## 1.Machine Translation.

1. Explain the different approaches for Machine Translation.

MT is classified into seven broad categories:

1. **rule-based**

* A Rule-Based Machine Translation (RBMT) system consists of
* a collection of rules, called grammar rules,
* a bilingual or multilingual lexicon, and
* software programs to process the rules.
* Having input sentences (in some SL), an RBMT system generates them to output sentences (in some TL) **on the basis of morphological, syntactic, and semantic analysis** of both the source and the target languages involved in a concrete translation task.
* 3 different approaches under RBMT category are
  + Direct Translation
  + Interlingua MT and
  + Transfer-based MT.

1. **knowledge-based**

* Emphasis is on functionally complete understanding of the source text prior to the translation into the target text.
* KBMT is implemented on the **Interlingua architecture**
* Once the SL is analyzed, it will run through the **Augmenter**. It is the knowledge base that converts the source representation into an appropriate target representation before synthesizing into the target sentence.
* KBMT systems provide high quality translations.
* They are quite expensive to produce due to the large amount of knowledge needed to accurately represent sentences in different languages.
* E.g. The English-Vietnamese MT system

1. **principle-based**

* PBMT Systems employ parsing methods based on the Principles & Parameters Theory of Chomsky‘s Generative Grammar.
* The parser generates a detailed syntactic structure that contains lexical, phrasal, grammatical, and thematic information.
* It also focuses on robustness, language-neutral representations, and deep linguistic analysis.
* PBMT systems are not efficient methods for languages applying different principles.
* E.g. UNITRAN (Universal translator used by YouTube, Netflix)

1. **Empirical MT (EMT) Systems**

* EMT systems rely on large parallelly aligned corpora.
* Empirical systems acquire the knowledge about a set of rules describing the translation process automatically from a collection of translation examples.
* It uses automatically induced rules.
* 2 Categories :
  + Statistical
    - Translation is based on the knowledge and statistical models extracted from bilingual corpora.
    - There are three different statistical approaches in MT named :
      * Word-based Translation
      * Phrase-based Translation
      * Hierarchical Phrase-based model.
  + Example-Based
    - An EBMT system is given a set of sentences in the SL and their corresponding translations in the TL, and uses those examples to translate other, similar SL sentences into the TL.
    - The basic logic is that, if a previously translated sentence occurs again, the same translation is likely to be correct again.

1. **hybrid-based**

* Takes advantage of both statistical and rule-based translation methodologies
* It is more efficient
* These systems are based on both rules and statistics.

1. **online interactive based**

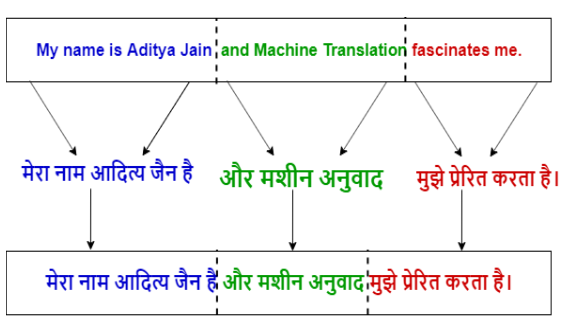
* In this interactive translation system, the user is allowed to suggest the correct translation to the translator online.
* **This approach is very useful in a situation where the context of a word is unclear and there exists many possible meanings for a particular word.**
* In such cases, the structural ambiguity can be solved with the interpretation of the user.

1. Explain in detail Rule Based Machine Translation

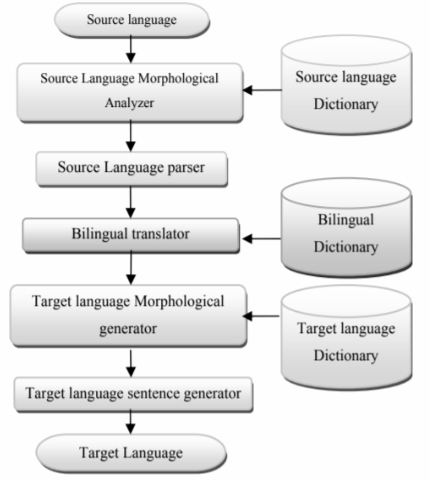
[Lec31\_Module06.pptx](https://docs.google.com/presentation/d/1xDCvWDlaiznjSKKrARCStFb34a7W5WGy/edit#slide=id.p33)

