Natural Language Processing

6th Semester B.Tech. (CSE)

Course Code:BTCS-T-PE-052

Nayan Ranjan Paul

Department of CSE

Silicon Institute of Technology

Unit -V

Information Retrieval

- Information retrieval(IR) deals with the organization, storage, retrieval and evaluation of information from document repositories relevant to a user's query.
- A user in need of information formulates a request in the form of a query written in natural language.
- The IR system responds by retrieving the document that seems relevant to the query.
- 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.

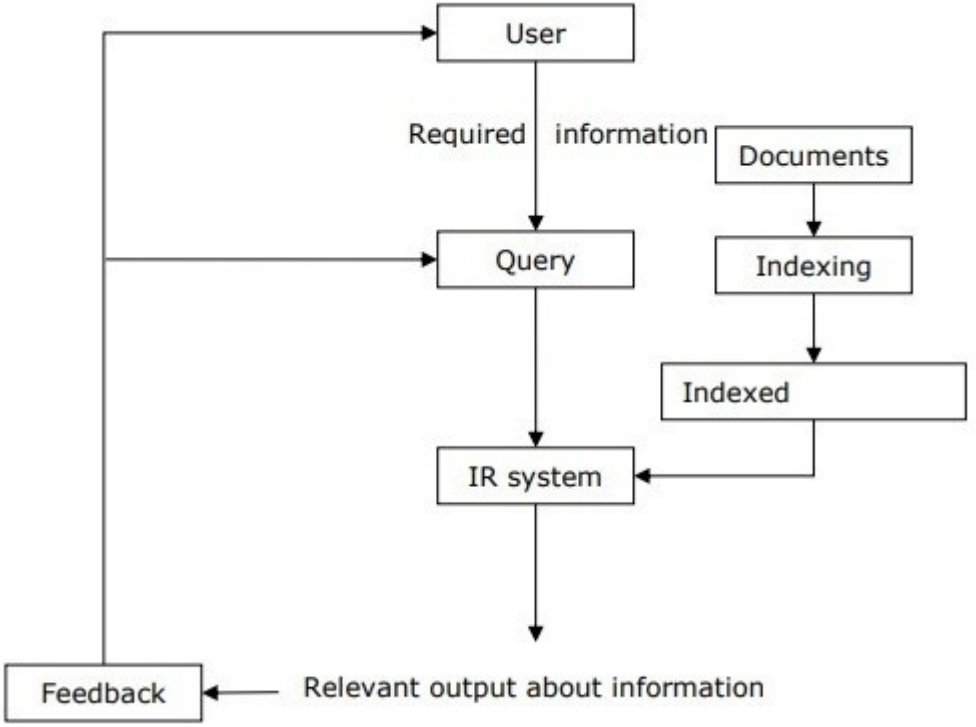
Information Extraction

- The main goal of IE is to extract meaningful information from document set.
- Here meaningful information contains types of information like events, facts, components, or relations.
- These facts are then usually stored automatically into a database
- which may then be used to analyze the data for trends, to give a natural language summary, or simply to serve for online access.
- More formally, Information Extraction gets facts out of documents while Information Retrieval gets sets of relevant documents.

Information Retrieval VS Information Extraction

Information Retrieval	Information Extraction
The goal is to find documents that are relevant to the user's information need	The goal is to extract pre-specified features from documents or display information.
Return set of relevant documents	Return facts out of documents
Real information is buried inside documents	Extract information from within the documents
Document Retrieval	Feature Retrieval
The long listing of documents	Aggregate over the entire set
Used in many search engines – Google is the best IR system for the web.	Used in database systems to enter extracted features automatically.

Process of Information Retrieval (IR)



Process of Information Retrieval (IR)

- The above figure illustrates the basic process of IR.
- It begins with the users information need.
- Based on this need, he/she formulates a query in natural language.
- The IR system returns documents that seem relevant to the query.

Process of Information Retrieval (IR)

- The basic question is:
  - What constitutes the information in the documents and queries.
  - This in turn is related to the problem of representation of documents and queries

- The retrieval is performed by matching the query representation with document representation

- The actual text of the document is not used in the retrieval process.

- Instead, documents in a collection are frequently represented through a set of index terms or keywords.

Process of Information Retrieval (IR)

- Keywords can be single word or multi-word phrases.
- The process of transforming document text to some representation of it is known as indexing.
- There are different types of index structures.
- One commonly used by IR system is the inverted index.
- An inverted index is simply a list of keywords, with each keyword carrying pointers to the documents containing that keyword.

Process of Information Retrieval (IR)

- The computational cost involved in adopting a full set of words to represent a document is high.
- Some text operations are usually performed to reduce the set of representative keywords.
- Two most commonly used text operations are
  - Stop word elimination
  - Stemming

Process of Information Retrieval (IR)

- Stop word elimination:
  - Stop words are high frequency words which have little semantic weight and are thus, unlikely to help in retrieval.
  - These words play important grammatical roles in language, but do not contribute to the semantic content of a document.
  - Typical example of stop words are articles and prepositions

Process of Information Retrieval (IR)

- Stemming:
  - Stemming normalizes morphological variants by removing affixes from the words to reduce them to their stem.
  - Eg: the words compute, computing, computes and computer, are be reduced to same word stem comput.

Information Retrieval (IR) Models

- The central objective of the model is to retrieve all documents relevant to a query.
- Main aspects of IR model:
  - How documents and user's queries are represented
  - How a system retrieves relevant documents according to user's queries
  - How retrieved documents are ranked
- The IR system consists of
  - a model for documents
  - A model for queries
  - A matching function which compares queries to documents
- Several different IR models have been developed

Information Retrieval (IR) Models

- These models can be classified as follows:
  - Classical model of IR
  - Non-classical models of IR
  - Alternative models of IR

Information Retrieval (IR) Models

- Classical model of IR
  - These models are on based mathematical knowledge that is easily recognized and well understood.
  - These models are simple, efficient, and easy to implement.
  - Almost all existing commercial systems are based on the mathematical models of IR.
  - That is why they are called classical models of IR
  - Three classical IR models are:
    - Boolean model
    - Probabilistic model
    - Vector space model

Information Retrieval (IR) Models

- Non-Classical model of IR
  - These models perform retrieval based on principles other than those used by classical models i.e similarity, probability, and Boolean operation.
  - These are the models based on
    - special logic technique,
    - situation theory, or
    - the concept if interaction

Issues in Information Retrieval

- The main issues of the Information Retrieval (IR) are
  - Document and Query Indexing,
  - Query Evaluation, and
  - System Evaluation
- Document and Query Indexing
  - Main goal of Document and Query Indexing is to find important meanings and creating an internal representation.
  - The factors to be considered are accuracy to represent semantics, exhaustiveness, and facility for a computer to manipulate.

Issues in Information Retrieval

- Query Evaluation
  - In the retrieval model how can a document be represented with the selected keywords
  - How are documents and query representations compared to calculate a score.
  - Information Retrieval (IR) deals with issues like uncertainty and vagueness in information systems.
  - Uncertainty :
    - The available representation does not typically reflect true semantics of objects such as images, videos etc.
  - Vagueness :
    - The information that the user requires lacks clarity, is only vaguely expressed in a query, feedback or user action.

Issues in Information Retrieval

- System Evaluation :
  - System Evaluation tells about the importance of determining the impact of information given on user achievement.
  - Here, we see if the efficiency of the particular system related to time and space.

Evaluation of Information Retrieval System

- The evaluation of IR systems is the process of assessing how well a system meets the information needs of its users.
- IR evaluation models can be broadly classified as
  - System driven models
    - It measures how well a system ranks documents.
  - User-centered models
    - It measures the user’s satisfaction.

Evaluation of Information Retrieval System

- Cleverdon listed the following six criteria that can be used for evaluation:
  - Coverage of the collection: the extent to which the system is covering
  - Time lag: the time that elapses between submission of a query and getting back the response.
  - Presentation format
  - User effort: the effort made by the user to obtain relevant information
  - Precision: the proportion of retrieved documents that are relevant
  - Recall: the proportion of relevant documents that are retrieved
- Of these criteria, recall and precision have most frequently been applied in measuring IR

Machine Translation

- Machine translation(MT) is the automatic translation of text from one language to another using computers.

- It is the earliest application of NLP

- The goal of achieving error-free translation that reads fluently in target languages is still far off; limited success has been achieved within a restricted domain.

- Example:
  - The well-known METEO system automatically translates hundreds of weather bulletins every day with 95% accuracy.

Problems in Machine Translation

- Achieving high quality translation is difficult.
- There are many structural and stylistic differences among languages, which make automatic translation difficult.
- Other factors are discussed below
- Word order
  - The arrangement of words in a sentence varies across languages.
  - Eg: English words are arranged in the order subject,verb, and object(SVO)

    Indian languages, the arrangement is subject, object and verb(SOV)
  - This makes a word by word translation impractical

Problems in Machine Translation

- Word sense
  - The sense of a word in one language may translate into a different sense with the words of another language.
  - This creates problem in target language word selection.
- Pronoun resolution
  - Resolving pronominal references is important for machine translation.
  - Unresolved references may lead to incorrect translation
- Idioms
  - A sentence involving idiomatic expressions is difficult to translate as idioms are composed of words that do not directly contribute to their meaning.

Machine Translation Approaches

- Machine translation approaches can be broadly classified into four categories:
  - Direct machine translation
  - Rule-based translation
  - Corpus-based translation
  - Knowledge-based translation

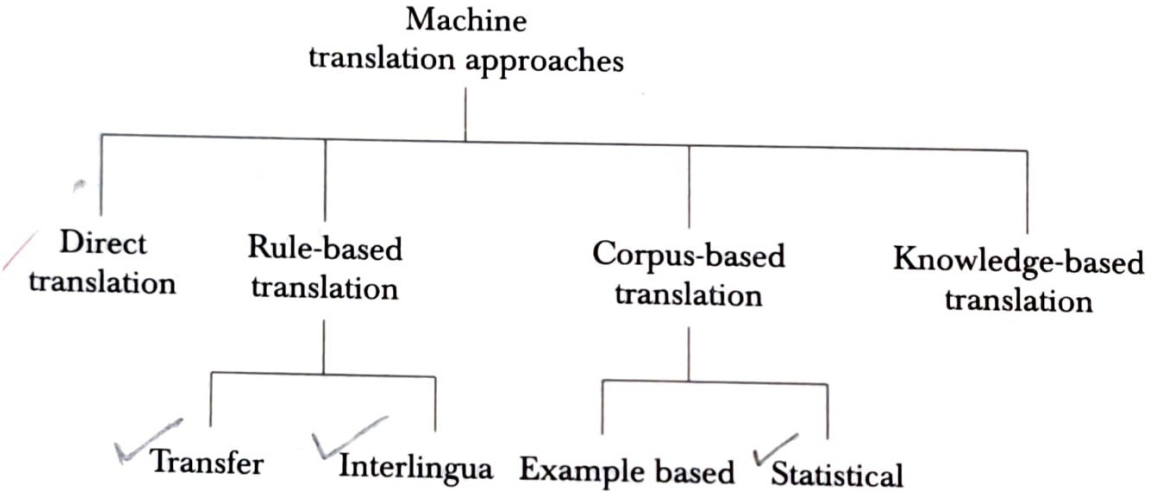


Figure 8.1 Machine translation approaches

Direct Translation Approaches

- Direct machine translation system provide direct translation i.e no intermediate representation is used.

- A direct translation system carries out word-by-word translation with the help of a bilingual dictionary, and usually followed by some syntactic rearrangement.

- These systems based on principle that an *"MT system should do as little work as possible"*.

Direct Translation Approaches

- It uses monolithic approach towards development- considers all the details of one language pair only.

