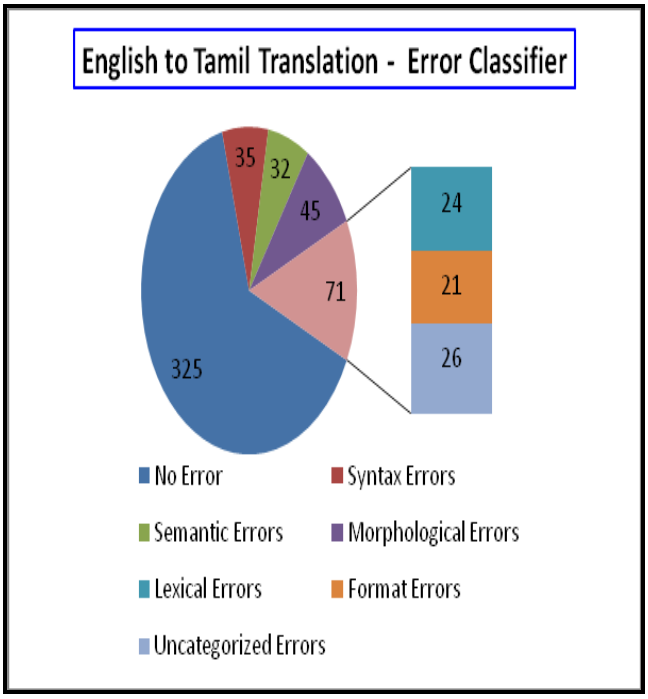

Fig. 2: English to Tamil Translation – Error Classifier

10. Assumptions and Inferences

The number and category of errors do not present any major disparities, which follows that these two languages (Tamil and Hindi) are inherently related to each other, and they keep up the same distance towards English, though it is widely accepted to be the crucial language of international communication [9].

Table 3. English to Hindi Translation

English to Hindi Translation	
Category	**Number**
No Error	403
Syntax Errors	20
Semantic Errors	17
Morphological Errors	30
Lexical Errors	15
Format Errors	12
Uncategorized Errors	11
Total Number of Test Instances	**508**

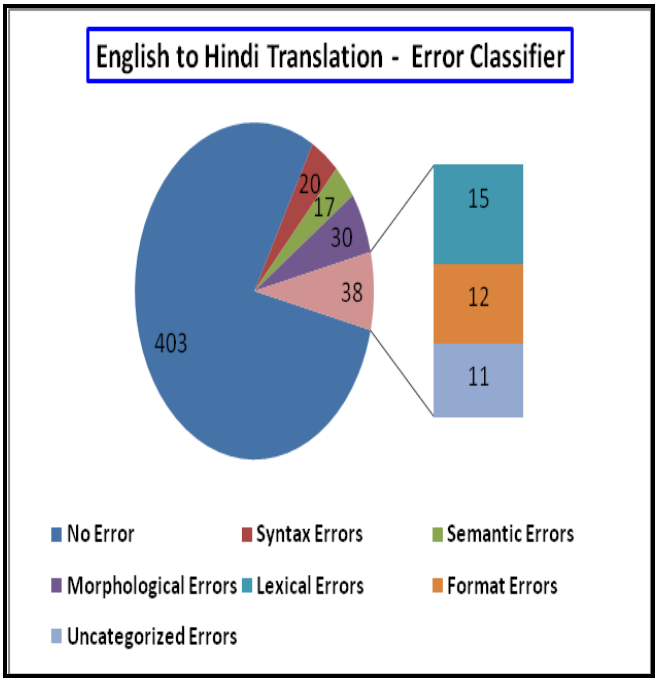

Fig. 3: English to Hindi Translation – Error Classifier

11. Conclusions

The research paper has mainly aimed at two aspects. The quality of the machine translation, lines of effort towards the improvement of the error reduction capability of the active translation by adapting back propagation and enriching the corpus of vocabulary in order to achieve greater efficiency in future. The multilingual approach adds more insight into the comparative study of the machine translation software and the nativity, proximity and natural affinity of different languages worldwide [10].

