# Enhancing Bi-Lingual Example Based Machine Translation Approach

Manish Rana, Mohammad Atique

[1]Research Scholar, Post Graduate Department of Computer Science & Engineering, Sant Gadge Baba Amravati University, Amravati, India.
[2]Associate Professor, Post Graduate Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, India.

***Abstract —*** *This research paper shows the implementation of the work carried in machine translation using machine learning algorithm taking in consideration bi-lingual i.e. English to Hindi translation based on fuzzy technique. Model is implemented in Python taking input in English and translating to Hindi as output. It consists of a trainee dataset containing English equivalent Hindi sentences. Initial program run to train the application with the training set data. Minimum one million datasets are taken based on Microsoft's vast collection of datasets .After implementation of the program and comparison with other techniques used in such research, the result found to achieve efficiency of 80% and above. The research done on the following machine translation shows a significant achievement in the relevant area. Further it opens a new gateway for improving the research on machine translation.*

***Keywords—Fuzzy Logic, Machine Translation, Machine Learning, NLP, etc.***

***General Terms: Fuzzy logic, Tokenization, machine translation etc.***

## I. INTRODUCTION: DESIGN OF SYSTEM EBMT USING FUZZY LOGIC

In this research work, the following use of the corpus consisting of minimum one million datasets and complex sentences, to carry out the experiments of machine translation use machine learning algorithm. **Example-based machine translation** (**EBMT**) is a method of machine translation often characterized by its use of a bilingual corpus with parallel texts as its main knowledge base at run-time. It is essentially a translation by analogy and can be viewed as an implementation of a case-based reasoning approach to machine learning.

At the foundation of example-based machine translation is the idea of translation using an approach known as translation (Use of Python). When applied to the process of human translation, the idea that translation takes place by analogy is a rejection of the idea that people translate sentences by doing deep linguistic analysis. Instead, it is founded on the belief that people translate by first decomposing a sentence into certain phrases, then by translating these phrases, and finally by properly composing these fragments into one long sentence. Phrasal translations are translated by analogy to previous translations. The principle of translation by analogy is encoded to example-based machine translation through the example translations that are used to train such a system.\

Example-based machine translation was first suggested by Makoto Nagao in 1984. B.Pevzner suggested the idea (looks like "translation memory") in 1975 year on seminar "Machine translation" in International center of scientific and technological information. (Moscow). Makoto Nagao pointed out that it is especially adapted to translation between two totally different languages, such as English and Japanese. In this case, one sentence can be translated into several well-structured sentences in another language; therefore, it is no use to do the deep linguistic analysis characteristic of rule-based machine translation. The figure below shows the POS tagger result for a good machine translation from English to Hindi.

*Table: 1.1 POS Tagger result.*

| | |
|---|---|
| 0 | Nonsense |
| 1 | Roughly understandable |
| 2 | Understandable |
| 3 | Good |
| 4 | Perfect |

## II. LITERATURE SURVEY

### 1) A bilingual machine translation system: English & Bengali

Natural language is a fundamental thing of human-society to communicate and interact with one another. In this globalization era, we interact with different regional people as per our interest in social, cultural, economical, educational and professional domain [1]. There are thousands of natural languages exist in our earth. It is quite tough, rather impossible to know all the languages. So we need a computerized approach to convert one natural language to another as per our necessity. This computerized conversion among multiple languages is known as multilingual machine translation. But in this paper we work with a bilingual model, where we concern with two languages: English and Bengali. We use soft computational approach where fuzzy If-Then rule is applied to choose a lemma from prior knowledge; Penn TreeBank PoS tags and HMM tagger are used as lexical class marker to each word in corpora.

**2) Bilingual Corpus Research on Chinese English Machine Translation in Computer Centres of Chinese Universities**

In recent years, monolingual or multilingual (primarily bilingual) [2] corpora are viewed as key resources in language information processing and language engineering projects. A Chinese English parallel corpus is being set up. This paper gives a brief discussion on construction, annotation, and alignment of the parallel corpus. And how it is used in Chinese English Machine translation.

