Example Based Machine Translation Using Natural Language Processing

Sunny Bhavan Sall
I/C HOD
(Department of Computer Engineering)
Sardar Vallabhbhai Patel Polytechnic,
Mumbai,India
sunny_sall@yahoo.co.in

Dr. Rekha Sharma
Deputy HOD
Department of Computer Engineering
Thakur College of Engineering &
Technology
Mumbai ,India
rekha.sharma@thakureducation

.org
Dr. R R Shedamkar
Dean Academicsand and HOD
Department of Computer Engineering
Thakur College of Engineering &
Technology
Mumbai ,India
rrsedamkar27@gmail.com

Abstract -

Machine translation (MT) research has come a long way since the idea to use computer to automate the translation process and the major approach is Statistical Machine Translation (SMT). An alternative to SMT is Example-based machine translation (EBMT). Among machine translation systems, traditional transformational methods are somewhat difficult to construct, as they basically involve hard coding the idiosyncrasies of both languages Natural Language Processing deals with the processing of natural language. The language spoken by the human beings in day to day life is nothing but the natural language. There are many different applications under NLP among which Machine Translation is one of the applications. In this paper, we describe the Example Based Machine Translation using Natural Language Processing. The proposed EBMT framework can be used for automatic translation of text by reusing the examples of previous translations. This framework comprises of three phases, matching, alignment and recombination.

Index Terms— Machine Translation, Example-Based Machine Translation, Natural Language Processing.

1 Introduction

Example based machine translation (EBMT) is one such response against traditional models of translation. 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). EBMT systems are attractive in that they require a minimum of prior knowledge and are therefore quickly adaptable to many language pairs. We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material. Machine translation (MT) research has come a long way since the idea to use computer to automate the translation process and the major approach is Statistical Machine Translation (SMT). An alternative to SMT is Example-based machine translation (EBMT) [1]. The most important common feature between SMT and EBMT is to use a bilingual corpus (translation examples) for the translation of new inputs. Both methods exploit translation knowledge implicitly embedded in translation examples, and make MT system maintenance and improvement much easier compared with Rule-Based Machine Translation.

On the other hand, EBMT is different from SMT in that SMT hesitates to exploit rich linguistic resources such as a bilingual lexicon and parsers. EBMT does not consider such a constraint. SMT basically combines words or phrases (relatively small pieces) with high probability [2]; EBMT tries to use larger translation examples. When EBMT tries to use larger examples, it can better handle examples which are discontinuous as a word-string, but continuous structurally. Accordingly, though it is not inevitable, EBMT can quite naturally handle syntactic information. Besides that, the difference in between EBMT and SMT, EBMT is not the replacement for SMT. SMT is a natural approach when linguistic resources such as parsers and a bilingual lexicon are not available. On the other hand, in case that such linguistic resources are available, it is also natural to see how accurate MT can be achieved using all the available resources. EBMT is a more realistic, transparent, scalable and efficient approach in such cases. The language spoken by the human beings in day to day life is nothing but the natural language. There are many different applications under NLP among which Machine Translation is one of the applications. The work on machine translation began in late 1947. Machine translation deals with translating one natural language to another.

The ideal aim of Machine Translation system is to give the possible correct output without human assistance. The example based machine translation use the former examples as the based for translating source language to target language. The database for the two languages is considered for translation. Example based machine translation is bilingual translation. Example based machine translation use the corpus of two languages, the target language and the source language. We are proposing the design and development of an EBMT system. In this system, the English text entered by the user in the box is to be converted to Hindi without any divergence. The sentence i.e. text entered at the source side will be fragmented and the fragmented text will be matched into the corresponding target text. This will be done by using the data mining and the tree formation of the source text. The output then obtained will be aligned and the sentence having proper structure and the meaning will be generated using the corpus. The Example based machine translation is one of the approaches in machine translation. The concept uses the corpus of two languages and then translates the input text to desired target text by proper matching. The different languages have different language structure of the subject-object-verb (SOV) alignment. The matching is then arranged to give proper meaning in target text language and to form proper structure.

modify the header or footer on subsequent pages.

