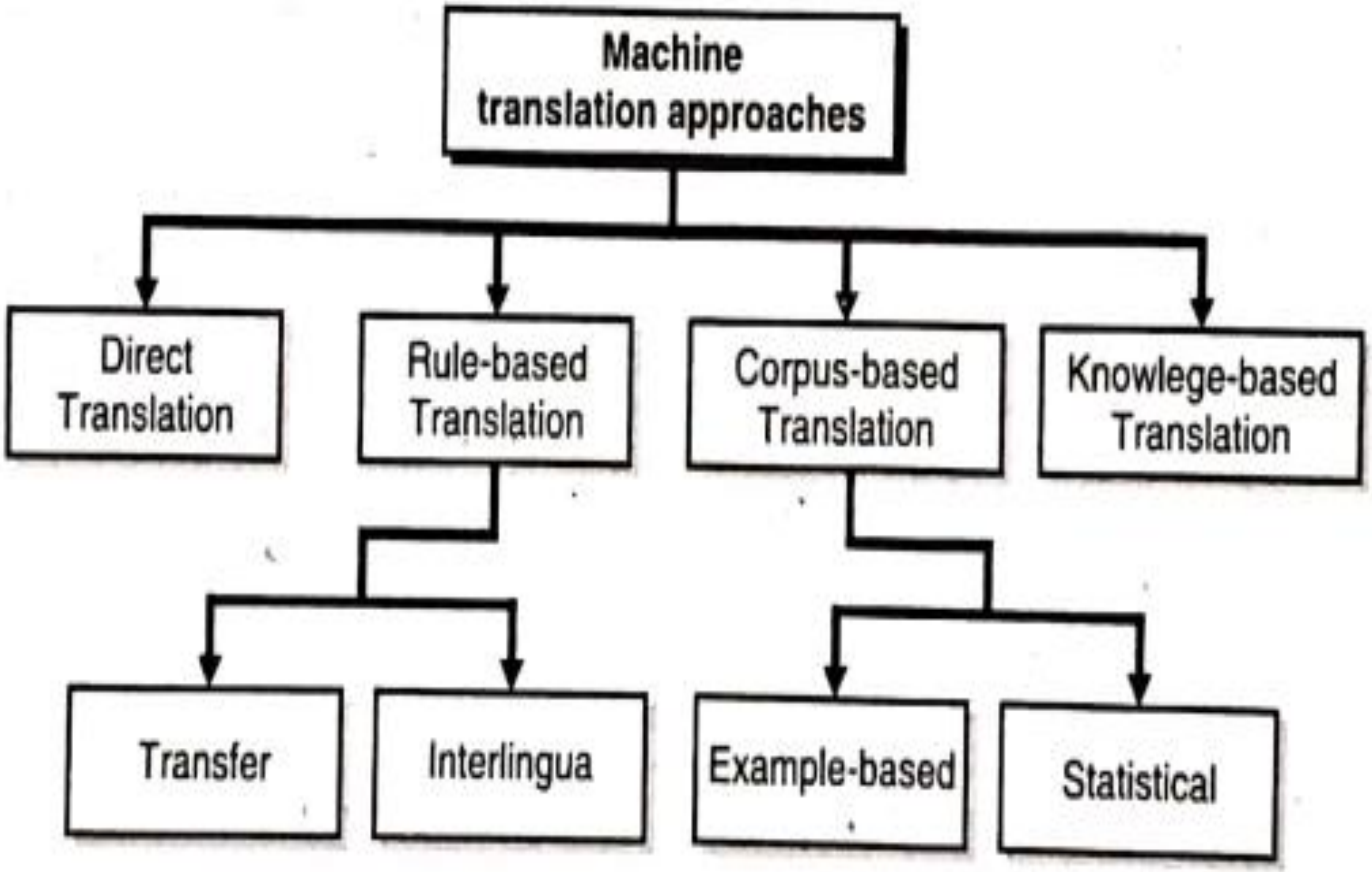
MODULE 6

APPLICATIONS OF NLP

1. MACHINE TRANSALATION

- Machine Translation (MT) is the process by which computer software translate text from one language into another without the assistance of a person.
- Machine Translation is the process of converting a text from one language to its equivalent in another.
- Challenges in MT
- 1. The wide range of alphabets, grammars and languages.
- 2. For computer, translating a sequence to a series is more difficult than dealing with numbers.
- 3. There is not just one right response.

