INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 2)
Available online at: www.ijariit.com

EBMT Sindhi to Hindi Sentence Translation System

Nisha S. Tathe
nishatathe@gmail.com
Datta Meghe Institute of Engineering, Technology, and Research, Wardha, Maharashtra

Jayasha S. Kriplani
kriplanijayesha@gmail.com
Datta Meghe Institute of Engineering, Technology, and Research, Wardha, Maharashtra

Shiwani S. Deshmukh
shiwanideshmukh51@gmail.com
Datta Meghe Institute of Engineering, Technology, and Research, Wardha, Maharashtra

Nikita K. Sadrani
nikitasadrani5@gmail.com
Datta Meghe Institute of Engineering, Technology, and Research, Wardha, Maharashtra

Monika A. Panjwani
panjwanimonika217@gmail.com
Datta Meghe Institute of Engineering, Technology and Research, Wardha, Maharashtra

ABSTRACT

Example-Based Machine Translation (EBMT) is the system which translates sentences from Sindhi to Hindi. It uses the parallel corpus for translating sentences. To handle a large volume of data the computational model is required. The requirement for a computational resource in EBMT is significantly less as compared to MT (Machine Translator). This makes the development of EBMT systems for Sindhi to Hindi translation feasible, where availability of large-scale computational resources is still scarce. Example based machine translation relies on the database for its translation. The software helps to get a sentential translation according to semantics and syntactic of the language. The frequency of word occurrence is important for translation in EBMT.

Keywords: *Translator, Database, Sentence, Language, Software, Sentential Translation.*

1. INTRODUCTION

A. Machine Translation

EBMT for Sindhi to Hindi Machine Translation is the software framework. EBMT is an example based machine translation. This Machine Translator is a corpus-based machine translation, which requires parallel-aligned three machine-readable corpora. Here, the already translated example serves as knowledge to the system. This approach derives the information from the corpora for analysis, transfer, and generation of translation. These systems take the source text and find the most analogous examples from the source examples in the corpora. The next step is to retrieve corresponding translations and the final step is to recombine the retrieved translations into the final translation and then to display the final desired output into the Target Text.

EBMT is best suited for sub-language phenomena like phrasal verbs; weather forecasting, technical manuals, air travel queries, appointment scheduling, etc. Since building a generalized corpus is a difficult task, the translation work requires annotated corpus, and annotating the corpus, in general, is very much critical and complicated task.The Example-Based Machine Translation uses in-build knowledge database to translate the Sindhi language into the Hindi language. The subtitles of source language as an input will be given by User. Each and every word will be assigned with tokens via training module. The training module will also check the most analogous examples in the stored database of the source language. After that, the corresponding examples of the target language from the database which is stored as the meaning of source language will also retrieve when the input strings and stored source language strings will match.

Accordingly, the analogous matching source text and their stored meaning in the target language will also be assigned with the precise tokens. Then the tokenize strings from input and target database will retrieve at the section. Then according to their tokens, the stings will be adapted and recombined.

The recombination of the strings will take place precisely according to semantics and syntactic. The work of matching of

the strings from source text to the input text is done by the testing module which serves by performing matching the strings. In the Machine translation after having the strings matched and retrieved the target text is displayed in output subtitles box according to the target text Grammer. Then we have finally got the desired output of Hindi language.

This is the methodology takes place in the task of Machine translation. In this, the process is certainly dependent on the knowledge database stored and provided for translation. This example-based machine translation helps to deal with the large amount data simultaneously. This makes the translation easy, efficient and reliable. If the input doesn't match with the source text database then the word or string will be directed towards the dictionary. In the dictionary, there will be words stored which is not present in the linguistic database of examples. So, the word will be directed toward dictionary and then dictionary will retrieve its meaning in the target language. This how the translation works from the perspective of dictionary use.

The need for machine translation can be briefly stated into following points briefly:

- Too much to be translated
- Boring for human translators
- The major requirement that terminology used consistently
- Increase speed and throughput
- Top quality translation not always needed
  Reduced cost

B. Example-Based Machine Translation

Translation is the communication of the meaning of a source-language text by means of an equivalent target-language text. Nowadays, machine translator is the massive topic for innovation in technology based on example-based machine translation. Sindhi is a rare language so there is no availability of a machine translator for it. The translator needs to be able to take the subtitles in the source text and transmits those very same subtitles in the target language. It is not simply a question of getting a dictionary and taking it one word at a time.