[Lec32.pptx](https://docs.google.com/presentation/d/1yQ5NtP1LelPBCDxfs6MBSVk4ocDDpWrk/edit#slide=id.g11e633baac4_0_18)

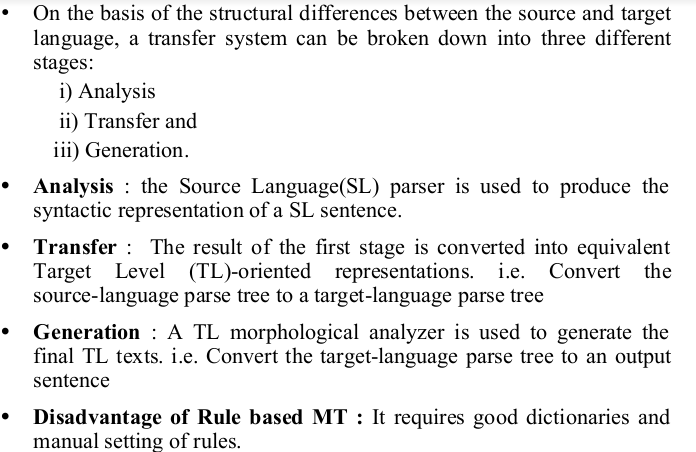
* A Rule-Based Machine Translation (RBMT) system consists of a collection of rules, called grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules.
* A RBMT system is always extensible and maintainable.
* Eg:



* Having input sentences (in some SL), an RBMT system generates them to output sentences (in some TL) on the basis of **morphological, syntactic, and semantic analysis** of both the source and the target languages involved in a concrete translation task.
* It applies a set of linguistic rules in three different phases: **analysis, transfer and generation.**
* It requires: **syntax analysis, semantic analysis, syntax generation and semantic generation.**
* RBMT generates the target text given a source text in the following the steps

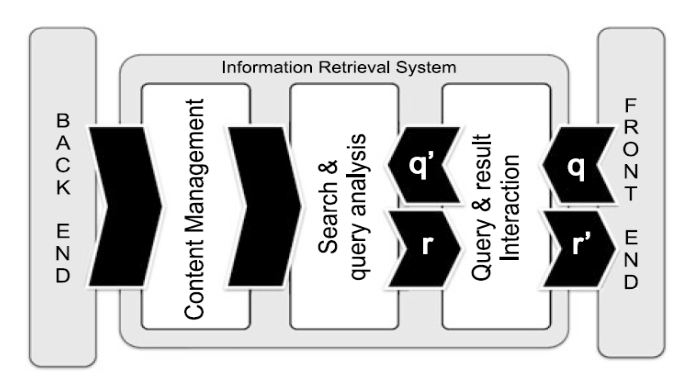
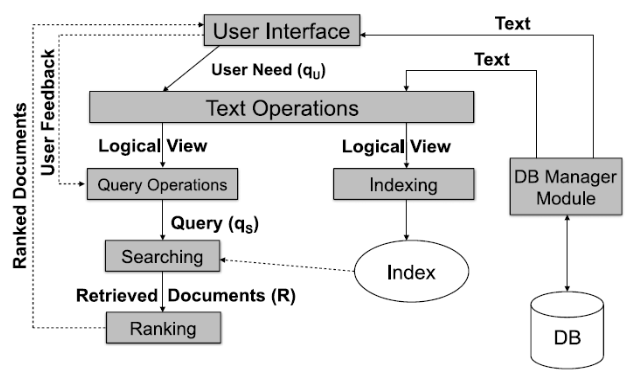