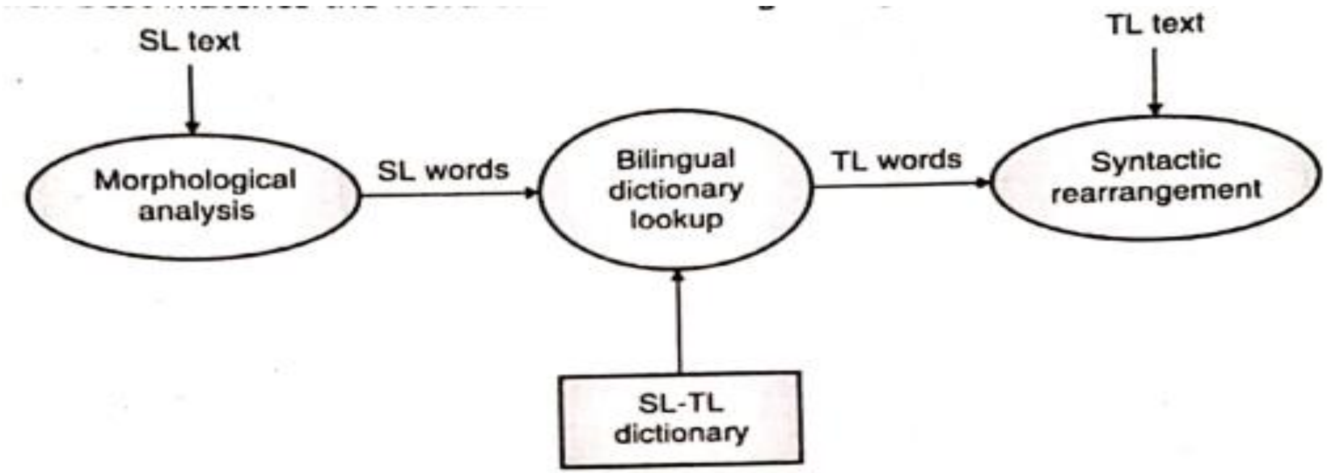
Types of Machine Translation



Direct Machine Translation

- Direct Machine Translation is one of the simplest machine translation approach.
- It translate the individual words in a sentence from one language to another using two-way dictionary (Bilingual dictionary).
- Direct MT has following characteristics:
- 1. Little analysis of source language.
- 2. No parsing
- 3. Depend on large two-way dictionary.

- The general procedure for DMT systems as shown in figure:



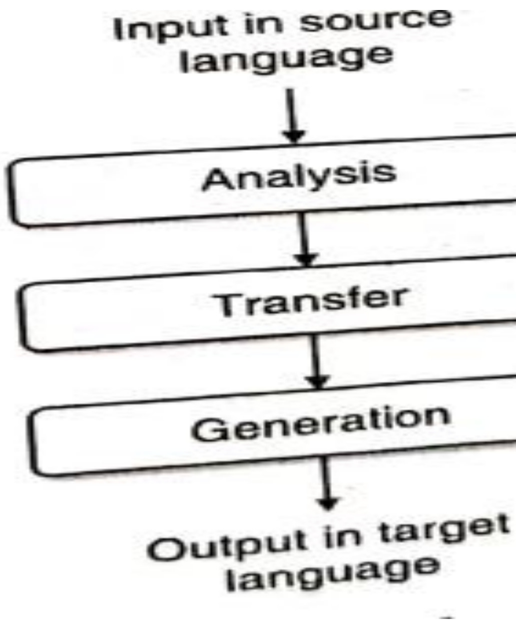
- 1. Morphological analysis removes morphological inflections from the words to get root form of the source language words.
- 2. A bilingual dictionary used to get the target language words corresponding to the source language words.
- 3. In syntactic rearrangement, the word order is changed to that which best matches the word order of the target language.