**3) Words to phrase reordering machine translation system in Myanmar-English using English grammar rules**

In machine translation (MT), one of the main problems to handle is word reordering [3]. This paper focuses to design and implement an effective machine translation system for Myanmar to English language.. The framework of this paper is reordering approach for English sentence. We propose an approach to generate the target sentence by using reordering model that can be incorporated into the Statistical Machine Translation (SMTS). Myanmar sentence and English sentence are not semantic. In this paper, we present Myanmar to English translation system that is our ongoing research. Input process, tokenization, segmentation, translation and English sentence generation include in this system. In this paper, we describe about the English sentence generation. The aim of this paper is to reassemble the English word into proper sentence. The resulted raw sentence from translation process is reassembling to form the English sentence. Subject/verb agreement process, article checking process and tense adjustment process will also be performed according to the English grammar rules. The English sentence generation is proper for Myanmar to English translation system

**4) A study to find influential parameters on a Farsi-English statistical machine translation system**

The aim of this paper is to analyze the Farsi-English statistical machine translation systems as a useful communication tool [4]. Improvement of the nation's communication increases the need of easier way of translating between different languages in front of expensive human translators. In this work, a statistical phrase-based system is run on Farsi - English pair languages and the effect of its parameters on the translation quality has been deeply studied. Using BLEU as a metric of translation accuracy, the system achieves an improvement of 1.84%, relative to the baseline accuracy, which is increment from 16.97% to 18.81% in the best case,

**5) Rule** Based **Machine Translation System from English to Tamil**

The main goal of this research is to develop English-Tamil machine translation system using rule-based approaches [5]. For rule based approach, considering the structural difference between English and Tamil languages, syntax transfer based methodology is adopted for translation. This translation engine is a parser, which analyzes the source text, and the corresponding target structure is generated through the transfer lexicon. Morphological generator for Tamil is required to generate the proper Tamil sentence.

## III. PROPOSED WORK

Example based machine translation (EBMT) [6] using NLP is one such response against traditional models of translation [1]. Like Statistical MT, it relies on large corpora and tries somewhat to reject traditional linguistic notions (although this does not restrict them entirely from using the said notions to improve their output). EBMTNLP systems are attractive in that they require a minimum of prior knowledge and are therefore quickly adaptable to many language pairs.

The particular EBMT [2] system that we are examining works in the following way [7]. Given an extensive corpus of aligned source-language and target-language sentences, and a source-language sentence to translate:

1. it identifies exact substrings of the sentence to be translated within the source-language corpus, thereby returning a series of source-language sentences
2. it takes the corresponding sentences in the target-language corpus as the translations of the source-language corpus (this should be the case!)
3. Then for each pair of sentences:
4. it attempts to align the source- and target-language sentences;
5. it retrieves the portion of the target-language sentence marked as aligned with the corpus source-language sentence’s substring and returns it as the translation of the input source-language chunk [3].

## IV. PYTHON PACKAGES USED

1. fuzzy-wuzzy
2. nltk (Natural Language Text Processing Toolkit)

## V. APPROACH: HOW IT IS DONE TECHNICALLY

**Algorithm:**

Step I: Prepare a training dataset containing English equivalent Hindi sentences.

Step II: Initial run to train the application with the training set data.

Step III: Minimum 1 million datasets required to achieve efficiency of 80% and above.

Step IV: Use of some manual dataset as well as from Microsoft's vast collection of corpora.

## VI. WORKING PROCESS

Capture the input text from the user through front end. Now tokenize the sentence using Natural Language Processing Carry out POS Tagging and label them accordingly [8]. Replace all the nouns and adjectives to Hindi directly. Now generate sentences with possible placements of prepositions in Hindi. Now these sentences will be compared for the grammatical precision with our datasets which contain predefined sentences. Now by matching the grammar patterns of our sentence and the dataset a probability ratio is generated. Out of all the sentences the one having highest ratio is selected and shown as output.

## VII. NATURAL LANGUAGE TOOLKIT

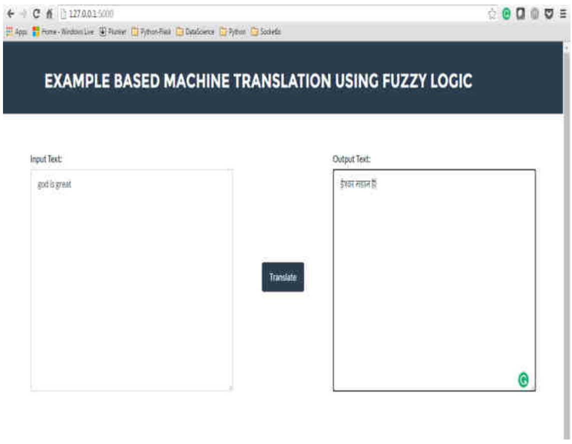
NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3**.**