- Besides dictionary translation, the analysis performed in this approach includes
  - Morphological analysis
  - Preposition handling
  - Syntactic arrangement
  - Morphological generation

Direct Translation Approaches

- The procedure for direct translation(English to Hindi) system can be summarized in following steps:
  - Remove morphological inflection from the words to get the root form of the source language(English) words
  - Look up a bilingual dictionary to get the target language words corresponding the source-language words.
  - Change the word order to that which best matches the word order of the target language.
    - Like in English-Hindi translation system, this would involve changing prepositions to post-positions and changing the subject-verb-object to subject-object-verb

Direct Translation Approaches

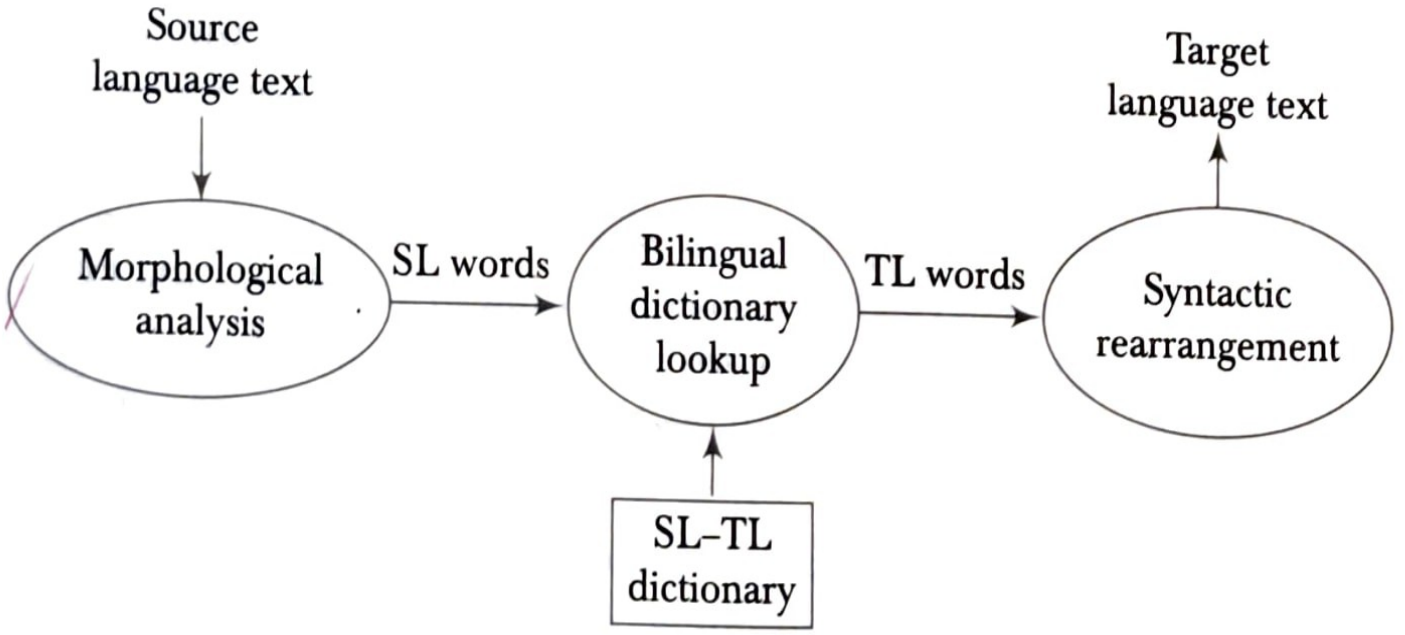


Figure 8.2 Direct machine translation systems

Direct Translation Approaches

- Example
  - Consider the English sentence:

    Khushbu slept in the garden.
  - A direct translation system will first look up a dictionary to get the target words appearing in the source-language.
    - Word translation: खुशबु सोयी में बाग
  - Then , the words are re-ordered to match the default sentence structure(SOV) of Hindi.
    - Syntactic arrangement: खुशबु बाग में सोयी

Direct Translation Approaches

- Example
  - Consider the English sentence:

    The boy gave the girl a book.

    - Word translation: लड़का दी लड़की एक किताब
    - Syntactic arrangement: लड़का लड़की एक किताब दी
    - Karaka handling and idiomatization: लड़के ने लड़की को एक किताब दी
  - In the above example we need to change the hindi word *ladka* to *ladke* simply to match it to the way Hindi is naturally used. This is termed as idiomatization

Disadvantages of Direct Translation Approaches

- It does not consider the structure and relationship between words, it does not attempt to disambiguate words, hence the quality of output is often not very good.

- A direct MT system is developed for a specific language pair and cannot be adapted for a different pair.

- It is quite expensive in a multilingual scenario, for example, in order to provide translation capability for n number of languages, we need to develop n(n-1) MT systems.

Rule Based Machine Translation

- Rule-based MT systems parse the source text and produce an intermediate representation.
- The intermediate representation may be a parse tree or some abstract representation.
- The target language text is generated from the intermediate representation.
- These systems make use of extensive lexicons equipped with morphological, syntactic and semantic information, and a large set of rules to map the input text to intermediate representation.

Rule Based Machine Translation

- Example of rule-based include the Ariane and SUSY systems.

- Depending on the intermediate representation used, these systems are further categorized as follows:
  - Transfer-based machine translation
  - Interlingua machine translation

Transfer-based Machine Translation

- These models transform the structure of the input to produce a representation that matches the rules of the target language.
- In order to get the structure of the input, some form of parse is needed.
- Therefore, transfer-based translation involves parsing of the source text.
- The source language parse structure is the transferred to target language structure
- The the target text is generated from the target language structure.

Transfer-based Machine Translation

- A transfer-based machine translation system has following component:
  - Analysis – To produce source language structure
  - Transfer – To transfer the source language representation to a target lavel representation
  - Generation – To generate target language text using target level structure.

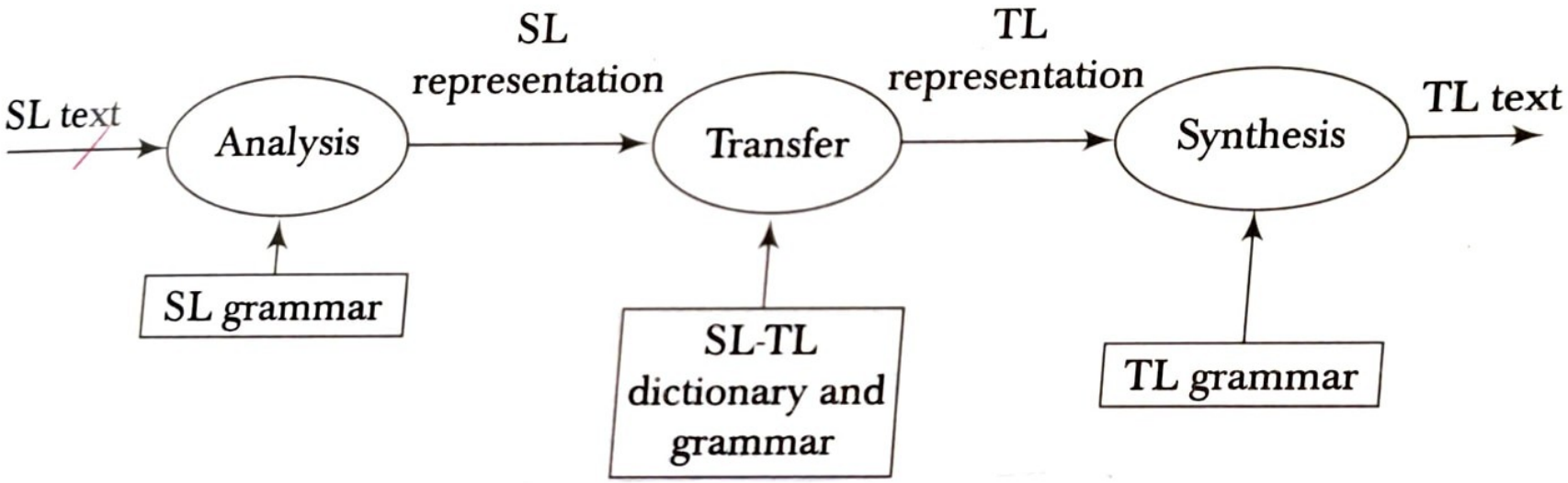


Figure 8.3 *Schematic diagram of the transfer-based model*

Transfer-based Machine Translation

- Analysis –
  - It analyses the source text and produces a structure confirming the rules of the source language.
  - It may involve morphological, syntactic, and semantic analyses.
  - This stage involves parsing
- Transfer –
  - The second stage transfers source language representation into target language representation
  - All the language -pair specific problems are handled by the transfer component.
- Generation –
  - The third stage is responsible for generating the actual target language text

Transfer-based Machine Translation

- Example –

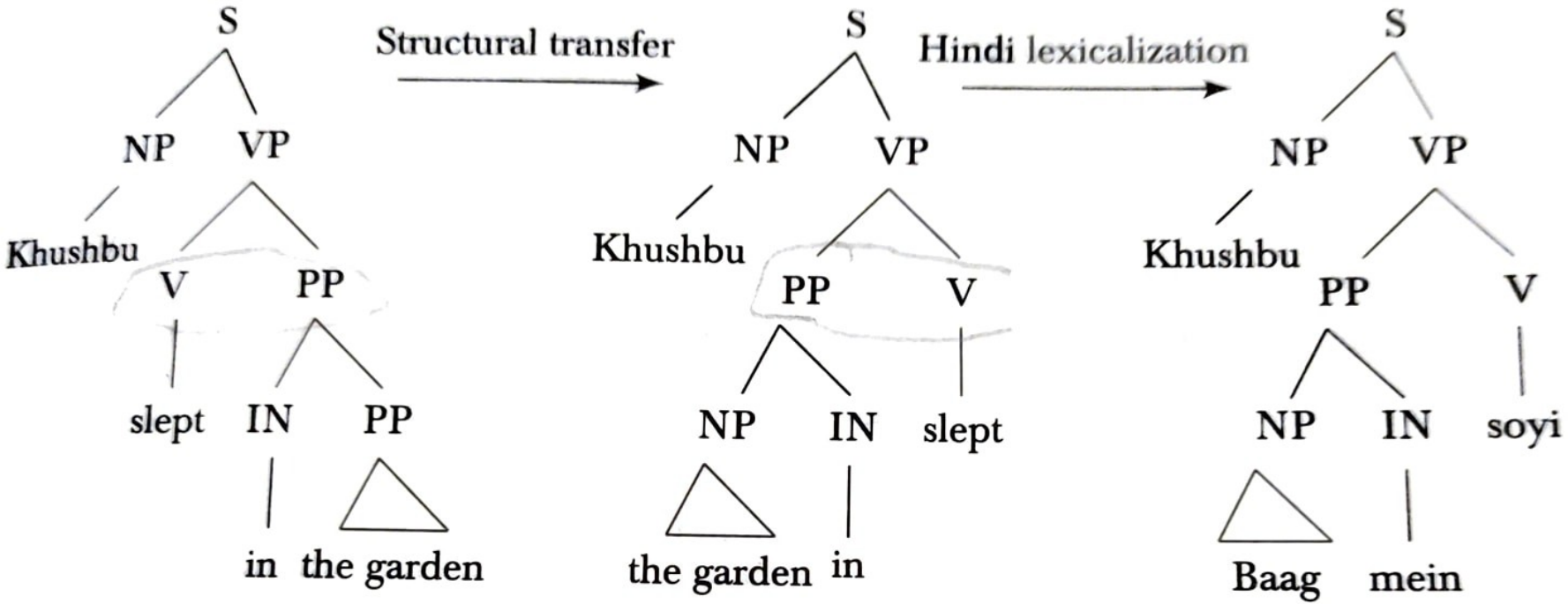


Figure 8.4 Structural transfer of sentence (8.1) from Hindi to English

Transfer-based Machine Translation

- Advantage
  - The main advantage of this approach is its modular structure.
  - The analysis of source language text (parser) is independent of target language generator
  - The transfer systems can easily handle ambiguities that carry over from ane language to another.