Rule-based Machine Translation

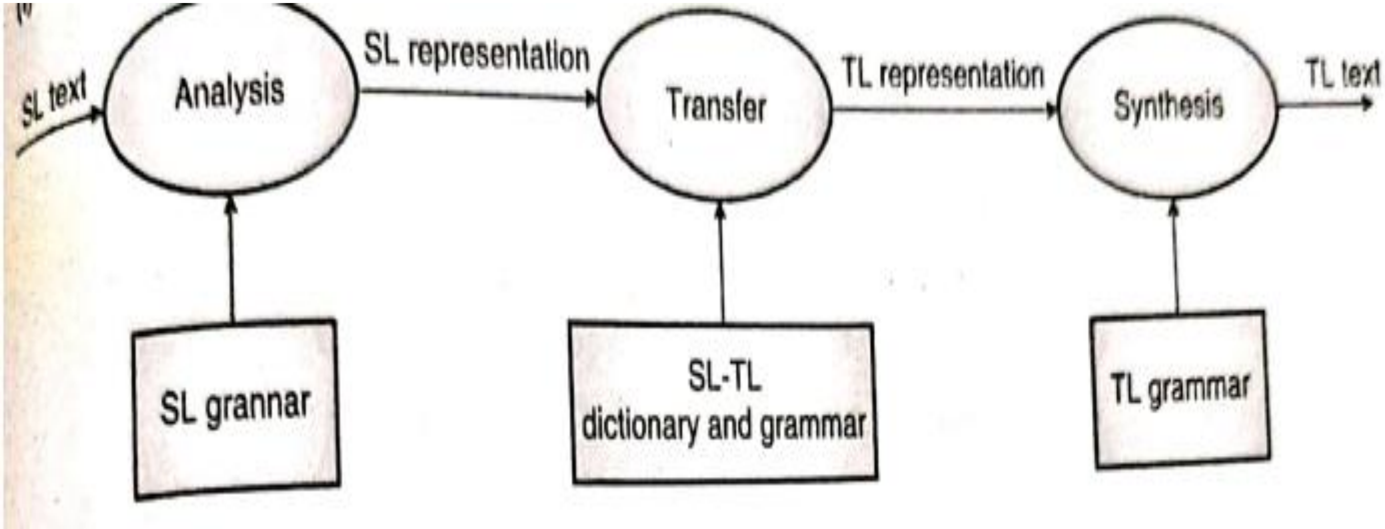
- RBMT(Rule-based machine translation) system consists of collection of various rules, called grammar rules, a bilingual lexicon or dictionary, and software program to process the rules.
- RBMT systems are depends on lexicons.
- Languages have to be added manually.
- It requires large mount of post-editing by humans.
- Rule Based systems are categorized as:
- 1. Transfer Based Machine Translation
- 2. Interlingua Machine Translation

1. Transfer Based Machine Translation

- Transfer based system can be broken down to three stages:
- 1. Analysis: It is used to produce source language structure.
- 2.Transfer: It is used to transfer source language representation to target level representation.
- 3. Generation: It is used to generate target language text using target level structure.



- General Model of Transfer Based Machine Translation is as Shown in figure:



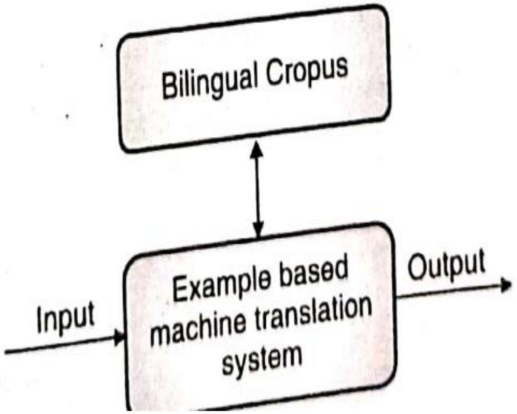
- This system use a database of translation rules to translate text from source to target language. Whenever a sentence matches one of the rules, it translated directly using a dictionary.
- It uses morphological and syntactic analysis to produce a sort of Interlingua on the base forms of the source language.
- From this it translates into base form of the target language.

2. Interlingua Based Translation

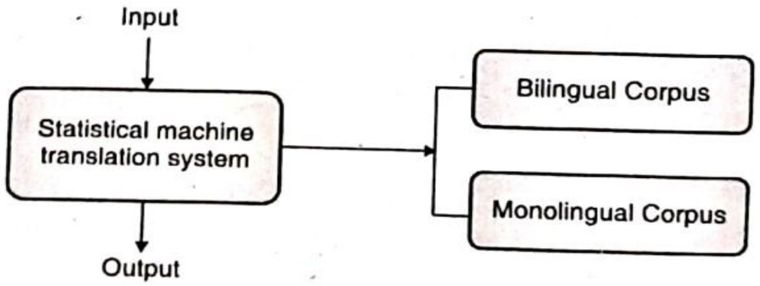
- In this approach, the translation consists of two stages, where the source language (SL) is first converted into the Interligua (IL) form.
- The main advantage of this approach is analyzer and parser of SL is independent of the generator for the Target language (TL).
- Then Interligua(IL) is converted into Target Language.

