A Survey of Machine Translation Techniques and Systems for Indian Languages

Paper Written By :-

- Sandeep Saini
- Vineet Sahula

Presented By : Riti Gupta

INTRODUCTION

- Machine Translation pertains to translation of one natural language to other by using automated computing.

MOTIVATION

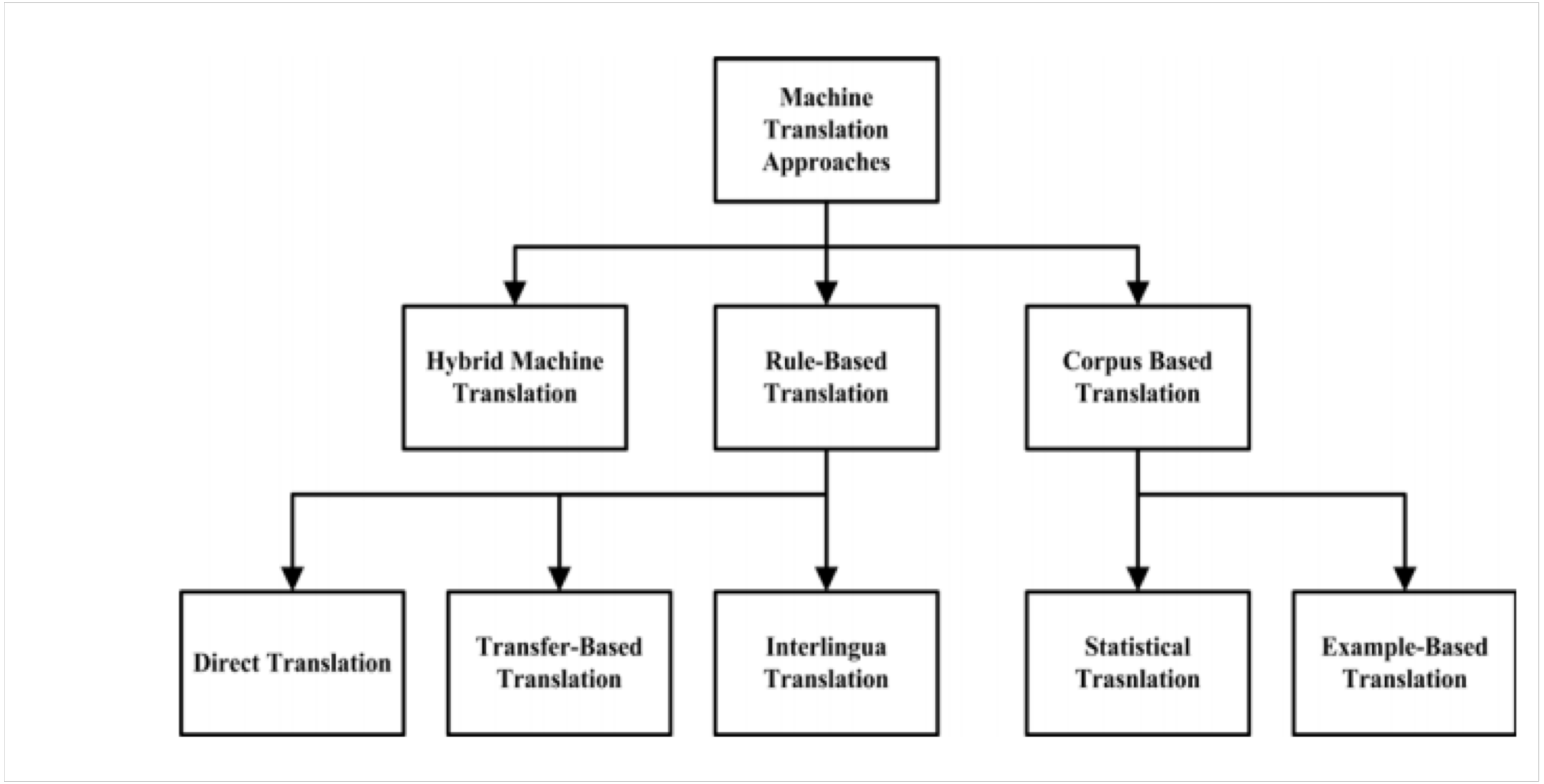
Avoiding digital divide : Most of the information available is in English understood by only 3% of the population