Interlingua-based Machine Translation

- In this approach, the source language text is converted into a language independent meaning representation called ‘interlingua’.
- An interlingua represents all sentences that mean the same thing in the same way regardless of the source language they happen to be in.
- From interlingua representation, text are generated into other language.
- Here translation is a two stage process
  - Analysis
  - Synthesis

Interlingua-based Machine Translation

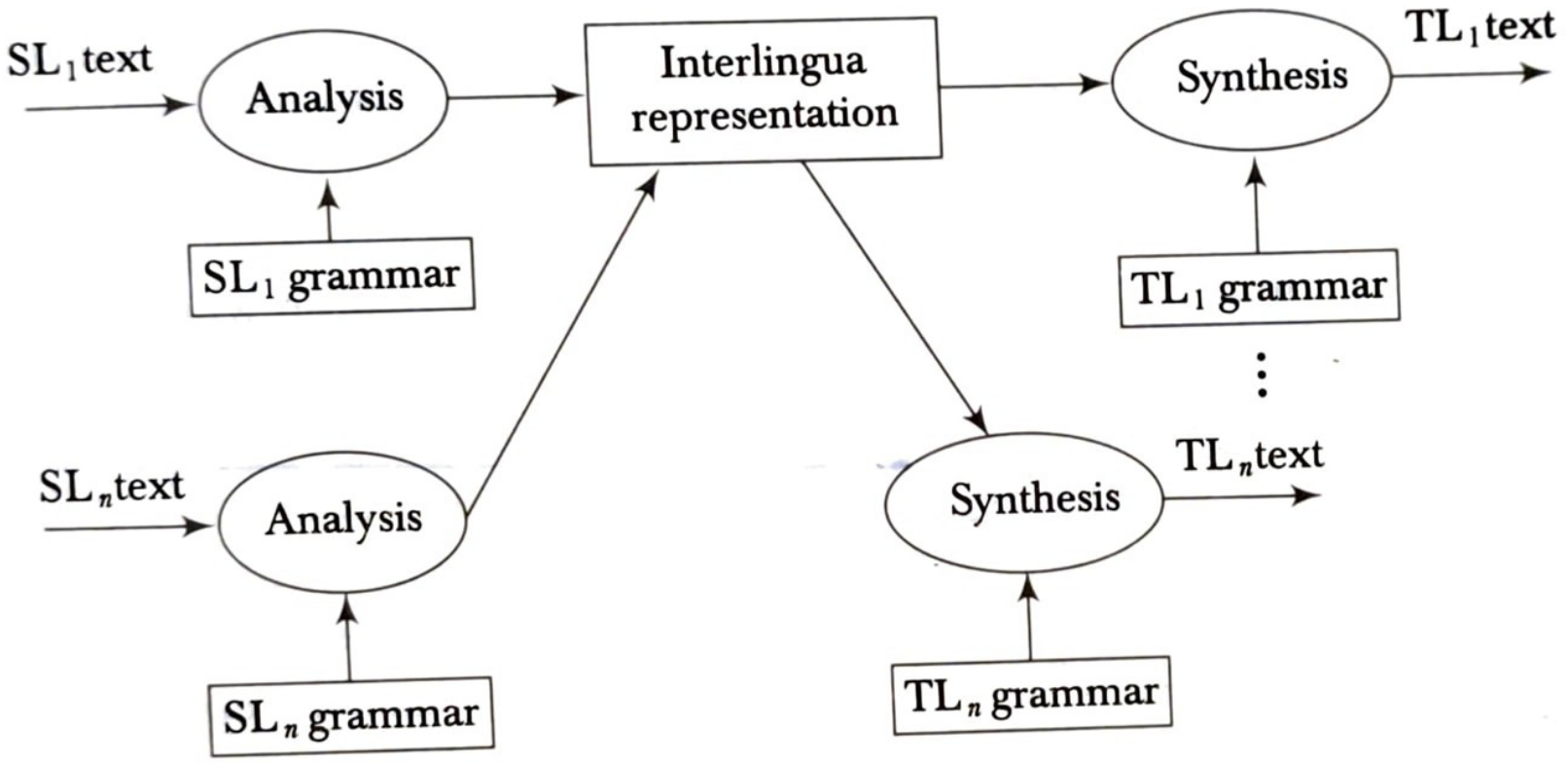


Figure 8.5 Interlingua-based translation model

Interlingua-based Machine Translation

- Analysis
  - In the first stage, source language text is represented in interlingua.
  - The analysis phase is specific to the source language text

- Synthesis
  - In the second stage, target language text is generated from the interlingua.
  - The synthesis phase is specific to the target language.

- It uses a semantic analyser as well as a syntactic parser to resolve all ambiguities.

Interlingua-based Machine Translation

- Example:
  - Consider the following English sentence

    Khushbu slept in the garden
  - The interlingua representation of the above sentence is

```
(*sleep
(tense past)
(mood declarative)
(punctuation period)
(subject (* Khushbu
         (number singular))
(Location (* garden
                  (reference definite)
                  (number singular)))
```

Figure 8.6 Structure of sentence (8.1) in interlingua

Advantages Interlingua-based Machine Translation

- This MT system can be used in multilingual environment.
  - The same analysis component can be used for more than one target language.

- To build a multilingual translation capability among n number of languages, it need only n analysis and n synthesis component as opposed to n(n-1) complete MT system in direct translation approach.

- The interlingua is a meaning-based representation and can be used in application like information retrieval.

Difficulty in Interlingua-based Machine Translation

- The major source of difficulty lies in defining a universal abstract interlingua representation which preserves the meaning of a sentence.
  - Because different languages conceptualize the world in different ways, defining a vocabulary for a universal interlingua is extremely difficult.

Corpus-based Machine Translation

- Corpus-based MT systems have gained much interest in recent years.
- These systems are fully automatic and require significantly less human labour than rule-based MT systems.
- These systems require sentence-aligned parallel text for each language pair.
- It is classified as:
  - Statistical MT
  - Example-based MT

Statistical Machine Translation

- Statistical MT systems are inspired by noisy channel model used in speech recognition systems.
- A noisy channel introduces noise that makes it difficult to recognize the input word.
- A recognition system based on it builds a model of the channel to identify how it modifies the input and recover the original word.
- The following figure shows the noisy channel model for English to Hindi translation.

Statistical Machine Translation

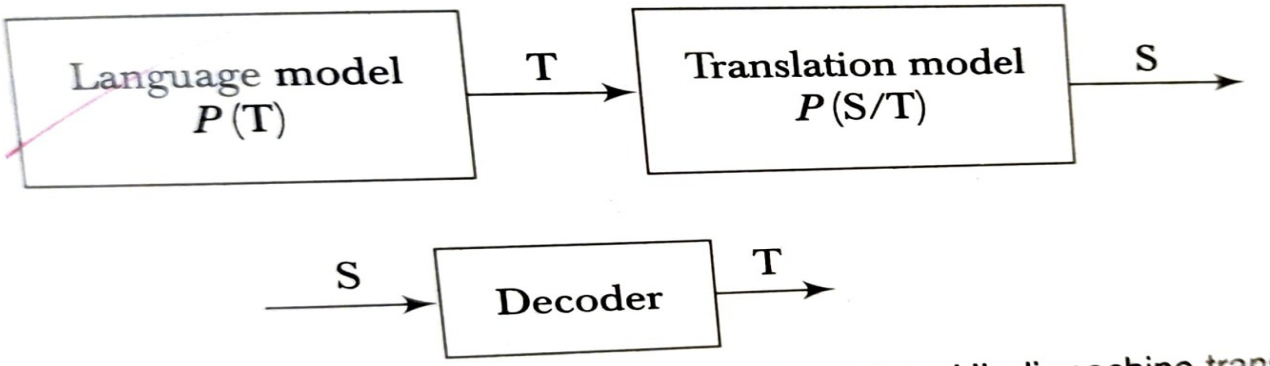


Figure 8.7 Noisy channel model for English to Hindi machine translation

- The input is considered as the distorted version of the target language sentence.
- The task is to find the most likely source language sentence, given the translation.
- This system models a target language sentence T, given a source language sentence S as the product of translation probability P(S|T) and the target language probability P(T).
  - P(S|T) – Accounts for the adequacy of translation contents
  - P(T) – Accounts for fluency of target construction.

Example-based Machine Translation

- Example-based machine translation (EBMT) system use past translation examples to generate translations for a given input.
- It is also called translation by analogy.
- An example-based MT system maintains an example-base consisting of translation examples between source and target languages.
- The EMBT systems rest on the idea that similar sentences will have similar translations.

Example-based Machine Translation

- When an input sentence is presented to it, the system retrieves a similar source language(SL) sentence from the example-base and its translation.

- It then adapts the example translation to generate the translation of the input sentence.

Example-based Machine Translation

- The EMBT system has two main modules
  - Retrieval
  - Adaptation

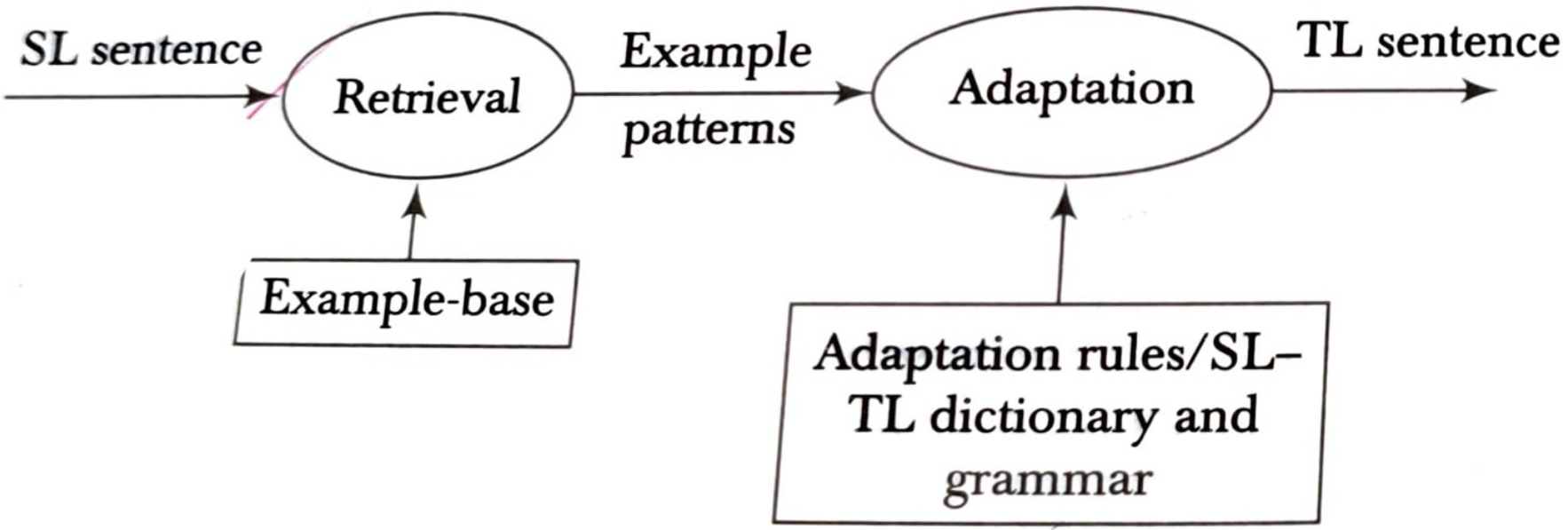


Figure 8.10 Example-based machine translation

Example-based Machine Translation

- Retrieval
  - The task of this module is to retrieve translation examples from the example-base for a given input.
  - It attempts to retrieve an example from the base which is similar to the input sentence.
  - It uses a similarity measure which can be based on word similarity or syntactic and semantic similarity.

Example-based Machine Translation

- Adaptation
  - This module is responsible for carrying out necessary modifications in the retrieved example pair to generate the translation of target sentence.
  - The modification may involve addition, deletion, and replacement of morphological words, constituent words or suffixes.
    - Addition – it involves insertion of a new chunk from the retrieved translation example
    - Deletion – it involves deletion of some chunk from the retrieved example pair.
    - Replacement – it involves replacement of some chunk in the retrieved example with some new chunk.

Example-based Machine Translation

- Example:
  - Consider the English-Hindi translation for the following input sentence:
    - Sheela sings a song.
  - Let the example base contains following sentences:
    - Rohit sings a song.
    - Sheela is playing
    - Sheela is singing a song
    - Sheela sings a Oriya song.
  - Let the retrieval component identify the example pair as similar to input:
    - Rohit sings a song.
    - Rohit gana gata hai

Example-based Machine Translation

- Example contd. :
  - Then the adaptation component generates the translation of input sentence by replacing Rohit with Sheela and taa with tii in the target language translation.
  - In this case the adaption operations used are word and suffix replacement.
  - The target sentence will be: Sheela gaana gaati hai.

Knowledge-based Machine Translation

- It is also called semantic-based MT.
- This approach requires a large knowledge-base that includes both ontological and lexical knowledge.
- The basic AI approach include:
  - Semantic parsing
  - Lexical decomposition into semantic networks
  - Resolution of ambiguity and uncertainties
- An example of this type of system is the KANT(Knowledge-based, accurate natural language translation) system.

Text Summarization

- **Text summarization** is the process of distilling the most important information from a text to produce an shorter version for a particular task and user.
- Important kinds of summaries:
  - **Outlines** of any document
  - **Abstracts** of a scientific article
  - **Headlines** of a news article
  - **Snippets** summarizing a web page on a search engine results page
  - **Action items or other summaries** of a (spoken) business meeting
  - **Summaries** of email threads

Text Summarization

- A summary can be loosely defined as
  - a text that is produced from one or more tools,
  - that conveys important information in the original text(s) and
  - that is no longer than half of the original text(s) and
  - usually significantly less than that.

- The goal of automatic summarization is
  - to take an information source,
  - extract content from it and
  - provide the most important content to the user in a condensed form, in a manner sensible to user's or application's needs.