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

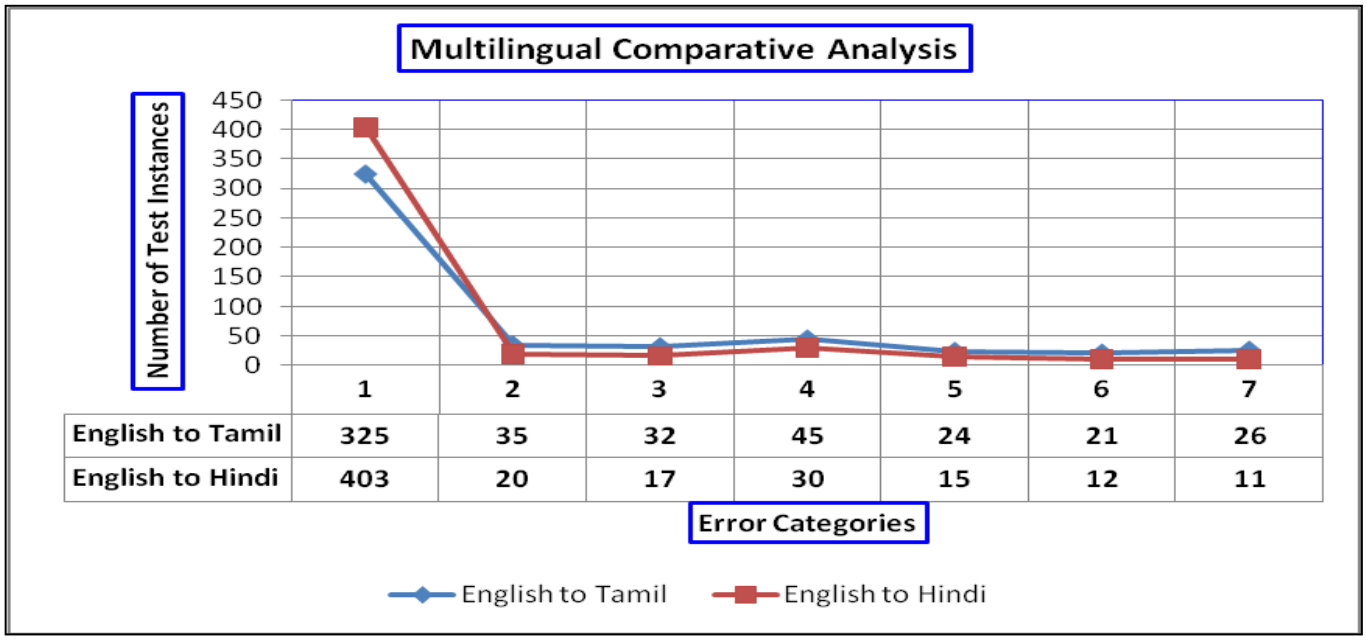

Fig 4. Multilingual Comparative Analysis

References

[1] GuillaumeWisniewski_, Natalie K¨ublery, Franc¸ois Yvon_LIMSI-CNRS, A Corpus of Machine Translation Errors Extracted from Translation Students, 91 403 Orsay, France, 75 013 Paris, France. http://www.lrec- conf.org/ proceedings/lrec2014/papers.html

[2] Anik Dey, Pascale Fung, A Hindi-English Code-Switching Corpus, Human Language Technology Center, Department of Electronic & Computer Engineering, HKUST http://www.lrec-conf.org/proceedings/lrec2014/papers.html

[3] Peter Exner, Pierre Nugues, REFRACTIVE: An Open Source Tool to Extract Knowledge from Syntactic and Semantic Relation, Lund University, Department of Computer science, Lund,Sweden.http://www.lrec-conf.org/ proceedings/lrec2014/papers.html

[4] Varvara Logacheva, Lucia Specia, A Quality-based Active Sample Selection Strategy for Statistical Machine Translation, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom, http://www.lrec-conf.org/proceedings/lrec2014/papers.html

[5] Axel-Cyrille Ngonga Ngomo, Norman Heino, Ren´e Speck, Prodromos Malakasiotis, A Tool Suite for Creating Question Answering Benchmark, Department of Computer Science, Department of InformaticsUniversity of Leipzig, Athens University of Economics and BusinessAugustusplatz 10, 04109, http:// www. lrec - conf. org /proceedings/ lrec2014/ papers.html

[6] Helen Hastie, Anja Belz Herio,t A Comparative Evaluation Methodology for N L G in Interactive Systems, Watt University, University of Brighton. http://www.lrec-conf.org/proceedings/lrec2014/papers.html

[7] A Multi-Cultural Repository of Automatically Discovered Linguistic and Conceptual Metaphors 1State University of New York – University at Albany, 2Sarah M. Taylor Consulting LLC, 3Polish Academy of Sciences E – mail : samirashaikh@gmail.com, tomek@albany.edu, http://www. lrec-conf.org/proceedings/lrec2014/papers.html

[8] Mohamed Morchid, Richard Dufour, Georges Linar`es, A LDA-Based Topic Classification Approach from Highly Imperfect Automatic Transcription, LIA - University of Avignon (France), fmohamed.morchid, richard.dufour, georges.linaresg@univ-avignon.fr, http://www.lrec-conf. org/ proceedings/lrec2014/papers.html

[9] Marion Baranes Beno ̂ıt Sagot A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon, http://www.lrec-conf.org/ proceedings/ lrec2014/papers.html
viavoo, 77 rue de Paris, 92100 Boulogne-Billancourt, France Alpage, INRIA & Universit´e Paris Diderot, b ̂atiment Olympe de Gouges, 75013 Paris, France, marion.baranes@viavoo.com, benoit.sagot@inria.fr, http:// www.lrec-conf.org/proceedings/lrec2014/papers.html

[10] Yves Scherrer, Benoît Sagot Alpage, A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages, http://www.lrec-conf.org/proceedings/lrec2014/ papers. html

[11] INRIA & Université Paris Diderot, Bâtiment Olympe de Gouges, 75013 Paris, France, . LATL-CUI, Université de Genève, 7 route de Drize, 1227 Carouge, Switzerland, yves.scherrer@unige.ch, benoit.sagot@inria.fr, http://www. lrec-conf.org/proceedings/lrec2014/papers.html