The translator is going to help the Sindhi known persons to convert their very own language into Hindi. So, that they can also learn Hindi too. Not only the people who know Sindhi but also the people who are facing problems in text kind data can also use this interface to get better results. The concept of EBMT relies on the database for its translation.

The natural language is used by every common man. The language which we all speak is termed as natural language. We need this language for communication. Natural language processing is operating i.e. processing, refining, modifying and translating on one of these natural languages. Therefore for understanding and making communication easy there is the basic need of a translator.

This translation can be done by humans; so why there is need of machine translation? The first reason is that text in "world of text" is huge. There are many large documents to be translated and it is not possible for a human to translate gigabytes of data in less time. To reduce the human efforts and to give the results quickly the translators are used which can translate the text from one language to another by just one click.

A second reason is that the whole technical materials are too boring to translate for human translators as humans do not like to translate them continuously and consistently. Hence they look for help from computers. Thirdly, as far as large corporations are concerned, there is the major requirement that terminology is used consistently; they want terms to be translated in the same way every time. Computers are consistent, but human translators tend to seek variety; they do not like to repeat the same translation and this is no good for technical translation. A fourth reason is that the use of computer-based translation tools can increase the volume and speed of translation throughput, and organizations like to have translations immediately. The fifth reason is that top quality human translation is not always needed. Computers do not produce good translations.

The term machine translation (MT) is a translation of one language to another. The ideal aim of machine translation system is to produce the best possible translation without human assistance Basically every machine translation system requires automated programs for translation, dictionaries, and grammars to support translation.

C. Problem Statements

- Lack of Machine Translators for the Sindhi to Hindi language translation
- All Sindhi people are not aware of Hindi Textual
- It is hard for Sindhi unknown person to translate a Sindhi textual data
- Lack of Human Translators
- A human can't translate a huge amount of data at a time

2. LITERATURE REVIEW

A. Machine Translation

The data collection is used as references to gain information during the research conducted. These literature reviews discuss the information gathered by reading research papers & websites. There are various IEEE papers and Journals but they have their own ideas and way of implementation which we have adopted for our project. None of the Journals have mentioned designing a portal for Sindhi to Hindi Machine Translation. So, this was our main aim to design a system of Machine Translation. Some of the Journals and Research Paper which we have referred are:

- This paper proposed a system named as Anglabharti to translate the English language into Hindi, Bangla, Asamiya, Punjabi, Marathi, Oriya, Gujarati, etc. This system works for example-based machine translation. It is a pattern directed rule-based system with context-free grammar like structure for English (source language).[1]

- A Bilingual Bengali in 2002 proposed Assamese automatic MT system for translating the news texts by using the Example-Based Machine Translation (EBMT) technique. In this the translation is done at sentence level. Some preprocessing and post-processing work has to be done for the translation. The longer sentences were fragmented at punctuation, which gives high quality translations. [6]

- In this paper, published an EBMT named as IBM-MTS which generates translation of a given sentence by retrieving similar past translation examples from its example-based system and then adapting them suitably to meet the current translation requirements. We have adapted an idea of creating and building an efficient example-based machine translator based on its knowledge from this paper published.[3]

- The paper focuses on Example-Based Machine Translation (EBMT) system that translates sentences from English to Hindi. Example based machine translation relies on the database for its translation. The frequency of word

occurrence is important for translation in EBMT in the following research. We have taken the idea of the modules used in this paper for our project. The modules used in this paper are very easy to understand and implement. So, this paper have a huge contribution in the system development. [2]

- In this paper, they use a Spectral Clustering for Example-Based Machine Translation to cluster words, and this is shown to result in as much as 29.08% improvement over the baseline EBMT system. The spectral clustering is kind of complex to implement but the idea of enhancing the machine translation via these path has let us think about the idea of more machine improvement. [5]

- It is a Hybrid Machine Translation System in 2004, a MT system for translating English news headlines to Bengali at Jadavpur University Kolkata. It was developed in 2005 developed EB-ANUBAD , a system for translating English to Bengali language and it shows 98% correct results although the output was in English language and this was used to understand the Bengali language. The approach taken for the translation was the hybrid approach of rule-based and transfer based with a parser for both morphological as well as for the Lexical parsing of the text. A brief description of the various machine translation systems.[4]