Translate literary works from any language into native languages

APPROACHES FOR MACHINE TRANSLATION(MT)

- Translation between structurally similar languages like Hindi and Punjabi is easier than between language pairs that have wide structural difference like Hindi and English.
- Fully automatic high quality machine translation system is difficult to build.
- Four Memories required for MT
  - knowledge of the language system
  - knowledge of language usage
  - knowledge of the world
  - knowledge of the situation
- Human translators use all four types of memory
- Machine Translation systems use some (but not all) of them.
  - Current MT systems all rely on grammatical and lexical competence. Rest is under investigation

CLASSIFICATION OF APPROACHES



A. RULE BASED TRANSLATION(RBMT)

- In rule-based systems, the source text is parsed and an intermediate representation is produced from which target language text is generated.
- The systems are based on linguistic information about source and target languages derived from grammars and dictionaries
- RBMT is further classified as depending on the intermediate representation
  - Direct Translation
  - Interlingua based Machine Translation
  - Transfer based Machine Translation

Direct Machine Translation:

- A direct word by word translation of the input source is carried out with the help of a bilingual dictionary and after which some syntactical rearrangement are made.
- Only one language pair is taken into consideration at a time.
- To get a target translation of any source sentence one would need:
  - A dictionary that will map each source word to an appropriate target word.
  - Rules representing regular source sentence structure.
  - Rules representing regular target sentence structure.

Direct Machine Translation

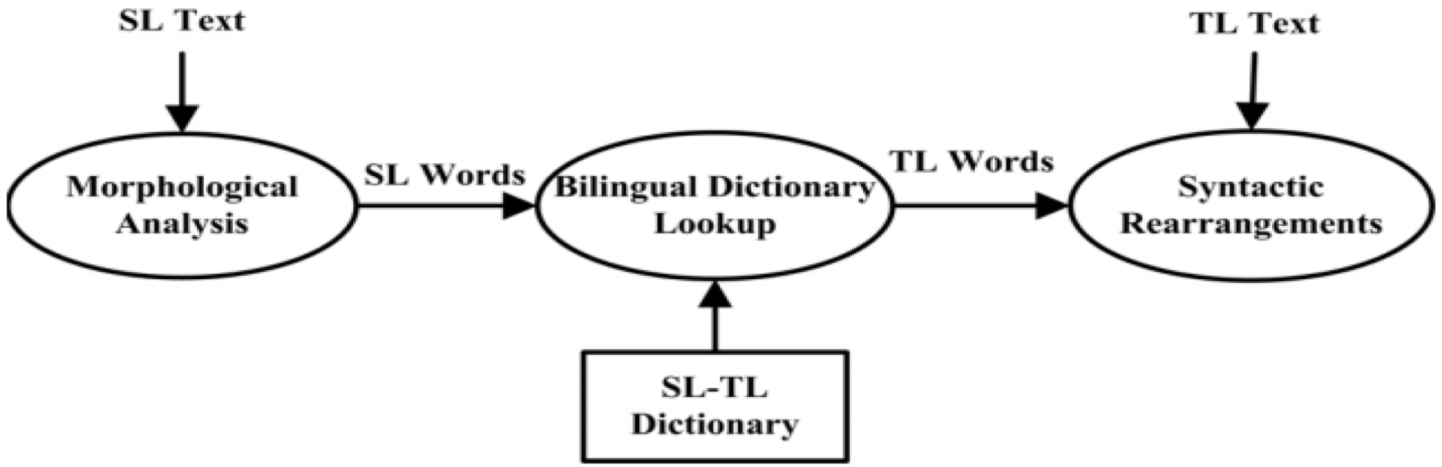


Fig. 2. Direct Machine Translation Approach

DMT PROS AND CONS

- PROS
  - Systems are easy to implement.

- CONS
  - Direct MT involves only lexical analysis. It does not consider structure and relationships between words.
  - Direct MT systems are developed for a specific language pair and cannot be adapted for different language pairs.