3. Corpus Based Approach

- The corpus based approach don’t required any explicit linguistic knowledge to translate the sentence, but a bilingual corpus of the source language and target language are required to train the system to translate a sentence.
- 1. Example Based Approach:
- It reused the examples of already existing translations.
- It uses bilingual corpus as its main knowledge base and it is essentially translation by analogy



- 2. Statistical Based Approach:
- Statistical machine translation is a data-oriented statistical framework which is based on knowledge and statistical models which are extracted from bilingual corpus.
- In this MT, Bilingual or multilingual corpora of the languages are required.
- In this a document is translated according to the probability distribution function which is indicated by p(w|h).
- Finding the best translation is done by picking the highest probability.



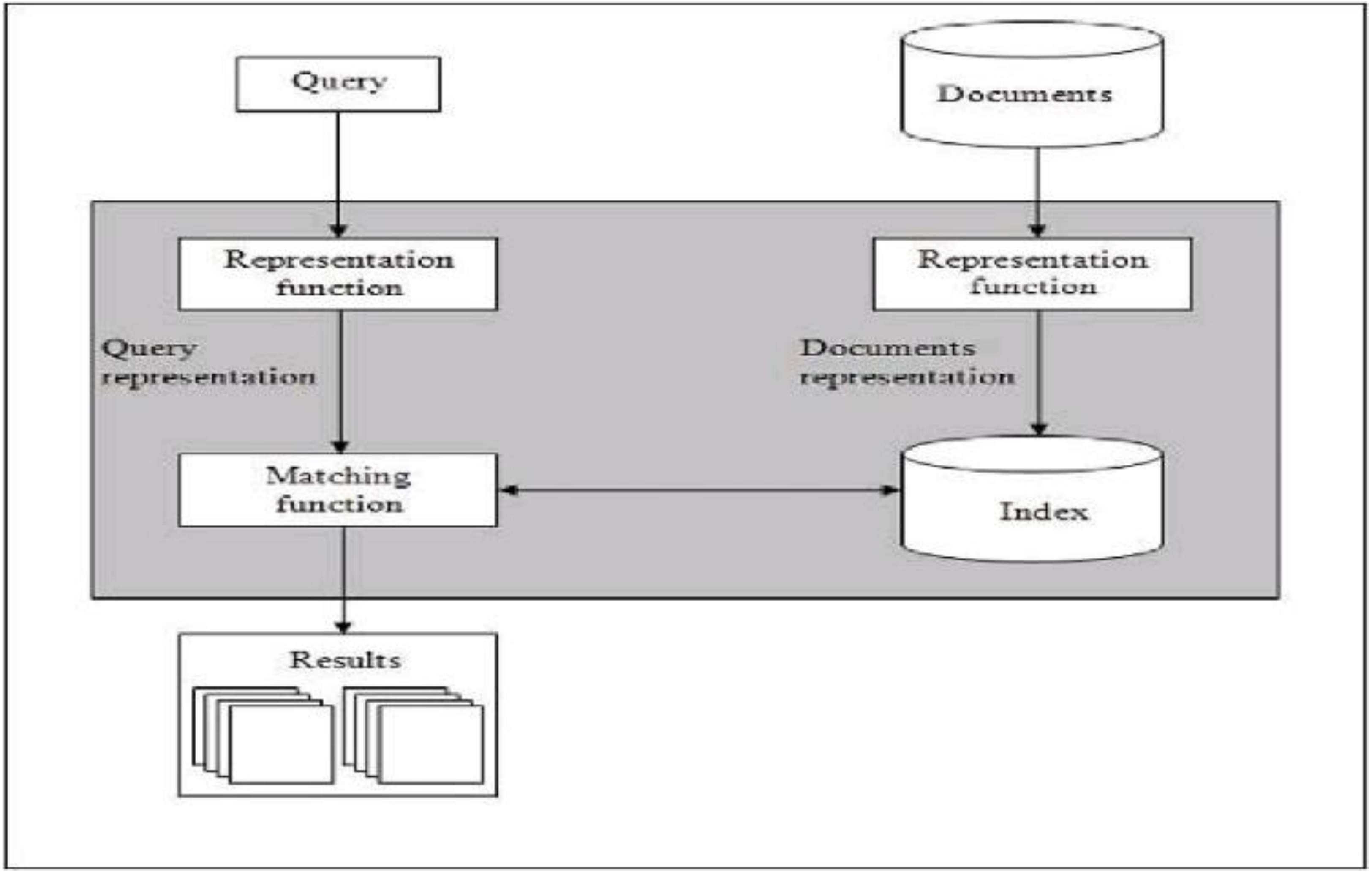
Algorithm of English to Marathi MT System