The various issues with MT systems

All The task of translation is needed in day to day life. Humans can also do the task of translation; but now-a-days there is too much data to be coped with. So the job becomes tedious; therefore there is need for a translator which gives proper results for a text without any human assistance. The text should be translated properly without any divergence in the translation; i.e. the output for translation should be proper and no meaningless translation should be done. The speed of translation can also be increased.

The translation done till now is not accurate , to give results with the divergence in conversion form source language to target language. There are certain drawbacks which does not give translation without human assistance. There is a genuine requirement of having a machine translation system which can overcome the limitations of existing machine translation systems , and provide the translated content with high relevance and precision. EBMT is trying to minimize the human assistance and still give a better translation.

Recently corpus based approaches to machine translation have received wide focus. They are namely Example Based Machine Translation (EBMT) [6] and Statistical Machine Translation (SMT) [7]. A combination of statistical and example-based MT approaches shows some promising perspectives for overcoming the shortcomings of each approach. Efforts have been made in this direction using the alignments from both the methods to improve the translation [8], to improve the alignment in the EBMT using the statistical information computed from SMT methods [9] etc. The results obtained have shown improvement in performance. However, these approaches cannot directly be applied to Indian languages due to the small size of the parallel texts available and sparse linguistic resources. Also some of the assumptions made in some of these approaches like marker hypothesis [10], cannot directly be applied to translate from english to Indian languages since word order in the source and target languages is very different and sequential word orderings between source and target sentences do not exist.

Machine translation of Indian Languages has been pursued mostly on the linguistic side. Hand crafted rules were mainly used for translation, [11], [12]. Rule based approaches were combined with EBMT system to build hybrid systems [13], [14] performs interlingua based machine translation. Input in the source language is converted into UNL, the Universal Networking Language and then converted back from UNL to the target language. Recently, Gangadhariah et al [15] used linguistic rules are used for ordering the output from a generalized example based machine translation [16]. While, in general in the machine translation literature, hybrid approaches have been proposed for EBMT primarily using statistical information most of which have shown improvement in performance over the pure EBMT system. [17] automatically derived a hierarchical TM from a parallel corpus, comprising a set of transducers encoding a simple grammar. [18] used example-based re-scoring method to validate SMT translation candidates. [19] proposed an example based decoding for statistical machine translation which outperformed the beam search based decoder [20]. Kim et al [9] showed improvement in alignment in EBMT using statistical dictionaries and calculating alignment scores bi-directionally. [8], [21] combined the sub-sentential alignments obtained from the EBMT systems with word and phrase alignments from SMT to make ’Example based Statistical Machine Translation’ and ’Statistical Example based Machine Translation’.

The EBMT module shares similarities in structure with three stages : analysis, transfer & generation as shown in the figure 1. The Vauquois Pyramid adapted for EBMT [22].

- Direct
- Transfer
- Interlingual

minimum of prior knowledge and are therefore quickly adaptable to many language pairs. The particular EBMT system that we are examining works in the following way. 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:
   i. It attempts to align the source- and target-language sentences;
   ii. 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

.

The above system is a specialization of generalized EBMT systems. Other specific systems may operate on parse trees or only on entire sentences. The system requires the following:

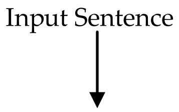
1. Sentence-aligned source and target corpora.
2. Source- to target- dictionary
3. Stemmer

The stemmer is necessary because we will typically find only uninflected forms in dictionaries. While it is consulted in the alignment algorithm, it is not consulted in the matching step—as stated before, those matches must be exact.

PROPOSED WORK

The Example based machine translation is one of the approaches in machine translation. The concept uses the corpus of two languages and then translates the input text to desired target text by proper matching.

The different languages have different language structure of the subject-object-verb (SOV) alignment. The matching is then arranged to give proper meaning in target text language and to form proper structure. In this paper, we describe the Example Based Machine Translation using Natural Language Processing. The proposed EBMT framework can be used for automatic translation of text by reusing the examples of previous translations. This framework comprises of three phases, matching, alignment and recombination.