Interlingua Translation

- Source language is transformed into Interlingua, which is an intermediate abstract language-independent representation. The target language is generated from this Interlingua.
- This approach is more efficient than direct translation as it is not merely a dictionary mapping of two languages. In this approach, linguistic rules specific to the language pair are used.

Interlingua Translation

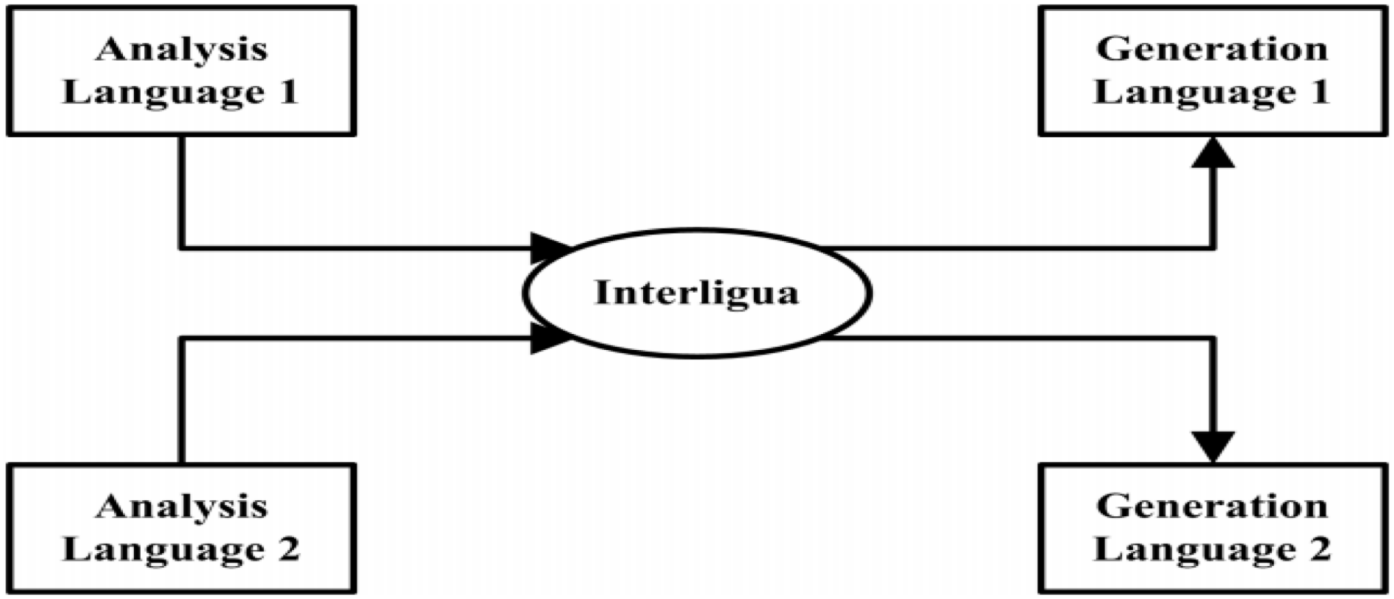


Fig. 3. Interlingua Translation [14]

Interlingua Translation : Pros and Cons

- PROS
  - An Interlingua system resolves all the ambiguities so that translation to any language can take place.
  - The system is more practical when several languages are to be interpreted

- CONS
  - Time efficiency is lower than Direct Machine Translation system.
  - Major problem lies in defining a universal abstract (Interlingua) representation which preserves the meaning of a sentence

Transfer Based Translation

- A database of translation rules is used to translate text from source to target language.
- Whenever a sentence matches one of the rules, or examples, it is translated directly using a dictionary. Source language dictionary, target language dictionary and a bilingual dictionary is used for this purpose.