* Source language morphological analyzer analyzes a source language word and provides the morphological information.
* Source language parser is a syntax analyzer that analyzes source language sentences.
* Translator is used to translate a source language word into target language. (Word Translation)
* Target language morphological analyzer works as a generator and it generates appropriate target language words for the given grammatical information.
* Also target language parser works as a composer and it composes a suitable target language sentence.
* This type of MT system needs a minimum of three dictionaries:
  + **Source Language Dictionary:** It is used by SL morphological analyzer
  + **Bilingual Dictionary:** It is used by the translator for translating source language into target language
  + **Target Language Dictionary:** It is used by a morphological generator to generate target language words.
* 3 different approaches under RBMT category are
  + Direct Translation
    - In this method, the SL text is analyzed structurally up to the morphological level and is designed for a specific source and target language pair.
    - After words are translated, simple reordering rules are applied – Example: move adjectives after nouns when translating from English to French
  + Interlingua MT
    - In the Interlingua approach, the translation is performed by first representing the SL (Source Language) text into an intermediary (semantic) form called Interlingua.
    - The advantage of this approach is that Interlingua is a language independent representation from which translations can be generated to different TLs.
    - Thus, the translation consists of two stages:
      * SL is first converted in to the Interlingua (IL) form
      * Translation from the IL to the TL(Target Language).
  + Transfer-based MT.



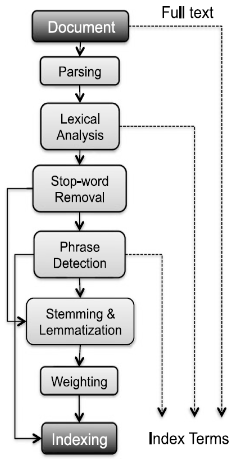
## 2.Information Retrieval

1. Explain with block diagram the working of Information Retrieval System

* IR deals with the representation, storage, and access of information and is concerned with the organization and retrieval of information from large database collections.
* 
* Here, the user issues a query q from the front-end application (accessible via, e.g., a Web browser)
* q is processed by a query interaction module that transforms it into a “machine-readable” query q’ to be fed into the core of the system, a search and query analysis module.
* This is the part of the IR system having access to the content management module directly linked with the back-end information source (e.g., a database).
* Once a set of results r is made ready by the search module, it is returned to the user via the result interaction module; optionally, the result is modified (into r ) or updated until the user is completely satisfied.
* The most widespread applications of IR are the ones dealing with textual data.
* As textual IR deals with document sources and questions, both expressed in natural language, a number of textual operations take place “on top” of the classic retrieval steps.
* The processing of textual queries typically performed by an IR engine are shown in figure.
* 
* The user need is specified via the user interface, in the form of a textual query qU (typically made of keywords).
* The query qU is parsed and transformed by a set of textual operations; the same operations have been previously applied to the contents indexed by the IR system. This step yields a refined query q’U .
* Query operations further transform the preprocessed query into a system-level representation, qS .
* The query qS is executed on top of a document source D (e.g., a text database) to retrieve a set of relevant documents, R.
* Fast query processing is made possible by the index structure previously built from the documents in the document source.
* The set of retrieved documents R is then ordered: documents are ranked according to the estimated relevance with respect to the user’s need.
* The user then examines the set of ranked documents for useful information; he might pinpoint a subset of the documents as definitely of interest and thus provide feedback to the system.
* **Text Processing in IR**
  + Noun words (words or noun phrase groups) are the most representative components of a document in terms of content.
  + When selecting candidate keywords, indexing must fulfill two different and potentially opposite goals:
    - one is exhaustiveness, i.e., assigning a sufficiently large number of terms to a document, and
    - The other is specificity, i.e., the exclusion of generic terms that carry little semantics and inflate the index.
  + Generic terms (conjunctions and prepositions) are characterized by a low discriminative power as their frequency across any document in the collection tends to be high.
  + In other words, generic terms have high term frequency, defined as the number of occurrences of the term in a document.
  + In contrast, specific terms have higher discriminative power, due to their rare occurrences across collection documents: they have low document frequency,
  + defined as the number of documents in a collection in which a term occurs.