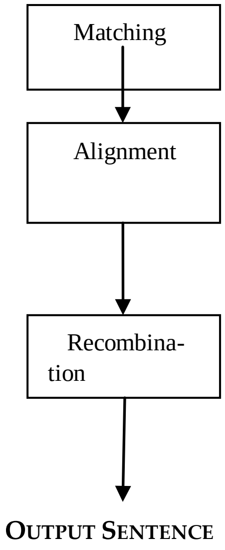


FIGURE 1 : PROPOSED SYSTEM

Matching Phase

Searching the source side of the parallel corpus for 'close' matches and their translations.

Alignment Phase

Determining the sub sentential translation links in those retrieved examples.

Recombination Phase

Recombining relevant parts of the target translation links to derive the translation.

ALGORITHM

3.1 Indexing

In order to facilitate the search for sentence substrings, we need to create an inverted index into the source-language corpus. To do this we loop through all the words of the corpus, adding the current location (as defined by sentence index in corpus and word index in sentence) into a hash table keyed by the appropriate word. In order to save time in future runs we save this to an index file.

3.2 Chunk searching and subsuming

1. Keep two lists of chunks: current and completed.

2. Looping through all words in the target sentence:
   i. See whether locations for the current word extend any chunks on the current list

   ii. If they do, extend the chunk.
   iii. Throw away any chunks that are 1-word. These are rejected.
   iv. Move to the completed list those chunks that were unable to continue
   v. Start a new current chunk for each location

3. At the end, dump everything into completed.

4. Then, to prune, run every chunk against every other:
   i. If a chunk properly subsumes another, remove the smaller one
   ii. If two chunks are equal and we have too many of them, remove one

3.3 Alignment

The alignment algorithm proceeds as follows:

1. Stem the words of specified source sentence

2. Look up those words in a translation dictionary

3. Stem the words of the specified target sentence

4. Try to match the target words with the source words—wherever they match, mark the correspondence table.

5. Prune the table to remove unlikely word correspondences.

6. Take only as much target text as is necessary in order to cover all the remaining (unpruned) correspondences for the source language chunk.

The pruning algorithm relies on the fact that *single* words are not often violently displaced from their original position..

EBMT IMPLEMENTATION

Example Based Machine Translation is based on the idea to reuse the previously done translations as examples. Following are three examples are given. EBMT system tries to translate the given input English Text into Hindi by using these previous translations.

Example 1
English : India won the match.
Hindi : Hkkjr us eWp ftrk
Example 2
English : India is the best
Hindi : Hkkjr loZJs"B gS
Example 3
English : Sachin plays well
Hindi : lfpu zvPNk [ksyrk gS
Input
English : Sachin is the best
Translation (Output)
Hindi : lfpu loZJs"B gS

Table 1 illustrates the comparison of three machine translation techniques, Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT) on the basis of various parameters such as Consistency, predictable quality, Quality of out of domain translation, Use of grammar, robustness, Fluency and performance.

Table 1. Comparison of various Machine Translation schemes

Parameter	RBMT	SMT	EBMT
Consistency	High	Low	Medium
Predictable Quality	Good	Similar	Very well
Out of Domain Quality	Medium	Low	High
Use of Grammar	Yes	No	No
Robust	Yes	No	Yes
Fluency	Less	Medium	High
Performance	Good	Medium	Good

We have also presented some translation results of some existing machine translation tools and our system in Table 2 .

Table 2. Comparison of translation of text from various Machine Translation tools

English Sentence	Hindi Translation by existing MT tools	Hindi translation by our EBMT translation
India is great	Hkkjr gS egku	Hkkjr egku gS
I am a boy	eS gw yMdk	eS yMdk gw
She is beautiful	Okks gS [kqclqjr	oks [kqclqjr gS
He was great	Okks Fks egku	os egku Fks
Where do you live ?	dgk gks vki jgrs ?	vki dgk jgrs gks ?
She reads book	oks i<rh gS fdrkc	oks fdrkc i<rh gS

Milk is white	nq/k gS lQsn	nq/k lQsn gS

.

MAJOR EXPECTED RESULTS

Based on this model expected result is as following:

Input: Source Language, **ENGLISH** (SL).
Output: Target Language, **HINDI** (TL).
The result obtained is with minimal human interface.