Text Summarization

- Types of summaries
  - Single document vs multi-document
  - Generic vs user-oriented
  - Indicative vs informative
  - Extracts vs abstracts

Text Summarization

- Single document vs multi-document
  - Based on the number of documents processed during summarization, a summary can be a single document or a multiple document sumary.
  - A single document summary is produced using information from a single document whereas a multiple document summary is produced by combining information from multiple documents.
- Generic vs user-oriented
  - A generic summary presents the author’s viewpoint on the document.
  - It considers all the information in the document to create a summary
  - A user-oriented summary on the other hand, considers only that information which is deemed relevant to a user query

Text Summarization

- Indicative vs informative
  - According to the function that a summary serves, it can be classified as indicative or informative.
  - An indicative summary merely points out what topics are addressed in the original document without giving any content.
  - It can be used to alert the user to the source document so that the user can decide which of the original documents to read.
  - The indicative summary does not, in any way, substitute the source document.
  - The informative summary on the other hand, provides a short hand version of the content as far as coverage of information is concerned.

Text Summarization

- Extracts vs abstracts
  - An extractive summary is formed by selecting (extracting) phrases or sentences from the document to be summarized.
  - The extracted text units can be included in the summary verbatim, or they can be processed further to smooth the text flow.
  - An abstractive summary, on the other hand, involves the identification of salient concepts in the source document and uses different words to describe the contents of the document.
  - Most current summarizers are extractive as it is easier.

Text Summarization approaches

- There are two basic approaches to text summarization
  - Shallow or knowledge poor approach
  - Deep or knowledge-rich approach
- Shallow or knowledge poor approach
  - It usually involves only syntactic level processing.
  - This approach typically generates extracts by extracting sentences directly from the source text using statistical analysis.
  - Post-processing is required to smooth out incoherence caused by extraction and make it more compact.

Text Summarization approaches

- Deep or knowledge rich approach
  - This approach involves semantic level and discourse level analysis, usually to produce abstracts.
  - Summary generation consists of
    - Identifying the most important information in the document,
    - Encoding it appropriately, then
    - Feeding it to a natural language generation(NLG) system which generates the summary.

Question Answering System

- Given a collection of documents and a natural language question posed by a user, a question-answering system attempts to find the precise answer or at least the precise portion of text in which the answer appears.

- The main difference between IR system and question answering system is that in IR system the entire document that seems to contain information relevant to the user is returned whereas in question-answering system the precise answer or precise portion of text in which answer appears is returned.

Question Answering System

- A question-answering system is different from an information extraction system in that the type of information to be extracted is unknown.

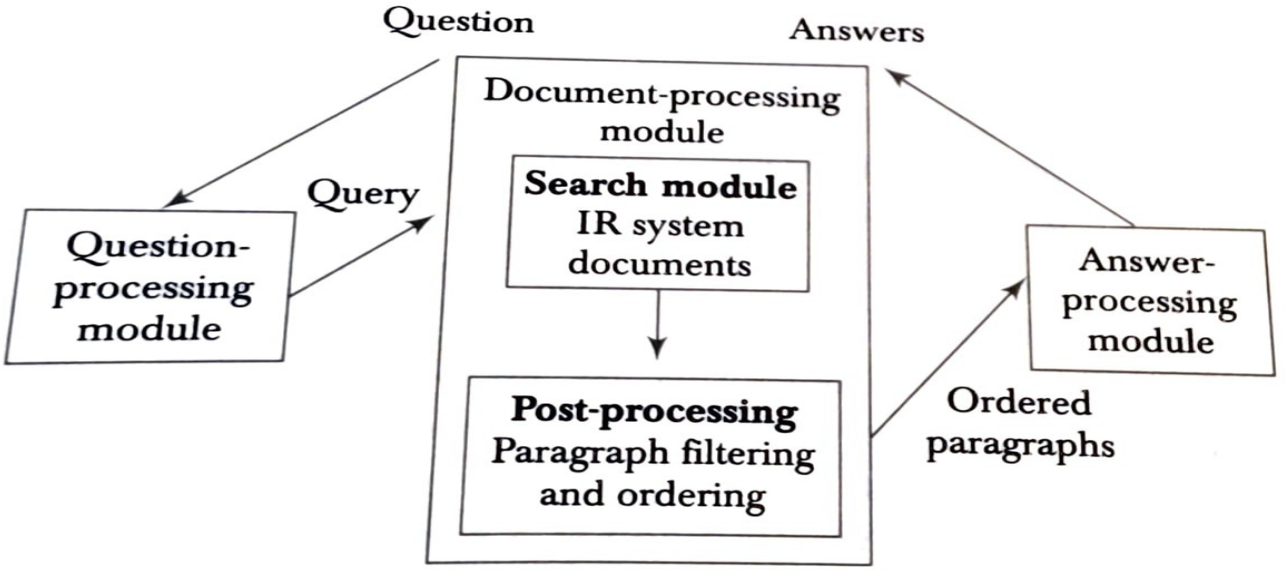
- In general, a question-answering system can use an information extraction system to identify entities in the text.

- It needs a precise analysis of questions and portions of texts and also semantic as well as background knowledge.

Question Answering System

- The current trend in question-answering is towards the processing of large volumes of open domain text.
- An open domain question answering system is supposed to answer questions on any possible topic.
- Such systems cannot rely on hand-coded domain-specific knowledge to provide correct answers.
- START is perhaps the first Web-based question-answering system.

Archicture of Open-Domain QA System



- It consists of three main modules
  - Question processing module / question analysis
  - Document processing module
  - Answer processing module

Question processing module

- Question processing module / question analysis
  - It extracts clues from the question, such as question category, expected-answer type, interesting terms, focus, and semantically related words.
  - The task performed by this module are as follows:
    - Identifying question type
      - The first step is to identify question type i.e. time,location, person, place, size etc.
      - Questions are usually categorized based on the answer type.
      - It involves morpho-syntactic analysis of the question.
      - Knowing question type defines what constitutes relevant data, which helps other modules to correctly locate the answer.

Question processing module

- Identifying answer type
  - The second step in question analysis is to identify the answer type.
  - Answer types often tied to the classes recognized by name entity recognizers e.g. method, explanation and recommendation.
  - The expected answer type is defined on the question terms, using rules as
    - Who X? → answer type = Person
    - Where X? → answer type = Place
    - When X? → answer type = Date
    - How X? → answer type = Method
    - How many X? → answer type = Number
    - Why X? → answer type = Explanation

Question processing module

- Identifying question topic and focus
  - The topic of the question is the object or event that the question is about.
  - The property of the topic that is the target answer for the question being asked is the focus.
  - Example: “What is the height of Qutub Minar?”

    Topic is “Qutub Minar”

    Focus is height.
- Transforming question into IR query
  - The last step in question analysis is to transform the question to an IR query
  - Which is used to retrieve candidate documents.

Document processing module

- The query generated by the question-analysis module is fed to a search engine to retrieve a set of candidate documents
- A question answering system will retrieve a document only when all the keywords are present in the document.
- The results from retrieval system are documents.
- For question answering system, a whole document is too large a unit

Document processing module

- Hence they are filtered to remove paragraphs that do not contain all the keywords.
- This filtering reduces the amount of text that is to be analysed in detail.
- The underlying assumption behind paragraph filtering is that the answer will appear in a few neighboring paragraphs.
- Using this heuristic a set of consecutive paragraphs containing all the question keywords will survive and the rest will be filtered out.

Paragraph Ranking

- The aim of paragraph ranking is to sort candidate passages according to the likelihood of their containing the answer.
- Different criteria can be used to rank:
  - Redundancy
  - Term statistics in passage
  - Keyword sequence
  - Separation between most distant keywords in the text segment etc.

Answer Processing Module

- The answer processing module extracts answers from ranked text segments passed to the answer-processing module by the document processing module.
- As input, the module also receives the expected answer type, query terms, and other question-related information extracted in the question processing phase.
- Based on this information, the exact answer should be carefully selected.
- First, the paragraphs containing the correct answer type are identified.
- The answer type and the focus supplied by question-processing module, guide the search for paragraphs containing the correct answer type.
- Named entity recognizer(NER) and POS is commonly used in this process.
- If no match is found, then best-ranked paragraph is usually returned.

Answer Extraction

- Once the potential paragraphs are selected, answer candidates are extracted and ranked.

- Shallow parsing (NER,POS) is required to extract answer candidates.

- The candidate answers are passed to a validation algorithm, which selects the best answers.

Question-Answering system Evaluation

- Mean reciprocal rank(MRR), confidence weighted score(CWS), or precision and recall can be used to evaluate a question-answering system.

Practce Questions - Question-Answering system

- Explain with a suitable example, how question-type detection plays an important role in extracting the answer of a query in a Question-Answering system.
- Differentiate between information extraction and question answering system.
- Discuss the usability of the question-answering systems with examples. Also, give your justifications on the major challenges in designing open domain and closed domain question answering systems.

- Explain the phases involved in designing a Question Answering System with associated issues.

Information Extraction(IE)

- An information extraction system captures and outputs factual information contained within a document.
- Like an IR system, it responds to a user's information need.
- Unlike IR system, the information need is not expressed as a keyword query.
- It is specified as a pre-defined template
- An IR system identifies a subset of documents in a large repository of text database, whereas an IE system identifies a subset of information within a document that fits the pre-defined templates

Information Extraction(IE)

- An information extraction system does not try to understand the text, as a question-answering system would.

- It tries to identify assertions about very specific kinds of facts.

- Instead of generating answers, the information is distilled into a structured form in which individual facts are accessible.

Information Extraction(IE)

- Two important characteristics of an information extraction system are as follows:
  - The desired information is expressed as a relatively simple and fixed format, usually a template or frame.
  - Only a small fraction of the text qualifies for filling slots in the template or frame; the rest are discarded

- The input to an information extraction system can be expressed as a database schema or template, specifying the output format.

Information Extraction(IE) Example

- Consider the following text taken from MUC-7

  "*A relevant article refers to a vehicle launch that is scheduled, in progress or has actually occurred and must minimally identify the payload, the date of the lunch, whether the lunch is civilian or military, the function of the mission and its status.*"

- The following figure shows the information sought by the text in the form of a template.

task: Launch Event
Vehicle:
Payload:
Mission_Date:
Mission_Site:
Mission_Type (Military, Civilian)
Mission_Function (Test, Deploy, Retrieve)
Mission_Status (Succeeded, Failed, In Progress, Scheduled)

Information to be extracted expressed as a template

Information Extraction(IE) Example

- The following figure shows sample text and the information extracted using the template in previous slide.

The second developmental test flight of India's Geosynchronous Satellite Launch Vehicle, GSLV, was successfully carried out this evening May 8, 2003 from Satish Dhawan Space Centre, Sriharikota, about 100 km north of Chennai, marking a major milestone in the Indian space programme. The payload is 573 pounds (260 kilograms) heavier than the one launched on the first GSLV flight in April 2001. With this launch, India has moved further in establishing its capability to launch geo-synchronous communication satellites.

Task:	Launch Event
Vehicle:	Geosynchronous Satellite Launch Vehicle
Payload:	573 pounds
Mission_Date:	May 8, 2003
Mission_Site:	Satish Dhawan Space Centre, Sriharikota: Chennai
Mission_Type (Military, Civilian):	Civilian
Mission_Function (Test, Deploy, Retrieve):	Test
Mission_Status (Succeeded, Failed, In Progress, Scheduled):	Succeeded

Sample text and extracted template

Design of an Information Extraction(IE) System

- An IE system can be designed using the
  - knowledge-engineering approach(rule-based approach) or
  - Trainable approach

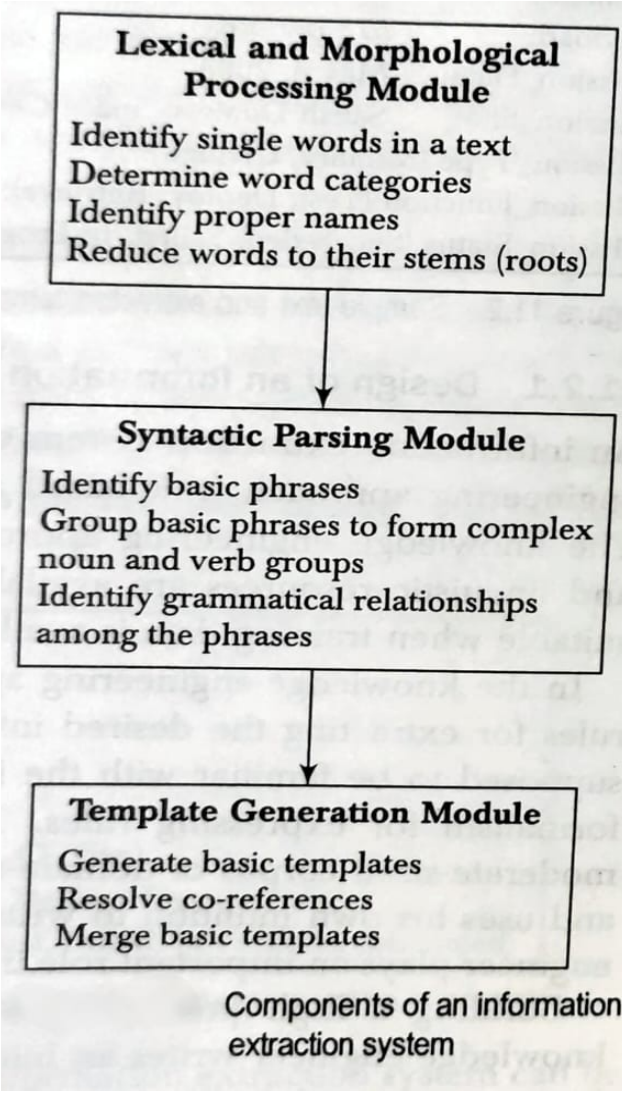
Design of an Information Extraction(IE) System

- Knowledge-engineering approach
  - In knowledge-engineering approach, a knowledge engineer writes rules for extracting the desired information by consulting domain experts and uses his own intuition.
  - The knowledge engineer writes an initial set of rules.
  - Then the system runs over a training corpus and the information is extracted.
  - By looking at the output, the knowledge engineer then makes the appropriate modifications, and iterate the process
  - The skill of the knowledge engineer plays an important role in the performance of the overall system.
  - It has the advantage that it results in a high performance system.
  - The disadvantage is that, it is fairly time consuming and labour intensive and require human expertise to write rules.