1. Explain the different steps in text processing for Information Retrieval

* The textual preprocessing phase typically performed by an IR engine, takes as input as a document and yields its index terms as output.
* The process of this extraction is given in the figure.



1. **Document Parsing.**
   1. Documents come in all sorts of languages, character sets, and formats and the same document may contain multiple languages or formats.
   2. e.g. A French email with Portuguese PDF attachments.
   3. Document parsing deals with the recognition and “breaking down” of the document structure into individual components.
   4. In this preprocessing phase, unit documents are created;
   5. e.g., emails with attachments are split into one document representing the email and as many documents as there are attachments.
2. **Lexical Analysis.**
   1. After parsing, lexical analysis tokenizes a document, seen as an input stream, into words.
   2. Issues related to lexical analysis include the correct identification of accents, abbreviations, dates, and cases.
   3. The difficulty of this operation depends much on the language at hand.
3. **Stop-Word Removal**
   1. A subsequent step optionally applied to the results of lexical analysis is stop-word removal, i.e., the removal of high-frequency words.
   2. The subsequent phases take the full-text structure derived from the initial phases of parsing and lexical analysis and process it in order to identify relevant keywords to serve as index terms.
4. **Phrase Detection**
   1. This step captures text meaning.
   2. Phrase detection may be approached in several ways, including rules morphological analysis syntactic analysis, and combinations thereof.
   3. A common approach to phrase detection relies on the use of thesauri
   4. **Thesauri usually contain synonyms and antonyms.**
   5. An alternative to the consultation of thesauri is to use machine learning techniques such as Key Extraction Algorithm (KEA)
5. **Stemming and Lemmatization**
   1. Stemming and lemmatization aim at stripping down word suffixes in order to normalize the word.
6. **Weighting**
   1. The final phase of text preprocessing deals with term weighting.
   2. The words in a text have different descriptive power; hence, index terms can be weighted differently to account for their significance within a document
   3. and/or a document collection.
   4. Such a weighting can be binary, e.g., assigning 0 for term absence and 1 for presence.

## 3.Question Answering System

1. Explain the different steps in the Question Answering System.

* A typical Question Answering System, consists of three main modules:
  + Question Analysis
    - takes natural language questions as input
    - is responsible for analyzing the question completely.
    - Aims to understand the question's purpose and meaning.
    - The question should be analyzed in different ways.
    - **Question Analysis**
      * Firstly, carry out the morph-syntactic analysis of words in the question.
      * This is done by POS tagging.
      * After POS tagging, find out the questioning information (what the question is looking for).
      * To get the meaning of the question, we need to classify the question **semantic type.**
    - **Question Classification**
      * Question classification process is used to generate possible question classes.
      * E.g. The question “Who was the first American in space?” is expecting that the person's name is in the answer
      * **The search space of reasonable answers will be definitely reduced here.**
      * Once the question type is recognized, the question analysis needs to recognize more constraints that the question's description type must meet.
      * This process is simple as taking out keywords from the question and finding candidate answer sentences.
      * These keywords may then be extended by using morphological and/or synonyms replacements or using query expansion techniques.
      * They form a representation of the question.
  + Answer Retrieval
    - It involves the following steps:
      * **Document Retrieval**
        + This module selects a set of relevant documents from a domain specific repository.
        + **Conceptual indexing** is used for the retrieval process since the keyword based indexing ignores the semantic content of the document collection.
        + Both the documents and queries can be mapped into concepts and these concepts are used as a conceptual indexing space for identifying and extracting documents.
      * **Document Processing**
        + The retrieved documents are processed for extracting candidate answer sets.
        + This module is responsible for selecting the response based on the relevant fragments of the documents.
      * **Syntactic Analysis**
        + The documents are analyzed syntactically using the NLP techniques such as POS tagging and NER.
        + Firstly the documents are tokenized into a set of sentences. Then the POS tagging and NER is performed.
        + Shallow parsing is performed to identify the phrasal chunks.
        + The chunks identified in the question analysis module are matched with those identified in the document and relevant sentences are retrieved.
      * **Semantic Analysis**
        + Shallow parsing can be performed for finding the semantic phrases or clauses.
        + Semantic roles are identified and mapped to semantic frames.
        + The sentences whose semantic frames map exactly to the semantic frames of the question are also extracted.
      * **Relation Identification**
        + The base ontology(relationship between the entities) is populated with the domain knowledge incrementally as we go through different sets of documents.
        + By this method a valid knowledge of any specialized discipline can be incorporated into the system.
        + The relations among different concepts are identified using the domain knowledge and the ontological information is obtained.
  + Answer Generation.
    - The filtering of candidate answer set and answer generation is performed.
    - The user is supplied with a set of short and specific answers ranked according to their relevance.
    - The different stages are:
      * Filtering
        + The extracted sentences are filtered and the candidate answer set is produced.
      * Answer Ranking
        + The answer set is ranked based on the semantic similarity.
      * Answer Generation
        + From the answer set, specific answers have to be generated in case the direct answers are not available.