- Input: Digital Document as input in English Natural language.
- Output: Translated document in marathi Natural language.
- Steps:
- 1. Accept a digital document as input.
- 2. For each sentence in put document apply POS tagging.
- 3. Generate the parse tree for each sentence.
- 4. Apply NER on each sentence.
- 5. Identifies names and relations between a verb and noun in the sentence.
- 6. Use bilingual dictionary to obtain appropriate translation.
- 7. Obtain the proper form of words using inflections.
- 8. Represent sentences in target language.

Information Retrieval

- Information retrieval (IR) may be defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information.
- The system assists users in finding the information they require but it does not explicitly return the answers of the questions.
- It informs the existence and location of documents that might consist of the required information.
- The documents that satisfy user’s requirement are called relevant documents.
- A perfect IR system will retrieve only relevant documents.
- **Google Search** is the most famous example of information retrieval.

- Basic IR System:



- From the above diagram, it is clear that a user who needs information will have to formulate a request in the form of a query in natural language.
- After that, the IR system will return output by retrieving the relevant output, in the form of documents, about the required information.
- The step by step procedure of these systems are as follows:
- 1. Indexing the collection of documents.
- 2.Transforming the query in the same way as the document content is represented.
- 3.Comparing the description of each document with that of the query.
- 4. Listing the results in order of relevancy.

- Retrieval Systems consist of mainly two processes:
- Indexing
- Matching
- **Indexing**
- It is the process of selecting terms to represent a text which involves tokenization of string, removing frequent words and stemming.
- **Matching**
- It is the process of finding a measure of similarity between two text representations.
- Relevance of document is computed based on parameters:
- **1.** TF: It stands for Term Frequency which is simply the number of times a given term appears in that document.
- **TF (i, j) = (count of ith term in jth document)/(total terms in jth document)**

- **2.** IDF: It stands for Inverse Document Frequency which is a measure of the general importance of the term.
- **IDF (i) = (total no. of documents)/(no. of documents containing ith term)**
- **Classical Problem in IR Systems**
- **Ad-hoc retrieval problem** is the classical problem in IR systems.
- In ad-hoc retrieval, the user must have to enter a query in natural language that describes the required information. Then the IR system will return the output as the required documents that are related to the desired information.
- **For Example**, suppose we are searching for something on the Internet and it gives some exact pages that are relevant as per our requirement but there can be some non-relevant pages too.
- This is due to the ad-hoc retrieval problem.

- Information Retrieval (IR) Model
- A model of information retrieval predicts and explains what a user will find in relevance to the given query.
- IR model is basically a pattern that defines the retrieval procedure and consists of the following −
- A model for documents.
- A model for queries.
- A matching function that compares queries to documents.
- Mathematically, a retrieval model consists of −
- **D** − Representation for documents.
- **R** − Representation for queries.
- **F** − The modeling framework for D, Q along with relationship between them.
- **R (q,di)** − A similarity function which orders the documents with respect to the query. It is also called ranking.

- **Types of Information Retrieval (IR) Model**
- An information model (IR) model can be classified into the following three models −

1. Classical IR Model

- It is the simplest and easy to implement IR model.
- This model is based on mathematical knowledge that was easily recognized and understood as well.
- Boolean, Vector and Probabilistic are the three classical IR models.

2. Non-Classical IR Model

- It is completely opposite to classical IR model.
- Such kind of IR models are based on principles other than similarity, probability, Boolean operations.

- Information logic model, situation theory model and interaction models are the examples of non-classical IR model.

3. Alternative IR Model

- It is the enhancement of classical IR model making use of some specific techniques from some other fields.
- Cluster model and fuzzy model are the example of alternative IR model.