Design of an Information Extraction(IE) System

- Trainable approach
  - This approach eliminates the need for human expertise.
  - But it requires an annotated corpus for training.
  - For example, to train a name recognizer, a corpus of texts annotated with the domain-relevant proper names is needed.
  - Once the annotated text is avaliable, the training algorithm is run, and statistics or rules are derived automatically from the training corpus, which is the used to analyze the new texts.

Components of Information Extraction(IE) System



Components of an information extraction system

Components of Information Extraction(IE) System

- It has 3 basic components
  - Lexical and Morphological processing module
  - Syntactic parsing module
  - Template generation module

Components of Information Extraction(IE) System

- Lexical and Morphological processing module
  - It has a tokenizer which divides the text to sentences and tokens
  - It handles inflectional variants of a word
  - It uses a lexicon to associate properties like POS tag and meaning with each word
  - Assign lexical features to items that have internal structures like date,time, proper names etc
  - Due to the requirement of extracting proper names, so every IE system includes Named-entity recognizer(NER).
  - Names often have internal structures. e.g. a company name can usually be identified by its final token.
    - <word><word> Company (Coca-Cola Company)
    - <word><word> Ltd. (Tata Tea Ltd.)

Components of Information Extraction(IE) System

- Syntactic parsing module
  - It does not require detailed parsing.
  - It identify basic phrases like noun groups, verb groups etc.
  - Usually finite state grammar is used for this.
  - Group basic phrases to form complex noun and verb groups.
  - Identify grammatical relationships among phrases
- Template generation module
  - It involves discourse level and domain specific processing.
  - For example, it must perform co-reference analysis which identifies the event or entities in individual sentences that refer to the same entity or event in the real world.

Practice Question-IE

- Differentiate between information extraction and question answering system.
- Differentiate between information retrieval and information extraction.
- What are the advantages and disadvantages of using the knowledge engineering based approach to design an Information Extraction system
- Discuss the approaches to design an Information Extraction system.
- Discuss the major components of an Information Extraction system and the role of NER tagging in it.

Practice Question-MT

- Explain the methodology used in the interlingua-based machine translation system with its advantages.

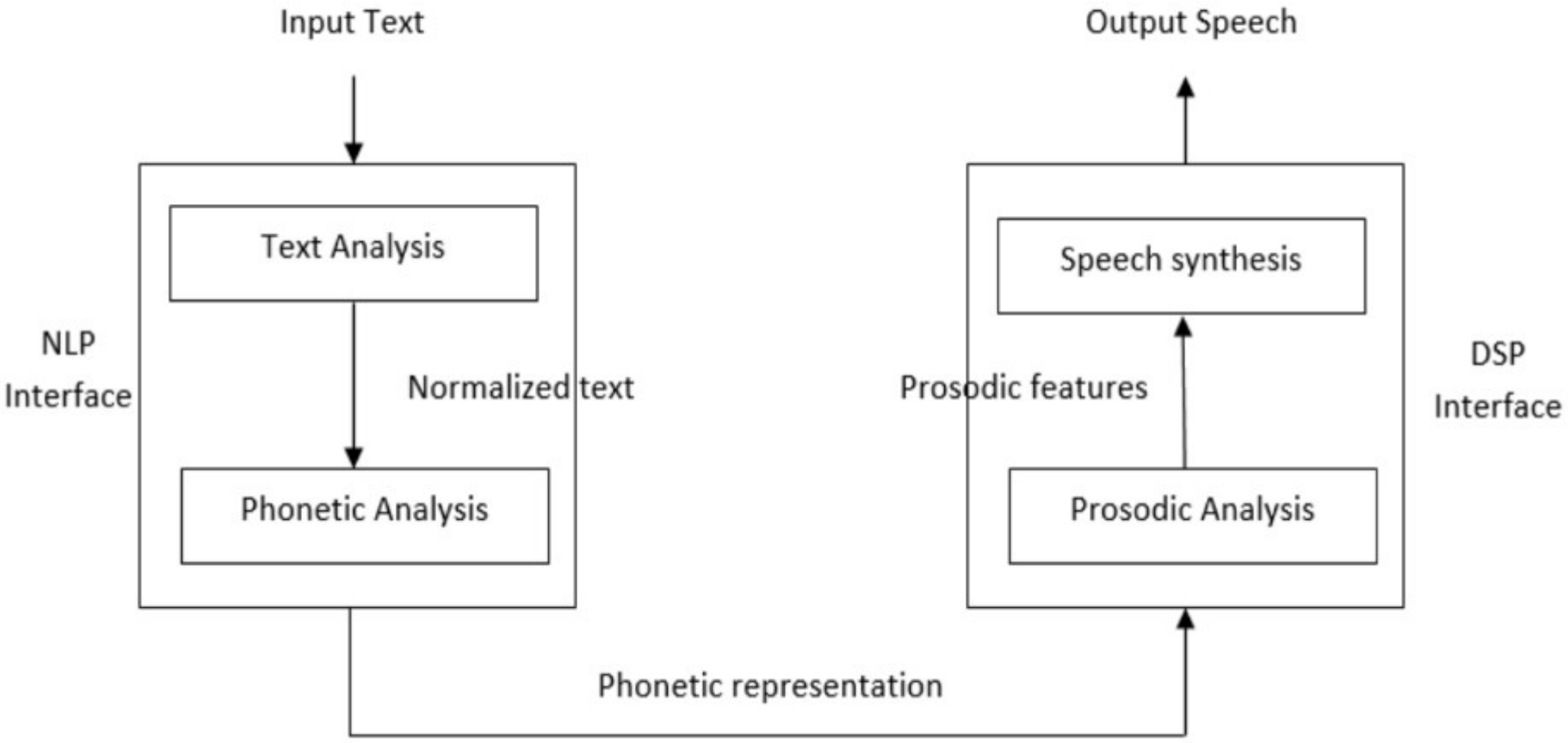
- Explain some useful applications of the Machine Translation System. Discuss the methodology used in the transfer-based machine translation systems. Explain with a suitable example how it is different from the direct machine translation techniques.
- Discuss the example-based and statistical machine translation approaches.
- Differentiate between direct machine translation and transfer-based machine translation systems with a suitable example on its working procedure.

Text to Speech Technology

- Speech synthesis is the artificial production of human speech
- A text-to-speech (TTS) system converts natural language text into speech

- Applications:
  - Speech synthesizers
  - Screen readers
  - Language Education
  - Telecommunications and Multimedia Applications
  - Games or Talking Toy
  - Other applications: E mail reader, PDF readers, pronunciation dictionaries

Overview of TTS System



Overview of TTS System

- A TTS system is composed of 2 parts:
  - NLP (Natural Language Processing) interface at front end and
  - DSP (Digital Signal Processing) interface at the back end.

- The front end process the input text and assigns phonetic transcriptions to each word.

- The back end also referred as the synthesizer converts the symbolic linguistic representation into sound.

Text Analysis

- The text analysis phase is the front end language processor of the TTS system which
  - accepts input text,
  - analyze it and
  - organize into manageable list of words.

- The input text may contain symbols, numbers, abbreviations, etc.
- The text analysis phase normalizes the input text to get the appropriate pronunciation.

Phonetic Analysis:

- It involves grapheme-to-phoneme conversion.
- i.e. the orthographical symbols are converted into phonological forms called the phonetic representation.

- Example:
  - the phonetic representation of the word ‘water’ may be derived as wa-ter if a syllable-based technique is used.

Prosodic Analysis:

- It involves analysis of different prosodic features for the synthesized speech that will be generated by the speech synthesis technique.

- Prosodic features contain pitch, duration, intonation, etc.

- With a good control over the features, emotion, age, gender, etc very natural sounding speech may be generated.

Speech Synthesis:

- It this phase, the output speech is generated by using a speech synthesis technique with respect to the prosodic features.

- Types of speech synthesis techniques:
  - Corpus based techniques: A speech corpus with pre-recorded speech segments for words, syllables, phones, etc. is used to produce the synthesized speech.
  - Rule-based techniques: Produce the speech based on a set of parametric rules for speech production.

Speech Synthesis Models

- Articulatory synthesis
- Formant synthesis
- Concatenative

- Hybrid synthesis
  - Statistical Parametric synthesis
  - Deep learning-based Models

Issues with Text to Speech Conversion

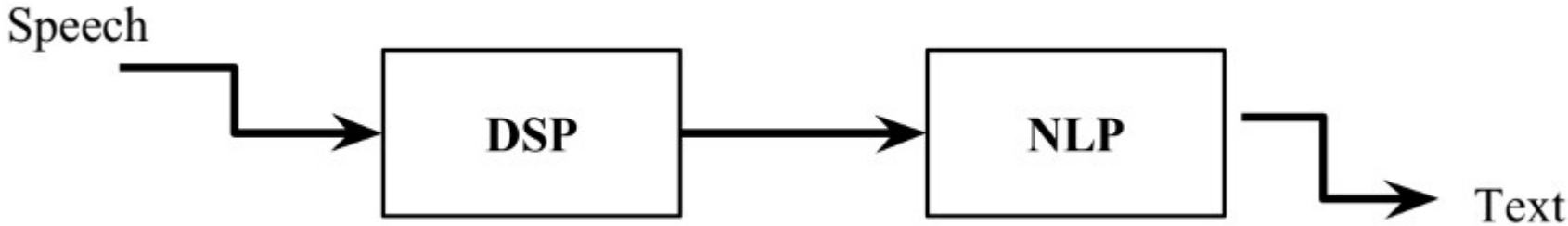
- Text Analysis and Normalization:
  - pronunciation of the word under different context
  - context based numeral reading
- Phonetic Analysis:
  - set of grapheme to phoneme conversion rules are needed
- Prosodic Analysis:
  - Appropriate identification of vocal features, and
  - emotional contents to produce highly natural sounding speech
- Selection of appropriate TTS technique

Speech To Text (STT) System

- Speech to text (STT) conversion is the process of converting spoken words into written texts.
- This is also called speech recognition.

- Applications:
  - Command and control
  - Dialog system
  - Text dictation
  - Audio document transcription

Overview of Speech to Text Technology

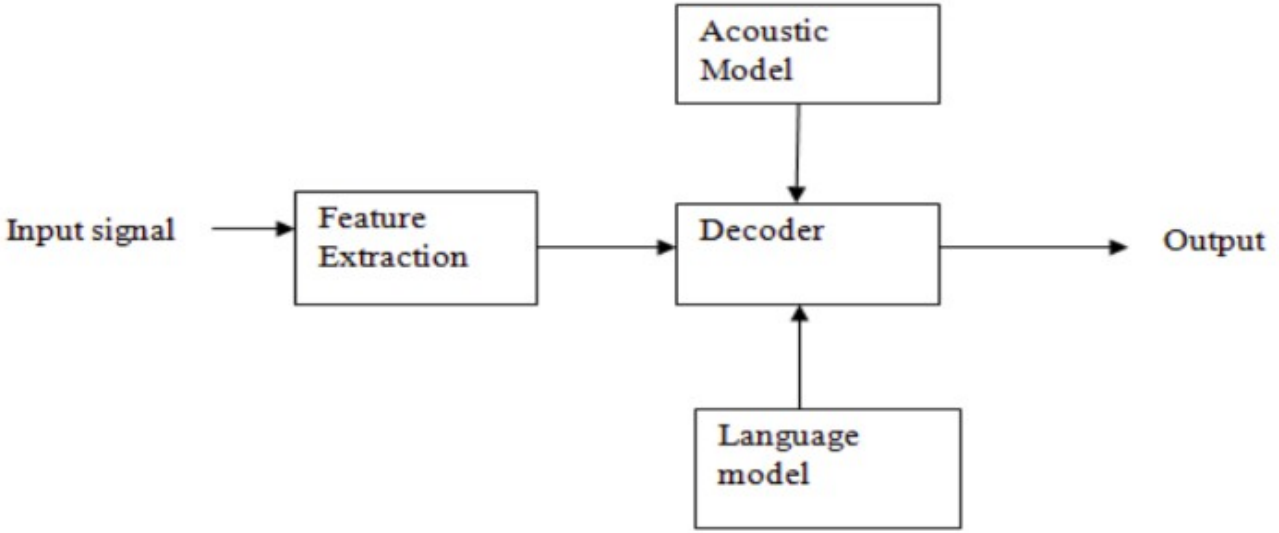


- The input to the STT system is spoken words or sentences in any natural language and the output is the text representation for the words.