- In 2004 two Machine Translation systems were developed jointly by IIIT Hyderabad, IIS Bangalore, and Carnegie Mellon University USA for translation from English to Hindi. The Shiva system uses example-based approach and the system uses rule-based approach with a statistical approach for Machine Translation. The Shakti system is working for three target languages Hindi, Marathi, and Telugu. The statistical approach used in this paper is better to understand so we have adopted this idea into our system.[7]

- In this paper, introduced a novel example-base machine translation method. The proposed method was motivated by an observation that the set of labels and messages of a software package possesses a lot of redundancy and structure. This redundancy might allow a very simple approach to somewhat reduce the job of the human translator by producing reliable translation by analogy with the already created ones. Experiments have shown, that it is indeed the case, and the translation of about a quarter of the messages in the considered example which can be derived by analogy from the other translations.[8]

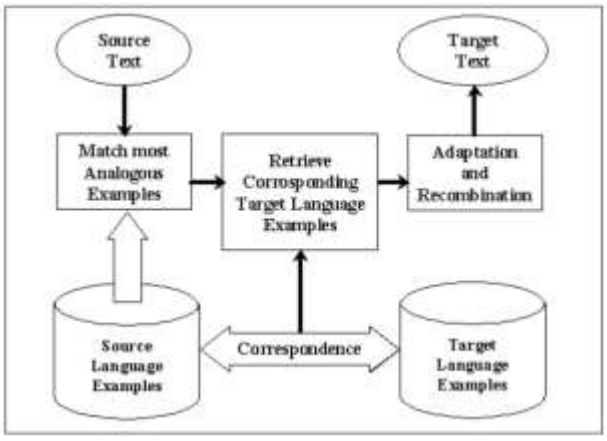
- Speech Synthesizer is the paper published to Text to-Speech (TTS) technology converts a given text into a corresponding speech waveform. TTS system based on unit selection approach using festival framework Text processing, G2P (grapheme to phoneme) and speech generation are the main components of a TTS system. Text processing component produces appropriate sequence of phonemic units for a given input text. These phonemic unit realized by the speech generation component using this kind of digital signal processing algorithm.[9]

3. SYSTEM ARCHITECTURE

The EBMT is the example-based machine translation 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. The Example-Based Machine Translation uses in-build knowledge database to translate the Sindhi language into Hindi language. The subtitles of source language as an input will be given by User. The knowledge base, parallel aligned corpora consist of two sections, one for the source language examples and the other for the target language examples. Each example in the source section has one to one mapping in the target language section. The Training and Testing modules performs the further operations of the translation from Sindhi to Hindi language

The subtitles of source language as an input will be given by user Each and every word will be assigned with a tokens via training module.

The training module will also check the most analogous examples in the stored database of source language After that the corresponding examples of the target language from the database will also retrieve when the input strings and stored source language strings will match**.**



The matching process may be syntactic or semantic level or both, depending upon the domain. On the syntactic level, matching can be done by the structural matching of the phrase or the sentence. In semantic matching, the semantic distance is found out between the phrases and the words. The corresponding translated segments of the target language are retrieved from the second section of the corpora. In the final phase of translation, the retrieved target segments are adapted and recombined to obtain the translation. It identifies the discrepancy between the retrieved target segments with the input sentences' tense, voice, gender, etc. The divergence is removed from the retrieved segments by adapting the segments according to the input sentence's features.

4. METHODOLOGY

A. Algorithm

There are two main algorithms used for this system development. The first aim is to train the database for translation and testing converts the sentence from Sindhi to Hindi.

- Training Algorithm
- Testing Algorithm

i) Training Algorithm

a) Matching String:

The matching for the word is done using two parts.
The word which is found in the database. The length which gives occurrence of the word. If a word occurs only once in the database, then the word is searched in the dictionary. The sentence will be compared to finding the string. If the common string is encountered it stores the string and appends it with the previous match. The word for word translation is performed. The Sentence to be translated we will write it as input in the first text box. The translation button is pressed to perform the translation. If we have taken an example "Kedha haal aahin".

The tokens are generated from the input sentence. The sentence is divided into tokens. Therefore the sentence tokens are "Kedha", "haal", "aahin".The occurrence of these words is checked from the database. If it occurs only one time, then we will check it from dictionary else if the search matched. Then the output retrieved is the word to word translation. So the output of the sentence will be "Kaise ho tum".