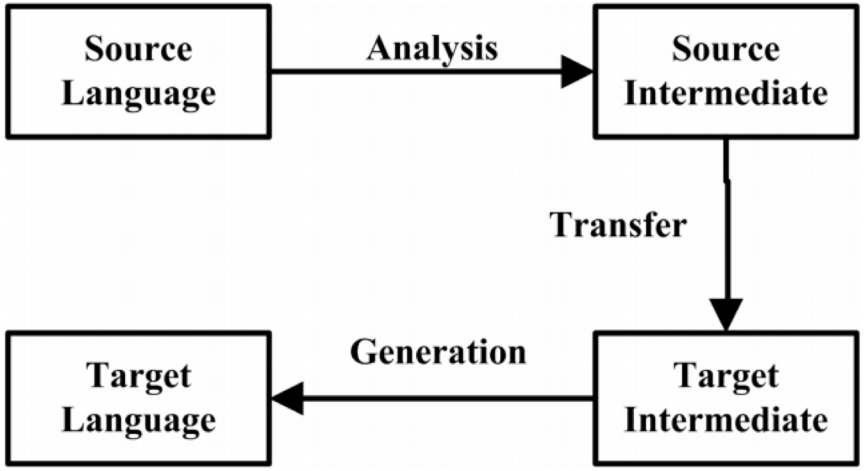


Fig. 4. Transfer Based Translation

TRANSFER BASED TRANSLATION: PROS AND CONS

- PROS
  - It has a modular structure.
  - The system easily handles ambiguities that carry over from one language to another.

- CONS
  - Some of the source text meaning can be lost in the translation

B. Corpus Based Translation

- Based on statistical analysis of source and target language corpus.
- Huge dataset is required for initial statistical analysis. This can be further classified as
  - Statistical Machine Translation (SMT)
  - Example Based Machine Translation system (EBMT)

Statistical Machine Translation

- In the first stage, many output sentence candidates in the target language are generated. The Language Model (LM) computes the probability of the target language T as probability P(T).
- In the second stage, The Translation Model (TM) computes the conditional probability of target sentences given the source sentence, P(T|S).
- Decoder maximizes the product of LM and TM probabilities.
  - $P(S, T) = \text{argmax} P(T) * P(S|T)$

Statistical Machine Translation

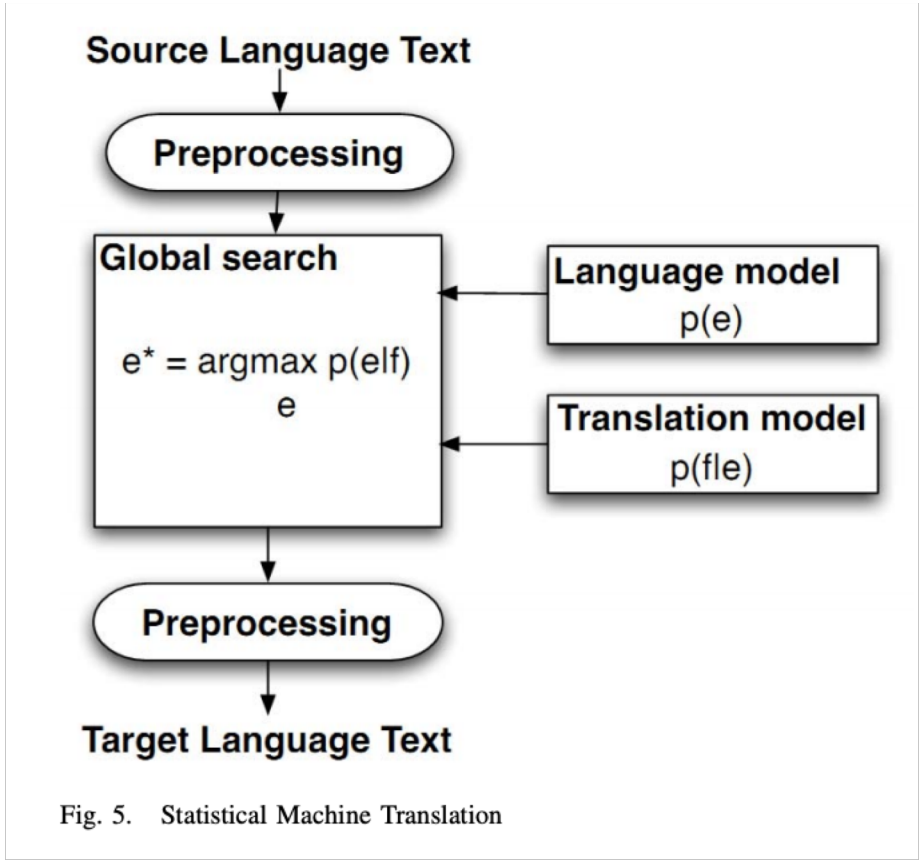


Fig. 5. Statistical Machine Translation

Statistical Machine Translation: Pros And Cons

- PROS
  - Not tailored to any specific pair of languages.
  - Rule-based translation systems require the manual development of linguistic rules, which can be costly and often do not generalize to other languages.
- CONS
  - Corpus creation can be costly for users with limited resources.
  - Statistical machine translation do not work well between languages that have significantly different word orders (e.g. Japanese and European languages).