## VIII. RESULT



EXAMPLE BASED MACHINE TRANSLATION USING FUZZY LOGIC

Input Text:

god is great

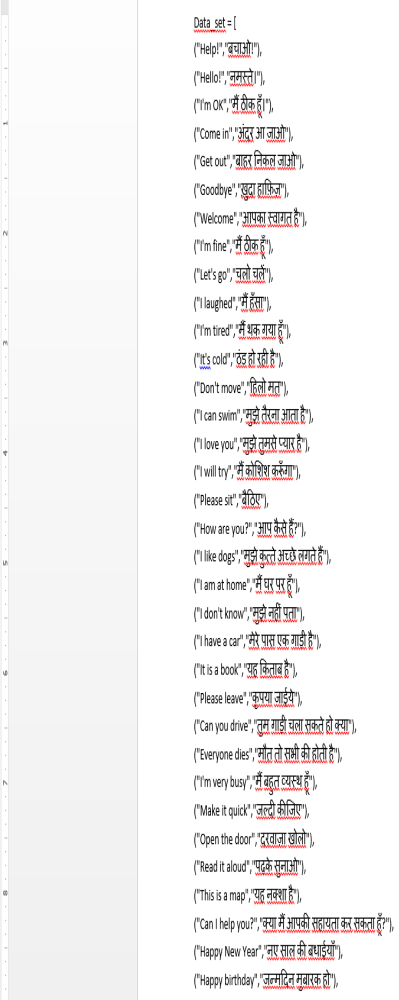
Translate

Output Text:

ईश्वर महान है

*Fig. 1.1*

*Table .1.2*



```
Data_set = [
("Help!","बचाओ!"),
("Hello!","नमस्ते।"),
("I'm OK","मैं ठीक हूँ।"),
("Come in","अंदर आ जाओ"),
("Get out","बाहर निकल जाओ"),
("Goodbye","खुदा हाफ़िज"),
("Welcome","आपका स्वागत है"),
("I'm fine","मैं ठीक हूँ"),
("Let's go","चलो चलें"),
("I laughed","मैं हँसा"),
("I'm tired","मैं थक गया हूँ"),
("It's cold","ठंड हो रही है"),
("Don't move","हिलो मत"),
("I can swim","मुझे तैरना आता है"),
("I love you","मुझे तुमसे प्यार है"),
("I will try","मैं कोशिश करूँगा"),
("Please sit","बैठिए"),
("How are you?","आप कैसे हैं?"),
("I like dogs","मुझे कुत्ते अच्छे लगते हैं"),
("I am at home","मैं घर पर हूँ"),
("I don't know","मुझे नहीं पता"),
("I have a car","मेरे पास एक गाडी है"),
("It is a book","यह किताब है"),
("Please leave","कृपया जाईये"),
("Can you drive","तुम गाडी चला सकते हो क्या"),
("Everyone dies","मौत तो सभी की होती है"),
("I'm very busy","मैं बहुत व्यस्थ हूँ"),
("Make it quick","जल्दी कीजिए"),
("Open the door","दरवाज़ा खोलो"),
("Read it aloud","पढ़के सुनाओ"),
("This is a map","यह नक्शा है"),
("Can I help you?","क्या मैं आपकी सहायता कर सकता हूँ?"),
("Happy New Year","नए साल की बधाईयाँ"),
("Happy birthday","जन्मदिन मुबारक हो"),
```

In Figure 1.1 shows the snapshot of the output for the translation as input in English giving output in Hindi.Table 1.2 shows the dataset used for training of the programme . Python was used for the creation of front end and background databased.

The output shows 80% accuracy in the result.

Example: Input Statement: in English:

“India is great”

Output in Hindi:

“भारत महान है”

## IX. CONCLUSION & FUTURE SCOPE

This research work proposes a new system, which is scalable, transparent and efficient. The entire system will convert the source language text into target language text using natural language processing. It will use the machine translation technique which is better than the existing tools available in the market.

The algorithm is such that, there is dictionary / corpus / vocabulary of English and **Hindi**. The parsing will be proper. The mapping technique will also be used. All the Literals will be separated using partitioning and stemming techniques. The root word will be identified using artificial intelligence and bilingual translation. We pursue the study of example based machine translation using natural language processing.