The Training System Algorithm is as shown below:

Step 1: Train the database.
Step 2: Read the input String.
Step 3: Tokenize the string. 13
Step 4: Parse the database for the first word that is tocken.
Step5:If the word is present only once in a database, use the dictionary to search for the translation else Go to 6.
Step 6: If the word is present twice in the database then it finds the common string in the Hindi language to find the match.
Step 7: The common string is founded.
Step 8: Stop.

ii) Testing Algorithm

The Testing is the module where the afterward work is taking place after training module is performed. The matched strings of the input from the Source language and the Target language is retrieved in this process. The process of testing refers to the arrangement of input and output sentential examples according to its semantics and syntactic of the Grammar. The input and output are being updated according to its grammar to get a meaningful sentence. This meaningful sentence which is retrieved is then displayed in the target language subtitles as an output.
The Testing System Algorithm has following steps as follows:

Step 1: Retrieve the matching strings
Step 2: Recombine the matching strings of the target language
Step 3: Arrange these matching strings according to its Grammar
Step 4: Display the meaningful sentence into target language subtitle box
Step 5: Stop

B. Result Analysis

i) Training Module:

The Language Converter project runs and gives a window for training and translation of sentences from Sindhi to the Hindi Language. The words and sentences both will translate efficiently.

Fig (4.1): Language Converter

The training algorithm stores the sentences in the source language and its target language translation. The sentences get store into the database.

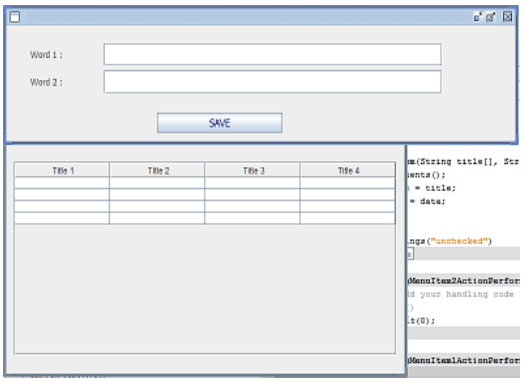


Fig (4.2): Training Algorithm

ii) Testing Module:

The Testing algorithm gives the window for testing the translations or getting the accurate translation of a sentence. The testing window shows the translation in Sindhi sentences with accurate semantics stored in the database.

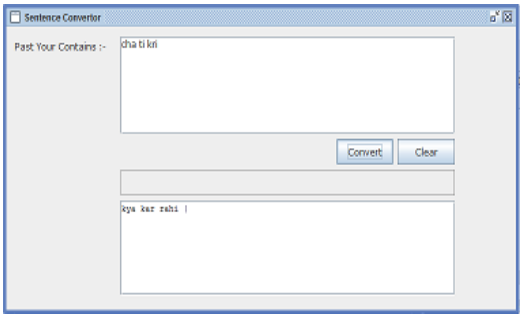


Fig (4.3): Testing Algorithm

The database with the source language text and the target language text with its proper translation is stored. The database is ready with the corpus data of sentences.

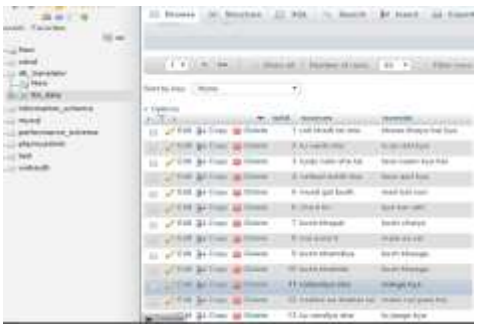


Fig (4.4): Database

5. FUTURE SCOPE

a) Real-Time Image Translation:

Real-time image translation means translation of a text from the image. The user can load the image or take the picture and get the translation of given text in the required form. This feature will make the user take the picture of an image for which user needs its translation. It should provide an output on present time only. In real time image translation feature, the text is retrieved and provide the correspondence meaning of the text.

It helps user/one to understand the meaning of given text in the image when they go out and see the boards with different language which is not familiar to them.
For Example:-

Fig(5.1) : Picture was taken from an image containing text in the source language

Suppose that we have to get the meaning of text given in the image, so we can take pictures of the image as shown in above figure and process it to get the meaning of the text

Fig(5.2): Translation of given text in image in the target language

b) Voice Translation:

Voice translation is the process in which user can speak in a particular language and get the translation in text form in familiar language. In voice translation, user can get the required translation in voice form also. Voice translation is the technology which enables speaker of different languages to communicate.