Example based Machine Translation (EBMT)

- Maintains a corpus consisting of translation examples between source and target languages.
- An EBMT system has two modules
  - Retrieval module : retrieves a similar sentence and its translation from the corpus
  - Adaptation module : adapts the retrieved translation to get the final corrected translation.
- Example :
  - Consider the English to Hindi translation for the sentence, "*Rohan eats a Mango*".
  - Retrieval module selects "*Seema eats a Mango*" and its translation "*Seema aam khaati hai*" as the closest one.
  - Adaptation module replaces "*Seema" with "Rohan*" and "*khaati*" with "*khaata*" and finally forms the translation, "*Rohan aam khaata hai*".

Example based Machine Translation (EBMT)

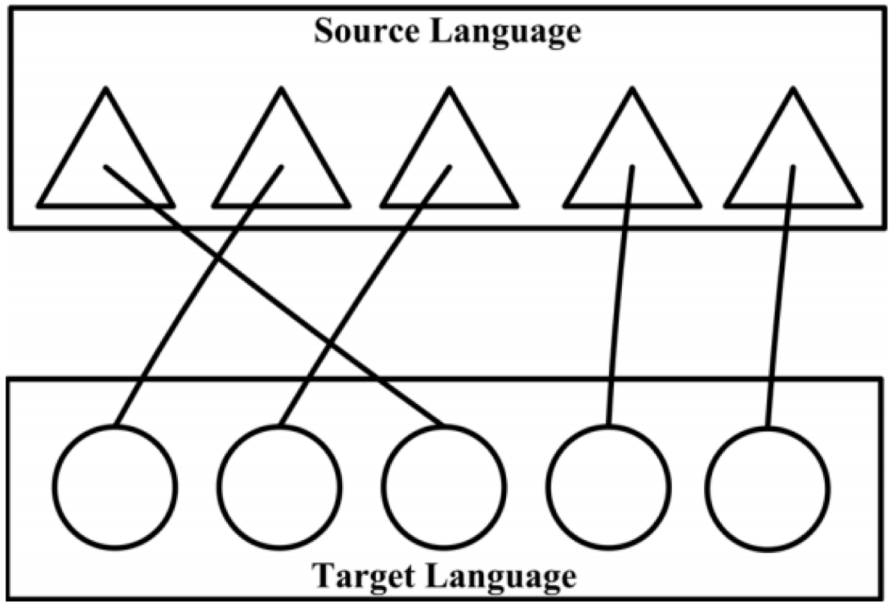


Fig. 6. Translation Template of a phrase in two different languages.

EBMT : PROS AND CONS

- A simple adaptation is required to replace the word and suffix replacements.
- This may not work in case of translation divergence where structurally similar sentences of the source language get translated into a different structure.

C. Hybrid Machine Translation

- In this technique, the objective is to get best possible translation by exploiting the features of various existing approaches.
  1) Statistical Rule Generation:
     1) Statistical data is used to generate lexical and syntactic rules.
     2) The input is processed with these rules as if it were a rule-based translator.
     3) Disadvantage: The accuracy of the translation would depend heavily on the similarity of the input text to the text of the training corpus. As a result, this technique has had the most success in domain-specific applications
  2) Multi-Engine MT:
     1) Combines the outputs of multiple MT engines using a statistical language model of the target language.
     2) Involves running multiple machine translation systems in parallel. The final output is generated by combining the output of all the sub-systems.

CONCLUSION

- Different formalisms best suited to the applications can be used.
- More research has to be done in these areas to overcome the language barrier being faced
- The MT systems so far developed have many shortcomings in terms of rule set, dictionary, translation methodology and further work is needed in MT to produce intelligible translations.