- The front end process the input speech and the back end produces the desired output text.

Overview of Speech to Text Technology

- The STT systems rely on two models:
  - Acoustic model: Convert audio into smaller acoustic units
  - Language/linguistic model: converts the acoustic unit into words or phrases
- In addition large vocabulary systems use a pronunciation model.

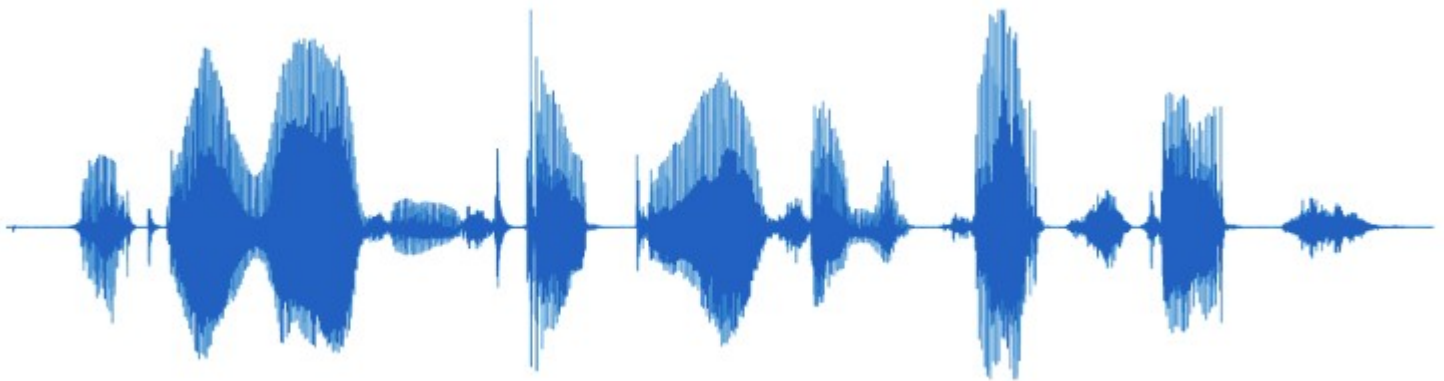


How STT System Works: Major Phases

- Input speech pre-processing
- Speech Segmentation
- Context Analysis and Output Text Generation

Input Speech Pre-processing

- Analog-to-Digital Converter (ADC):
  - This detects the sound vibrations as you speak and converts them to a digital format that the computer can understand.
- Background noise is filtered out and the sound is separated into different frequency bands.
- The sounds are also normalized and adjusted to a constant volume and speed level.



Speech Segmentation

- The sound signal is chopped into small fragments
- These fragments are then matched to known phonemes of the language.
- A phoneme is the smallest pronounceable unit in a language
- Example: As per linguists, the English language has approximately 40-44 phonemes.

Context Analysis and Output Text Generation

- The converter program then examines the order of the phonemes and runs complex mathematical models to analyze the context.

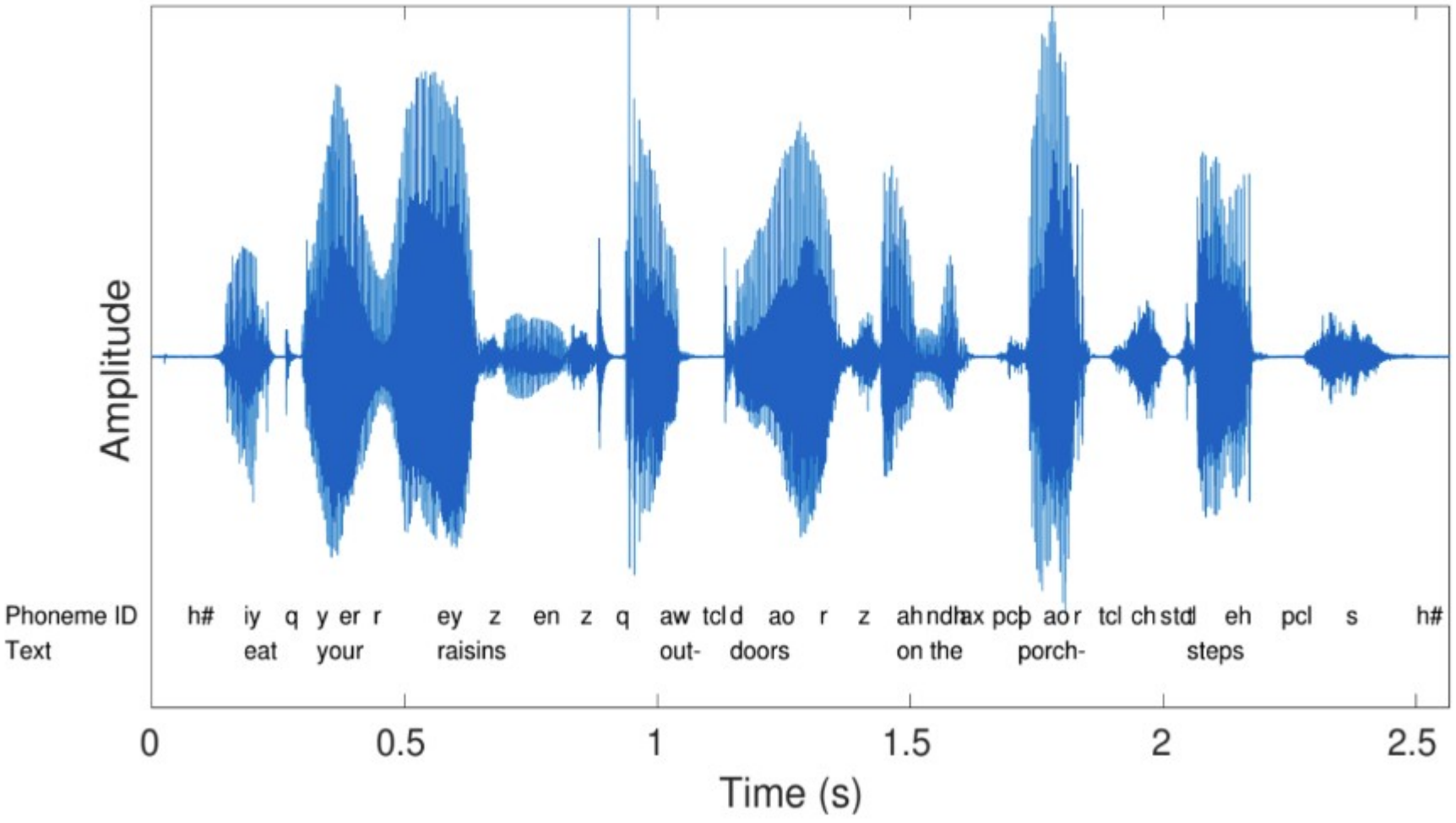
- It also runs them through a database of known words, sentences, and phrases to determine with a high probability what the user is saying.

- Based on the similarity match, the corresponding text is generated.

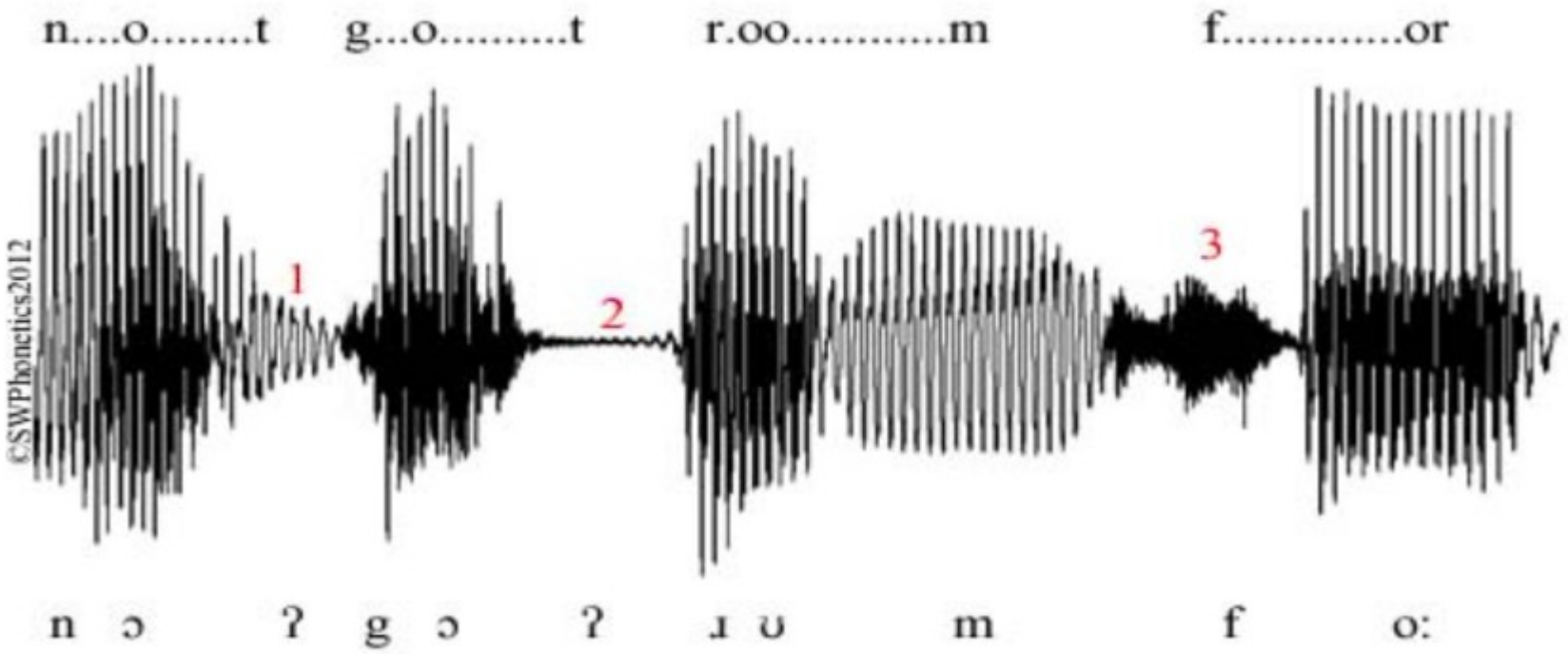
Models for Speech to Text Conversion

- Hidden Markov Model (HMM) based statistical models
- Machine Learning models
- Deep learning