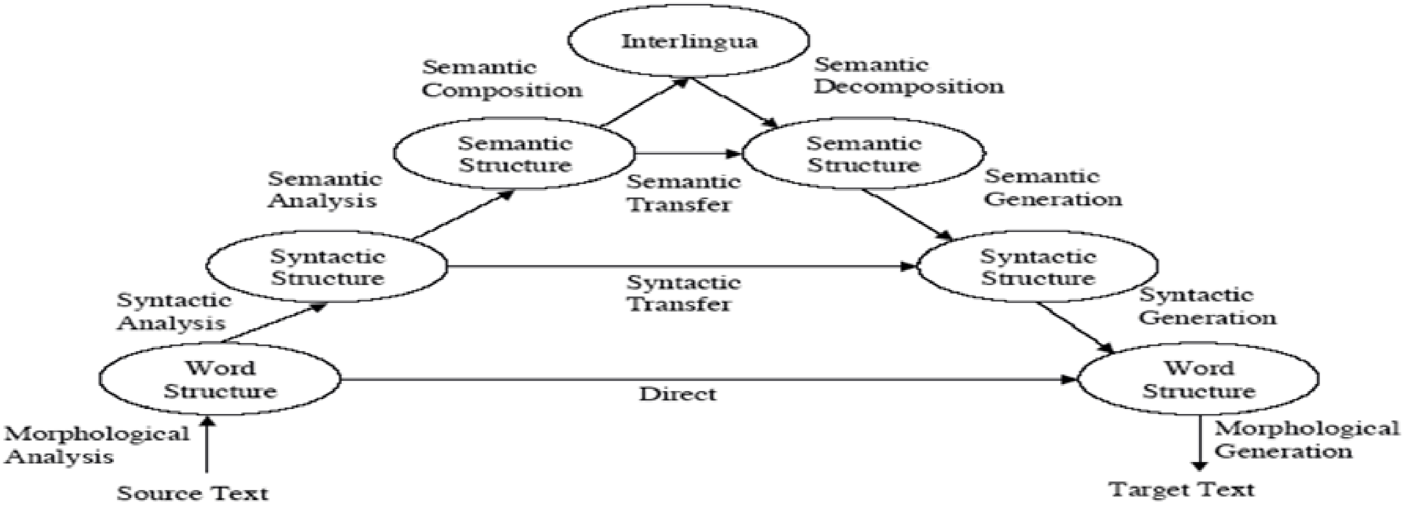
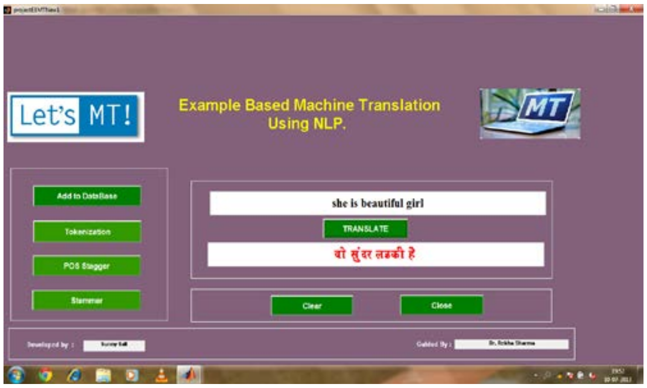


Figure 2 The Vauquois Tringle for MT

RESULT

The work presented in this paper shows the results "Example Based Machine Translation Using Natural language Processing" as shown in the snap shot figure 3 and figure 4. The paper shows simple translation of English sentences in Hindi sentences.



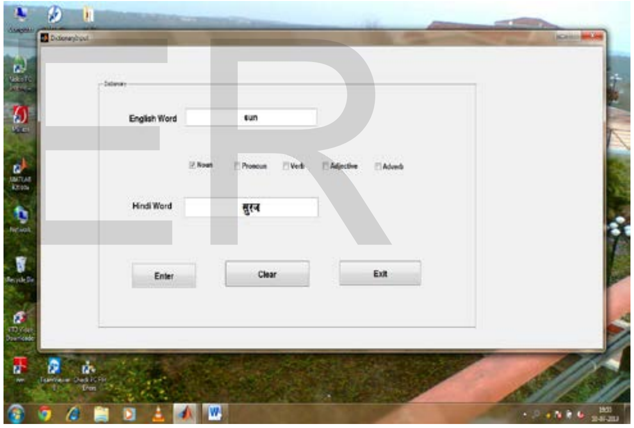
she is beautiful girl

वो सुंदर लड़की है

Figure : 3. Translating English sentences to Hindi sentences.