1. Explain the classification of approaches in Question Answer System

There are **eight** criteria in support of classifying the available large number of QASs.

1. **Application domains for which QASs are developed**
   1. The task of generating answers of questions is related to the type of questions asked.
   2. Some users may require **general information** on a general topic
   3. Some may require **specific information** from a particular application domain.
   4. Therefore, selection of the domain as a basis of classification of QASs may be a natural choice.
   5. In this, ***QASs answer domain independent questions.***
   6. **GENERAL DOMAIN(OPEN DOMAIN) QAS’s**
   7. It searches for answers within a large document collection.
   8. Here, the quality of answers delivered is not high, and questions are asked by casual users.
   9. **Pros of general domain QASs**
      1. Users don’t need to acquire knowledge of domain specific keywords for formulating questions.
   10. **Cons of general domain QASs**
       1. The quality of answers is low.
       2. The answer's satisfaction depends upon the users.
   11. **RESTRICTED DOMAIN (CLOSED DOMAIN) QAS’s**
   12. It answers domain specific questions.
   13. Answers are searched within domain specific document collections.
   14. The quality of answers is expected to be high.
   15. **Pros of restricted domain QASs**
       1. The quality of answers generated by restricted domain QASs is high
       2. The level of satisfaction of the users depends on their domain knowledge.
   16. **Cons of restricted domain QASs**
       1. There is a limited repository of domain specific questions; such QASs can answer a limited number of questions.
2. **Types of questions asked by the users**

The different categories are

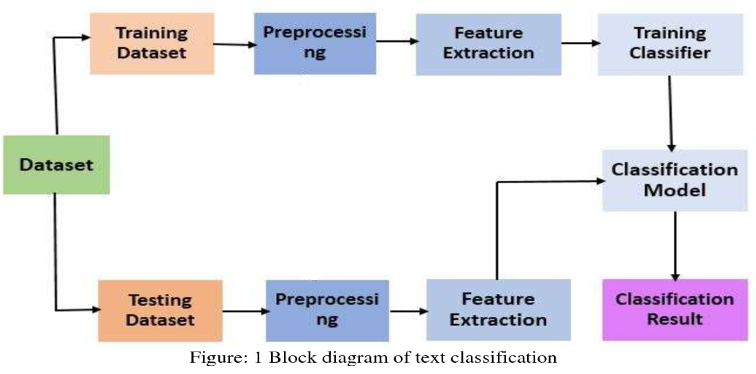
* **Factoid type questions [what, when, which, who,how]**
  + These questions are simple and fact based that require answers in a single short phrase or sentence.
  + The factoid type questions generally start with wh-word.
  + Current QASs have got a satisfactory performance in answering factoid type questions.
* **List type questions**
  + The list questions require a list of entities or facts in answers
  + e.g., – list names of employees getting salaries more than 5K?
  + QASs consider such questions as a series of factoid questions which are asked ten times one after the other.
* **Hypothetical type questions**
  + Hypothetical questions ask for information related to any hypothetical event.
  + They generally begin with ‘what would happen if’. QASs require knowledge retrieval techniques for generating answers.
  + Moreover, the answers are subjective to these questions.
  + There are no specific correct answers to these questions.
* **Confirmation questions**
  + Confirmation questions require answers in the form of yes or No.
  + Systems require inference mechanisms, world knowledge and common sense reasoning to generate answers.
* **Causal questions [how or why]**
  + Causal questions require explanations about an entity.
  + The answers are not named entities as observed in the case of factoid type questions.
  + QASs require advanced natural language processing techniques to analyze the text at pragmatic and discourse level for generating answers.