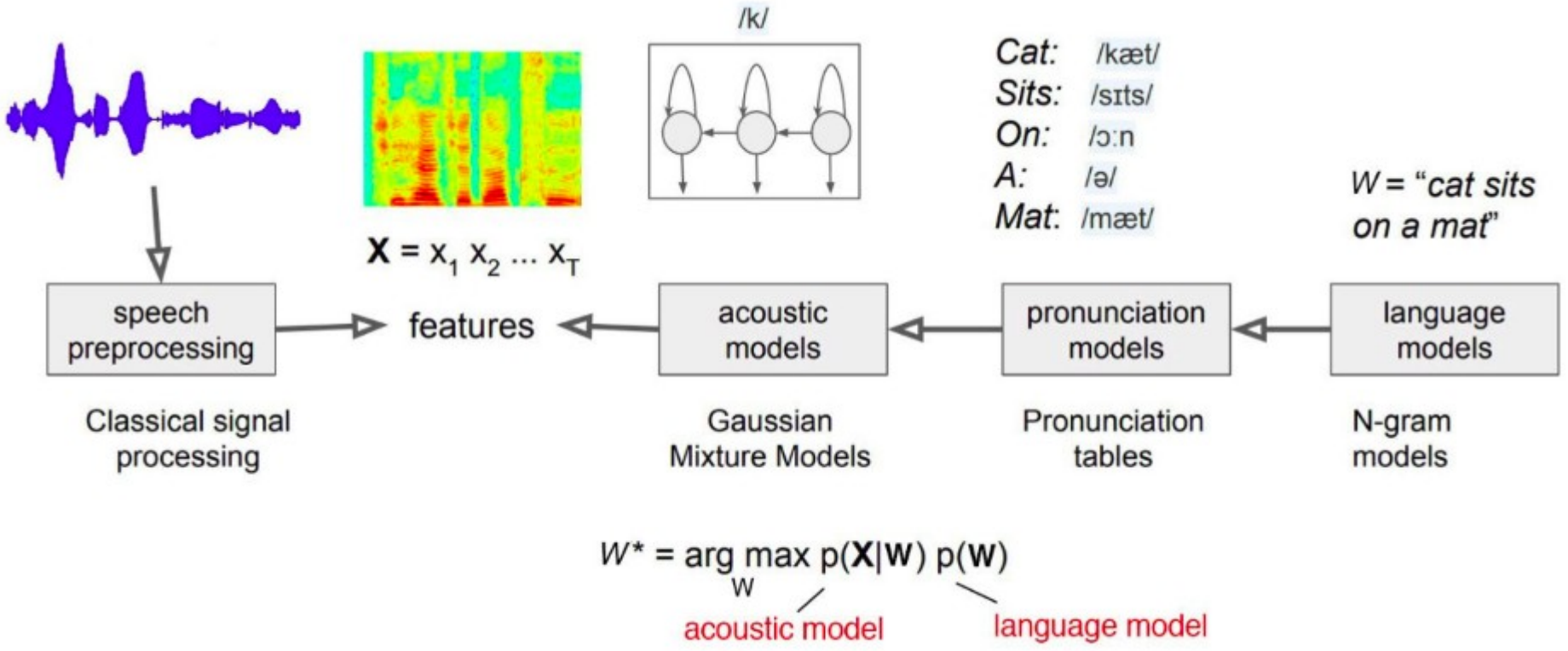
Example Waveform



Example Waveform



Example of a STT Conversion Model



Powerful STT Models

- Google Docs Voice Typing
- Apple Dictation:
  - built-in speech to text converter software that uses Siri’s servers to capture voice notes
- Windows Speech Recognition

Challenges in Speech to Text Conversion

- Imprecise interpretation
- Time
- Accents
- Background noise and loudness
- Training data set for low resourced languages

What is a Chatbot?

A chatbot is a computer program governed by a set of pre-defined rules or artificial intelligence that grants it the capabilities to communicate with and like a human.

A set of commands is fed into the system that makes it smart enough to interpret and react to user-inputted queries.

History of Chatbots

The idea of a bot came to light back in 1950 when Alan Turing's article about artificial intelligence was published, called as- "**Computing Machinery and Intelligence**".

ELIZA (1966) and **PARRY** (1972) were the first two chatbots that were intended to understand and simulate human language.

Since then the industry has grown bigger and better especially with the emergence of messenger bots that integrate with platforms like Facebook, WhatsApp, Telegram, WeChat and alike.

Types of Chatbots

There are two types of chatbots:

#1. Based on a set of rules-

Can be referred to as a pre-programmed bot with limited capabilities. Responds to a specific set of commands and fails to do so if the inputted query does not match the database.

#2. Based on machine learning-

It has an artificial brain that's powered by AI (Artificial Intelligence). It not just understands commands but a complete language. It learns continually with the conversations of past that eventually makes it smarter and better.

Types of Chatbots

- Weather bot
- Life advice bot
- Grocery bot
- Ticket booking bot
- Financial advice bot
- News bot
- Scheduling bot
- And the list goes on

Some of The Chatbots Prominent Today Include The Following:

How Does Chatbot Work?

A chatbot powered by AI or the one which relies on Natural Language Processing (NLP) combines several steps of code to transform text or speech into structured data, which is then used to generate a relevant response.

This is what happens:

- **Tokenization**- A chatbot bisects the string of words into tokens.
- **Identification of the entity**- Identifies and classifies the words, like a product's name, a person's name or an address.

How Does Chatbot Work?

- **Normalization-** Searches for common errors like spelling or typographical mistakes.
- **Speech identification-** Recognizes and understands parts of speech and sentence structures.
- **Dependency parsing-** Recognizes subject, object and dependent phrases in a sentence.
- **Sentiment analysis-** Analyzes the quality of the conversation in regard to a real human.

Benefits of Chatbots

With endless possibilities, chatbots can transform the thinkable into a reality. From assisting users in buying shoes to scheduling a meeting, everything can be handled with ease.

- 24x7 instant customer support
- Helps in generating leads by answering queries in real time

- Multiple platform support
  (like Facebook, WeChat, Skype, Telegram and more)
- Performs multiple tasks with supreme accuracy and efficiency

Technology Behind

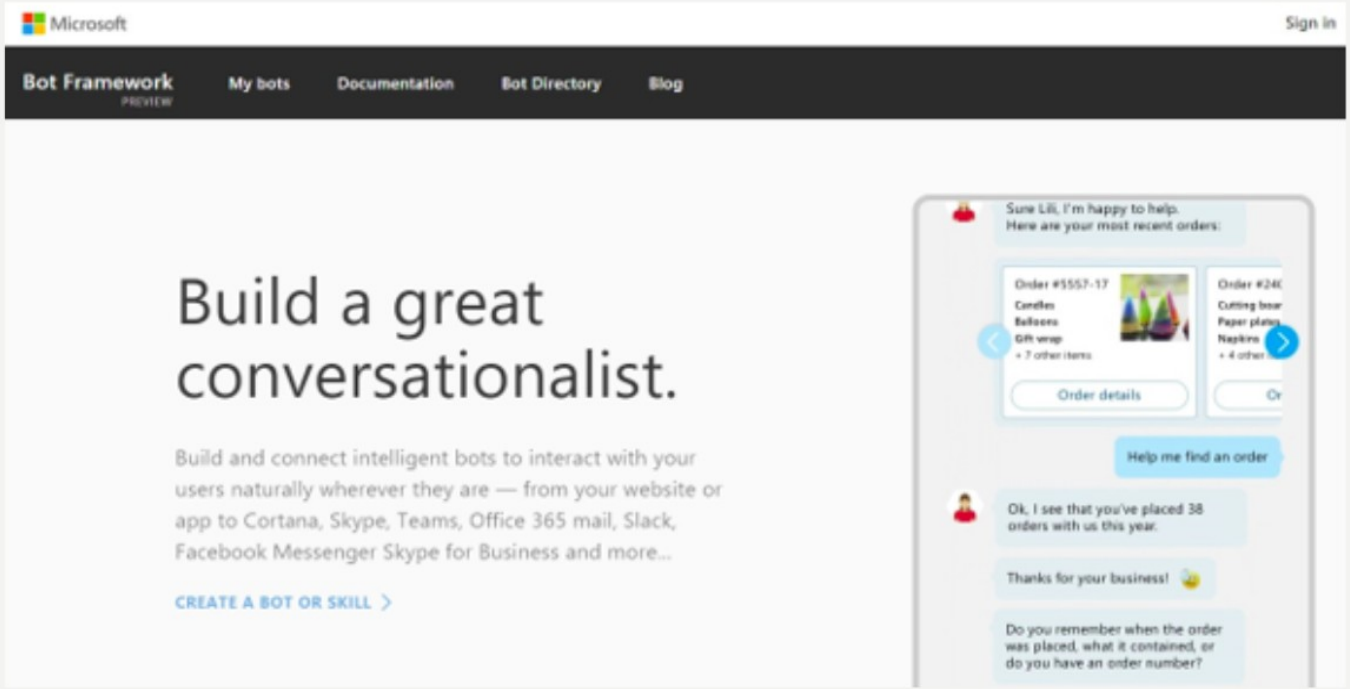
Chatbots operate either through APIs or learn to respond like humans via machine learning.

- **TUI (Textual User Interface)**- Bots recognize the text commands and respond to users' requests accordingly

- **Machine Learning**- Bots are smarter with the ability to learn patterns of natural language, interpret data of the past and respond as intended

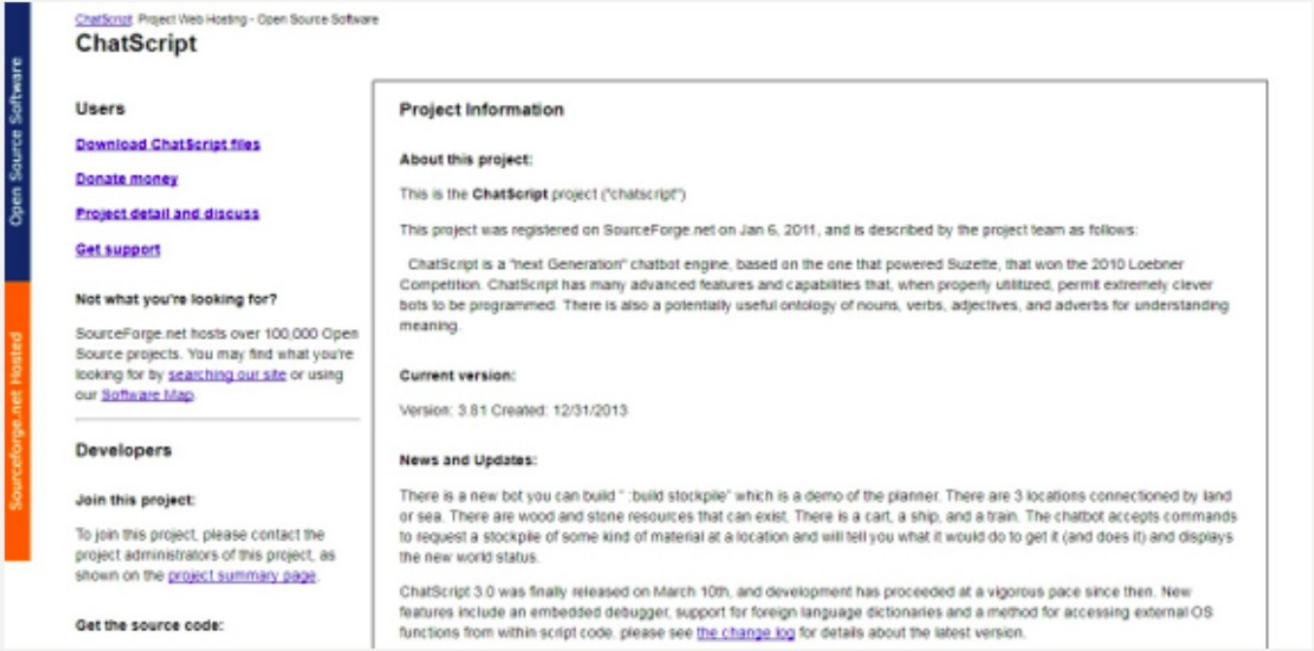
#1 Microsoft's Bot Platform

Build, test and publish your own bots, using frameworks like Express and Restify.



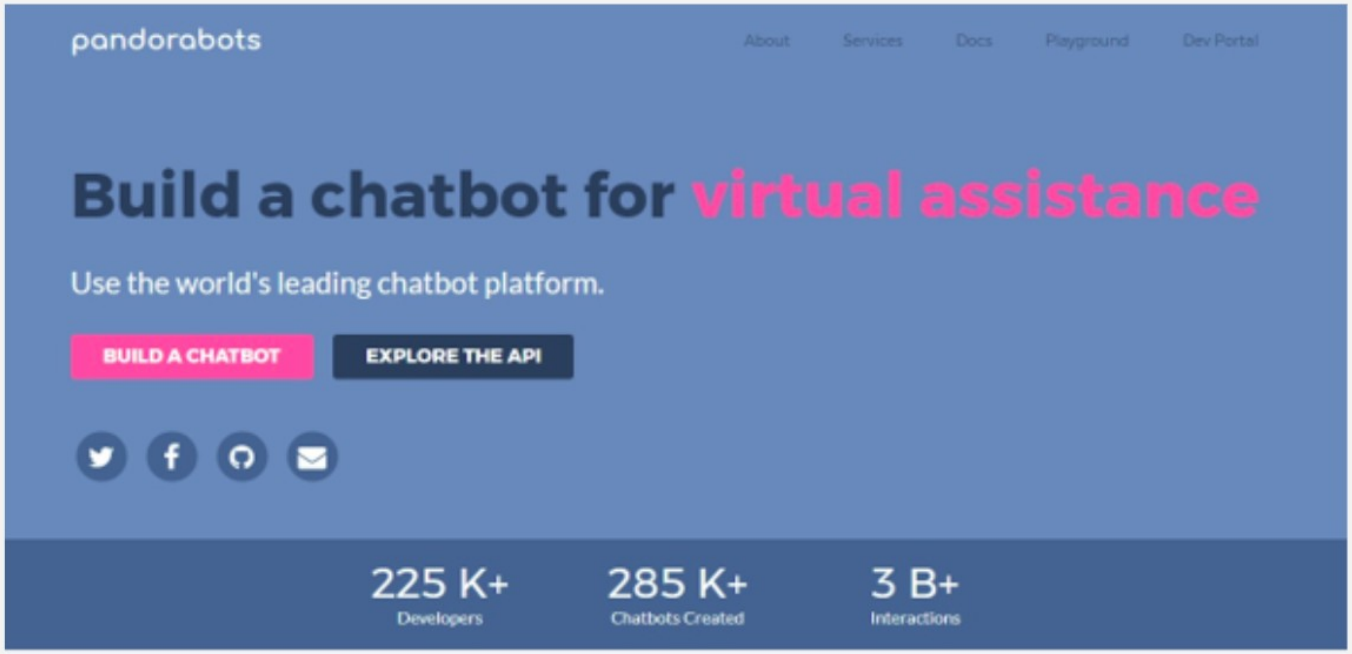
#2 Chatscript

Four times Loebner Prize winner, it's open source and easy to build and deploy chatbots.