Voice translation can be built for sindhi translation which will translate the sindhi language voice into Hindi. It can provide the output in text form or in the voice form.
For Example:-

Fig(5.3) : Voice translation of the source language into Target language

The voice translation shown in above figure can be done with the sindhi language as a source language and hindi language as a target language. In above figure translation is done with voice to voice translation, but same can be done in voice to text form/translation.

6. CONCLUSION

This system will provide a proper translation of Sindhi language into Hindi language. It will give the meaning of full sentences in Hindi.The sentence will be in structured form. This system will be introduced as an aid for the Sindhi persons who don't know Hindi language. The example based machine translation is providing an efficient way of translating in fastest time. This research focuses on simple way of comparing sentence to extract the translation. The research concludes the translator gives the proper expected output to some extent by comparing sentences. The research can be taken to next level by using preprocessed database. The translator presented in the research work does not structure the output sentence.

7. REFERENCES

[1] R.M.K. Sinha,"An Engineering Perspective of Machine Translation", AnglaBharti-II and AnuBharti-II Architectures. In proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS-2004). November 17-19. Tata McGraw Hill, New Delhi. pp. 134-38, 2004.
[2] A Pure EBMT Approach for English to Hindi Sentence Translation System by prof. Ruchika A. Sinhal and Kapil O. Gupta.
[3] D. Gupta, N. Chatterjee,"Identification of Divergence for English to Hindi EBMT", In proceedings of MT SUMMIT IX. New Orleans, Louisiana, USA. pp. 157-162, 2003.
[4] Sivaji Bandyopadhyay,"Use of Machine Translation in India", AAMT Journal, 36. pp. 25-31, 2004.
[5] Indranil Saha et.al. (2004). Example-Based Technique for Disambiguating Phrasal Verbs in English to Hindi Translation. Technical Report KBCS Division CDAC Mumbai.
[6] Vijayanand Kommaluri, Sirajul Islam Choudhury, Pranab Ratna, "VAASAANUBAADA-Automatic Machine Translation of Bilingual Bengali-Assamese News Texts", *Langu-age Engineering Conference. Hyderabad, India.* 2002. [Online] Available: http://www.portal.acm.org/citation.cfm?id=788716
[7] R. Moona Bharati, P. Reddy, B. Sankar, D.M. Sharma, R. Sangal,"Machine Translation: The Shakti Approach. Pre-Conference Tutorial", *ICON-2003*. [Online] Available:http://www.ebmt.serc.iisc.ernet.in/mt/login.html,[Online]Available:http://www.gdit.iiit.net/~mt/shaki
[8] R. Brown. Example-based machine translation in the pangloss system. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pages 169–174, 1996. URL http://www.cs.cmu.edu/ ralf/papers.html.
[9] Speech enabled Integrated AR-based Multimodal Language Translation Mahesh Bhargava, Pavan Dhote, Amit Srivastava , Ajai Kumar AAI, Centre for development of Advanced Computing(C-DAC), Pune, India
mbhargava@cdac.in , pavand@cdac.in, asrivastava@cdac.in
[10] Machine Translation from Technology Development for Indian Languages (TDIL), Department of Electronics & Information Technology (DeitY), Ministry of Communications & Information Technology, Government of India. Available: http://tdildc.in/components/com_mtsystem/CommonUI/homeMT.php

[11] K. Kintzley, A. Jansen, and H. Hermansky, "Text-to-speech inspired duration modeling for improved whole-word acoustic models," in Proceedings of INTERSPEECH. ISCA, 2013

[12] Steven Bird, Ewan Klein and Edward Loper, Natural Language Processing with Python, 1st ed., O'Reilly Media, June 2009.

[13] Raghavendra Udupa, Tanveer A. Faruquie,"An English-Hindi Statistical Machine Translation System", *In proceedings of First International Joint Conference*, Hainan Island, China, March 22-24, pp. 254-262, 2004.

[14] B. K.Murthy, W. R.Deshpande,"Language technology in India: past, present and future", *In proceedings of MLIT Symposium 3. GII/GIS for Equal Language Opportunity*. Vietnam. October 6-7. Pp.134-137, 1998

[15] D. Arnold, L. Balkan, S. Meijer, L.L. Humphreys, L. Sadler: *Machine Translation: an Introductory Guide*. Blackwells-NCC, London, Great Britain, 1994.

[16] Hutchins 95 J. Hutchins: Reflections on the history and present state of machine translation. In Proc. of *Machine Translation Summit V*, pp. 89–96, Luxembourg, July 1995.