## REFERENCES

[1] Chandranath Adak “A bilingual machine translation system: English & Bengali” 1-2 Feb. 2014 Page(s):1 - 4 Print ISBN:978-1-4799-3893-3,INSPEC Accession Number:14268273, Conference Location :Hooghy DOI:10.1109/ACES.2014.6808033 Publisher:IEEE

[2] Chan-Juan Liu; Su Han “Bilingual Corpus Research on Chinese English Machine Translation in Computer Centres of Chinese Universities” Published in: ComputerDate of Conference:11-13 Aug. 2012 Page(s):1720 - 1723 Print , ISBN:978-1-4673-0721-5 , INSPEC Accession Number:13227134 , Conference Location :Nanjing DOI:10.1109/CSSS.2012.430, Publisher: IEEE.

[3] Aye Thida Win “Words to phrase reordering machine translation system in Myanmar-English using English grammar rules” Published in:Computer Research and Development (ICCRD), 2011 3rd International Conference on (Volume:3 ) Date of Conference:11-13 March 2011 Page(s):50 - 53 ,Print ISBN:978-1-61284-839-6 , INSPEC Accession Number:11975817 Conference Location :Shanghai DOI:10.1109/ICCRD.2011.5764243, Publisher:IEEE

[4] Somayeh Bakhshaei; Shahram Khadivi; Noushin Riahi; Hossein Sameti “A study to find influential parameters on a Farsi-English statistical machine translation system” Published in:Telecommunications (IST), 2010 5th International Symposium on Date of Conference:4-6 Dec. 2010 Page(s): 985 - 991 Print ISBN: 978-1-4244-8183-5 INSPEC Accession Number:11875599 Conference Location :Tehran DOI:10.1109/ ISTEL. 2010.5734165, Publisher :IEEE.

[5] M. Kasthuri; S. Britto Ramesh Kumar “Rule Based Machine Translation System from English to Tamil” Published in:Computing and Communication Technologies (WCCCT), 2014 World Congress on Date of Conference:Feb. 27 2014-March 1 2014 Page(s): 158 - 163 Print ISBN: 978-1-4799-2876-7 INSPEC Accession Number: 14220410 Conference Location :Trichirappalli DOI:10.1109/WCCCT.2014.50 Publisher:IEEE

[6] Manish Rana, Mohammad Atique, “Example Based Machine using fuzzy logic from English to Hindi” Int'l Conf. Artificial Intelligence , ICAI'15 , pp 354-359.

[7] Manish Rana, Mohammad Atique, “Example Based Machine using various soft computing techniques review" *IJSER* “International Journal of Scientific & Engineering Research”, Volume 6, Issue 4, April-2015, ISSN 2229-5518E pp.1100-1106.

[8] Manish Rana, “Review: Machine using various soft-computing Tools “International Conference on communication computing & Virtualization Vol 6 Issue 2, Feb 23 & 24, 2015 pp. 813- 816..

[9] H. H. Owaied, M. M. Qasem, “Developing Rule-Case-Based Shell Expert System,” *Proc.of Int. MultiConf. of Engineers & Scientists*, vol. 1, 2010.

[10] M. G. Tsipouras, C. Voglis, D. I. Fotiadis, “ A Framework for Fuzzy Expert System Creation - Application to Cardiovascular Diseases,” *IEEE Transactions Biomedical Engg*., vol. 54, no. 11, pp. 2089-2105, 2007.

[11] Jort F. Gemmeke*, Student-Member, IEEE, Tuomas Virtanen, Antti Hurmalainen “*Exemplar-based sparse representations for noise robust automatic speech recognition”*This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publicationCopyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, Permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.Pages: 1-14

[12] Mehrez Boulares1 and Mohamed Jemni2, Research Lab. UTIC, University of Tunis, 5, Avenue Taha Hussein, B. P. : 56, Bab Menara, 1008 Tunis, Tunisia *"Toward an example-based machine translation from written text to ASL using virtual agent animation"* IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012

[13] Stephen J. Wright, Dimitri Kanevsky, Senior Member, IEEE, LiDeng, Fellow, IEEE,, Xiaodong He, Senior Member, IEEE, Georg Heigold, Member, IEEE, and Haizhou Li, Senior Member, IEEE *"Optimization Algorithms and Applications for Speech and Language Processing"*IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 11, NOVEMBER 2013 Pages: 2231-2243

[14] Hongshen Chen, Jun Xie, Fandong Meng, Wenbin Jiang Qun Liu "*A Dependency Edge-based Transfer Model for Statistical Machine Translation*" Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, August 23-29 2014, pages 1103–1113.