1. **Types of analyses performed on users’ questions and source documents**
2. **Types of data consulted in data sources**
3. **Characteristics of data sources**
4. **Types of representations used for questions and their matching functions**
5. **Types of techniques used for retrieving answers**
6. **Forms of answers generated by QASs**

## 4.Text Categorization/Classification

1. Explain with diagram the steps in Text Categorization

* The goal in automatic text classification is to assign a document to a category by evaluating its text components.
* Given is the general block diagram of a text categorization system.
* The dataset is split into training and testing dataset for the classification process.
* Then starts the actual processing of text data.



* Text Preprocessing
  + The main objective of pre-processing is to obtain the key features or key terms from stored text documents
  + Pre-processing step is crucial in determining the quality of the next stage, that is, the classification stage.
  + It is important to select the significant keywords that carry the meaning and discard the words that do not contribute to distinguishing between the documents.
  + The pre-processing phase of the study converts the original textual data in a data mining ready structure.
  + **methods for text cleaning and preprocessing text data sets.**
    - **Tokenization**
      * Tokenization is a pre-processing method which breaks a stream of text into words, phrases, symbols, or other meaningful elements called tokens.
      * EG: “I hate to study” => [“I”, “hate”, “to”, “study”]
    - **Stop Word Removal**
      * Text and document classification includes many words which do not contain important significance to be used in classification algorithms, such as {“a” “the”}
    - **Capitalization**
      * Text and document data points have a diversity of capitalization to form a sentence.
      * The most common approach for dealing with inconsistent capitalization is to reduce every letter to lowercase.
    - **Slang and Abbreviation**
      * Slang and abbreviation are other forms of text anomalies that are handled in the pre-processing step.
      * An **abbreviation** is a shortened form of a word or phrase which contains mostly first letters from the words, such as SVM which stands for Support Vector Machine.
      * **Slang** is a subset of the language used in informal talk or text that has different meanings such as “lost the plot”, which essentially means that they’ve gone mad.
      * A common method for dealing with these words is **converting them into formal language.**
    - **Noise Removal**
      * Most of the text and document data sets contain many unnecessary characters such as punctuation and special characters.
      * Critical punctuation and special characters are important for human understanding of documents, but it can be detrimental for classification algorithms
    - **Spelling Correction**
      * Spelling correction is an optional pre-processing step.
      * Typos (short for typographical errors) are commonly present in texts and documents, especially in social media text data sets (e.g., Twitter).
    - **Lemmatization**
      * Lemmatization is a NLP process that replaces the suffix of a word with a different one or removes the suffix of a word completely to get the basic word form (lemma).
* Feature Extraction
  + The common techniques of feature extractions are
    - Term Frequency-Inverse Document Frequency(TF-IDF)
    - Term Frequency (TF)
    - Word2Vec
    - Global Vectors for Word Representation (GloVe).

1. Explain the different classification techniques used in text categorization

[Lec36.pptx](https://docs.google.com/presentation/d/19kKAXZ1PnCvKtHspxE4RDjTuoefSq3Gx/edit#slide=id.g11e633baac4_0_402)

## 5.text summarization

1. What is text summarization? Which are the different types of text summarization?

Text summarization refers to the technique of shortening long pieces of text.

The intention is to create a coherent and fluent summary having only the main points outlined in the document.

There are two main types of how to summarize text in NLP:

* **Extraction-based summarization**

The extractive text summarization technique involves pulling keyphrases from the source document and combining them to make a summary.