CONCLUSION

We proposed 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



English Word: sun

☑ Noun □ Pronoun □ Verb □ Adjective □ Adverb

Hindi Word: सूरज

Figure : 4. Adding English word to the Data Base.

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.

.

ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of this research work. We are thankful to Prof. R.R.Sedamkar for his constant guidance and support. We also thankful Dr. B.K.Mishra for providing us environment for research

REFERENCES

[1] Allen, J. (1994 a). Natural Language Understanding, 2nd ed. Benjamin/Cummings, Redwood City, California.Allen, J. (1994 a). Natural Language Understanding, 2nd ed. Benjamin/Cummings, Redwood City, California.

[2] Deepa Gupta, Niladri Chatterji, Identification of divergence for English to Hindi EBMT, 2010, in proceedings of Machine translation SUMMIT III. 141—148.

[3] Eiichiro Sumita Example-based machine translation using DP-matching between word sequences 2001, ATR Spoken Language Translation Research Lab.,Kyoto, Japan

[4] Allen, J. (1994 a). Natural Language Understanding, 2nd ed. Benjamin/Cummings, Redwood City, California.

[5] Allen, J. (1994 b). Linguistic Aspects of Speech Synthesis. In Voice Communication Between Humans and Machines (D. B. Roe, and J. G. Wilpon, eds.), pp. 135-155. National Academy of Sciences, Washington, District of Columbia.

[6] Alexandersson, J., Reithinger, N., and Maier, E. (1997). Insights into the Dialogue Processing of Verbomil. In Proceedings of Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Morgan Kaufmann, San Francisco, California, pp. 33-40.

[7] Nagao, 1984. A framework of a mechanical translation between Japanese and English by analogy principle, in Proceedings of the international NATO symposium on Artificial and human intelligence, 1984, pp. 173–180.

[8] R. D. Brown, 1993. Example – Based Machine Translation in Pangloss system ,International workshop on EBMT.

[9] Groves and Way, 2006. A memory based classification approached to marker based EBMT, Dublin, Ireland.

[10] J. D. Kim, 2010. Chunk based EBMT in annual conference of European Association for machine translation (EAMT, 2010).

[11] Gough, 2005. Example based controlled translation, Valetta, Malta, pp. 35-42.

[12] Sinha and A. Jain, 2002, A English to Hindi Machine Aided Translation System Based.

[13] L.K. Bharati, 1997.,Constraint Based Hybrid Approach to Parsing Indian Languages, IIIT Hydrabad.

[14] Viren Jain, 2003., A Smorgasbord of features for Statistical Machine Translation.

[15] Dave , 2010. Machine translation System in Indian Perspective, Lucknow. Journal of computer science 6(10) : 111-1116.

[16] Gangadharaiah and Balakrishnan, 2010., Generalised -EBMT, in 23rd COLING .

[17]] Brown, 2000. International workshop on EBMT. Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*,19(2):263–311

[18] Vogel and Ney, 2000. A New Approach for English to Chinese Named Entity Alignment

[19] Paul, 2003. The Mathematics of SMT : Parameter Estimation.

[20] Imamura, 2003 .EBMT Based on Syntactic Transfer with Statistical model.

[21] Koehn, 2006. Statistical Machine translation, Cambridge University Press. C. Callison – Brunch, M. Osborne, and P. Koehn, Reevaluating the role of BLEU in MT research in Proceedings of 11th Conference of EACL, 2006, pp. 249-256.

[22] Groves and Way, 2005, Hybrod example based SMT in ACL.

[23] Somers, 2007. International workshop on EBMT, Duplin, Ireland, pp. 53-60