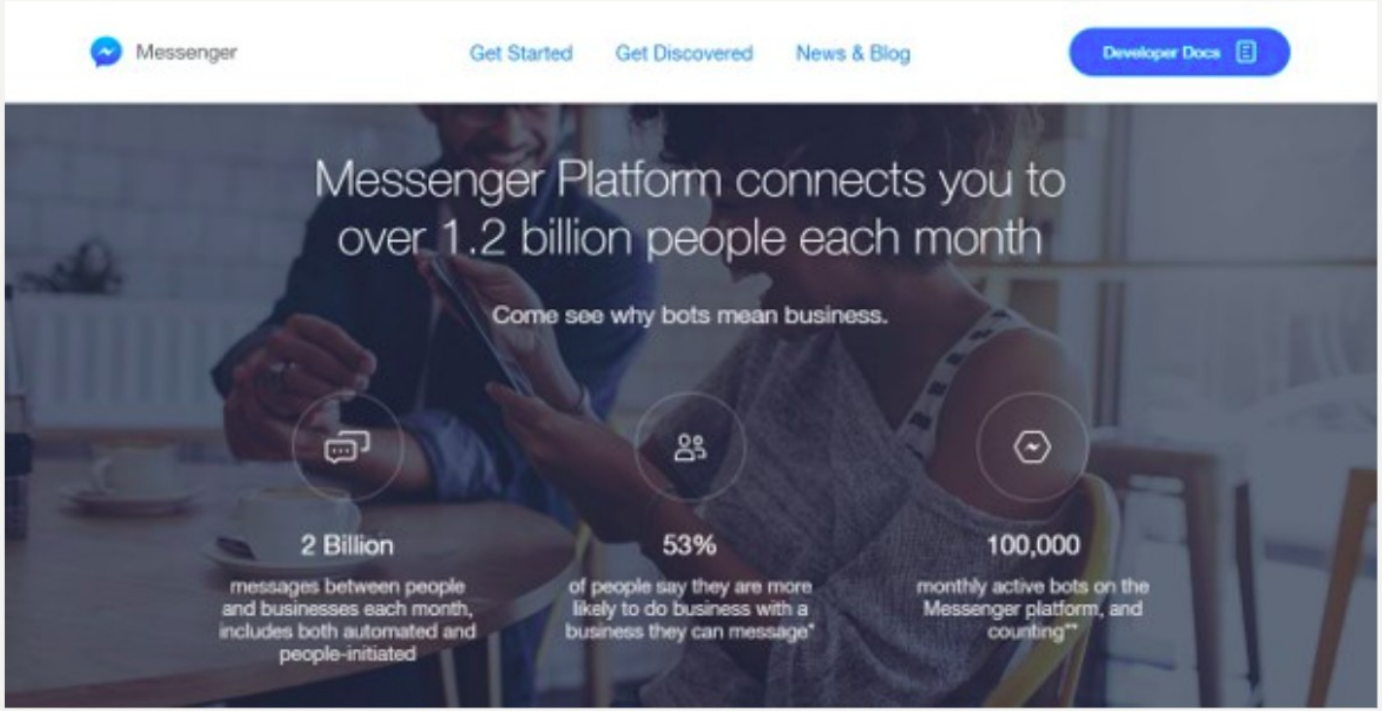
#3 Pandorabots

It uses AIML (Artificial Intelligence Markup Language) and offers an integrated development environment for chatbot development.



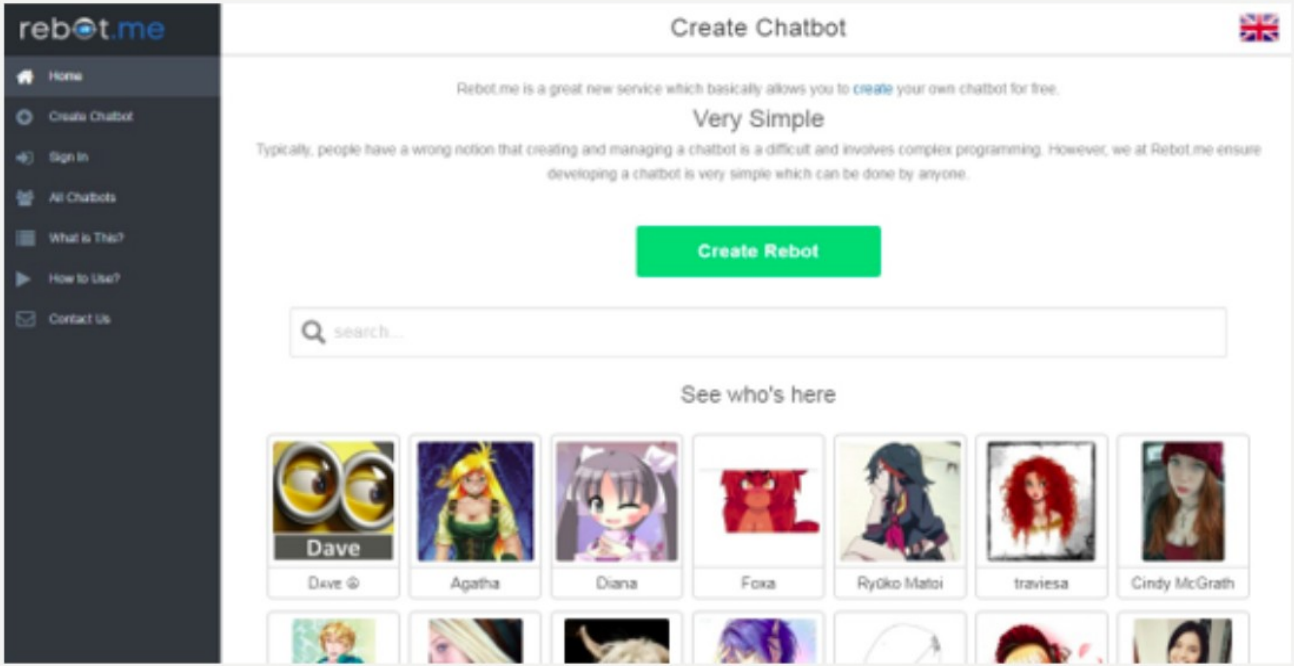
#4 Facebook's Bots for Messenger

It helps build bots, submit them and wait for experts to approve.



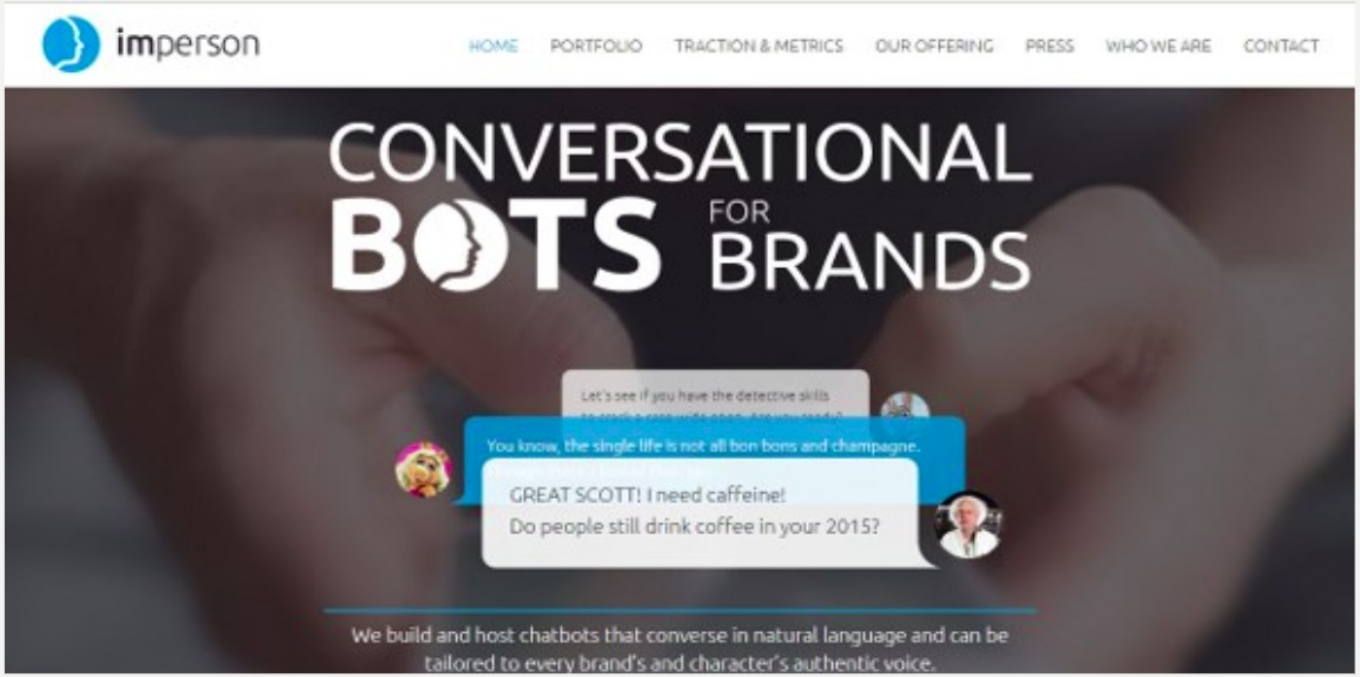
#5 Rebot.me

You don't require programming skills to build a chatbot using Rebot.me as it learns when you interact with it.



#6 Imperson

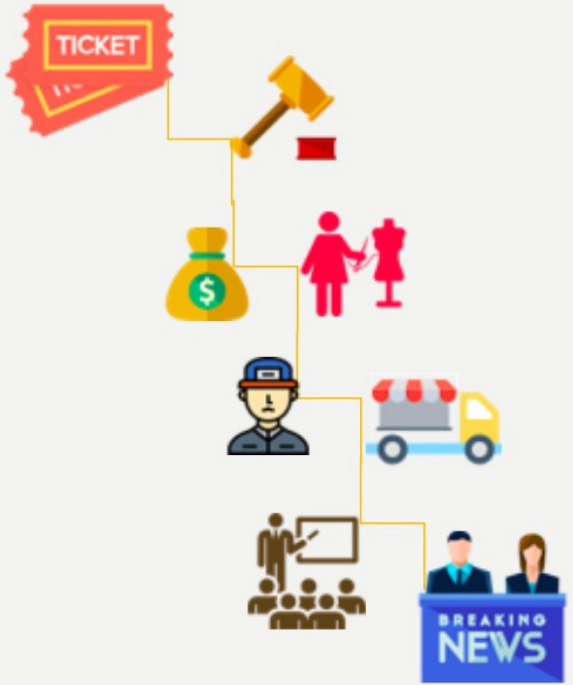
It specializes in character bots helping digital agencies to develop bots for enhanced and engaging relationships.



Uses of Chatbots

With an unprecedented hype lately, chatbots usage has grown significantly. Most popular uses of chatbots include:

- As a ticket booking assistant
- As a legal consultant
- As a personal financial advisor
- As a fashion stylist
- As a life coach
- As a food ordering partner
- As a teacher
- As a newsreader
- And more…



Future of Chatbots

In the coming years, chatbots will rule a variety of business verticals due to the following undeniable reasons:

- Ease of creating and deploying chatbots due to comprehensive tools and platforms
- An alternative to a lot of native applications which are expensive to develop
- Bots can learn, become better and do tasks 24x7
- Helps in enhancing user experience and support
- Easy integration into messaging apps eradicates development and designing cost of UI

Future of Chatbots

The future of chatbots looks bright as more and more companies are investing time, money and effort to achieve their business goals via modern technologies.