The extraction is made according to the defined metric without making any changes to the texts.

Here is an example:

Source text: Joseph and Mary rode on a donkey to attend the annual event in Jerusalem. In the city, Mary gave birth to a child named Jesus.

Extractive summary: Joseph and Mary attend event Jerusalem. Mary birth Jesus.

* **Abstraction-based summarization**

The abstraction technique entails paraphrasing and shortening parts of the source document.

It can overcome grammar inconsistencies of extractive method.

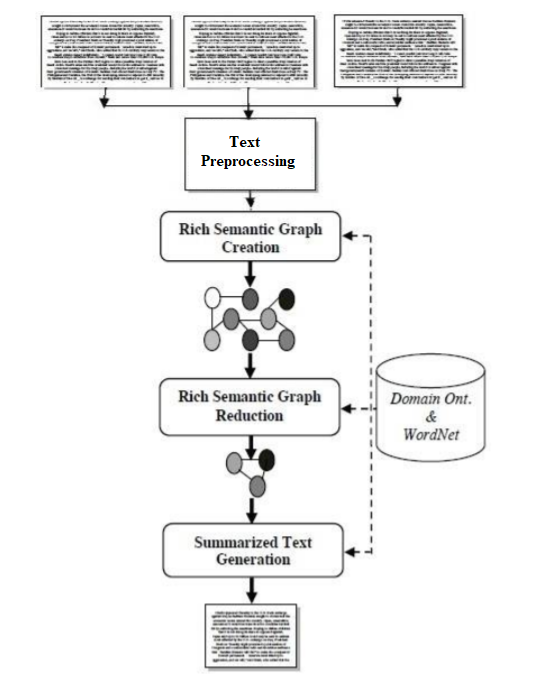
These algorithms create new phrases and sentences just like humans do.

The abstraction based text summarization algorithms are more difficult to develop.

Example:

Abstractive summary: Joseph and Mary came to Jerusalem where Jesus was born.

1. Explain with diagram general working of a text summarization system.



**Text Preprocessing Module**

Preprocessing module is responsible for accepting the input text, and converts it to preprocessed sentences.

It consists of four main processes: named entity recognition, morphological and syntactic analysis, cross-reference resolution, and pronominal resolution processes.

The **named entity recognition** process locates atomic elements into predefined categories such as person names, organizations, etc.

In **morphological analysis**, each word is divided into morphemes and figures out its grammatical categories, the syntactic analysis parses the whole sentence to describe each word syntactic function and build the parse tree, and typed dependencies expresses syntactic knowledge in terms of direct relationships between words.

Co-reference and pronominal resolution reference resolution processes identify co-reference named entities and resolve pronominal references in the whole input text.

Co-reference is defined as the identification of surface terms (words within the document) that refer to the same entity.

Rich Semantic Sub-graphs Generation Module

The Rich Semantic Sub-graphs Generation module is responsible to transform each preprocessed sentence to a set of ranked rich semantic subgraphs.

The main objective of the Rich Semantic Sub-graphs Generation module is to generate multiple rich semantic sub-graphs for each input preprocessed sentence.

This module includes three processes: Word Senses Instantiation, Concepts Validation, and Semantic Sentences Ranking processes.

Word Senses Instantiation process: For each input preprocessed sentence, this process instantiates a set of word concepts for both noun and verb senses based on the domain ontology.

Concept Validation Process: In this process, for each preprocessed sentence, the sentence concepts instantiated are interconnected and validated to generate multiple rich semantic sub-graphs.

Sentences Ranking Process: It aims to rank and to threshold the highest ranked rich semantic sub-graphs for each sentence. To generate single rich semantic graph and to keep the semantic consistency for the whole sentence, the process considers the first ranked rich semantic sub-graph only. The ranking method is based on deriving the average weight of each concept (word sense). The weight of the word concept is derived according to its usage popularity (Wordnet usage popularity).

The Rich Semantic Graph Reduction Phase

This phase aims to reduce the generated rich semantic graph of the original document to a more reduced graph.

In this phase, a set of heuristic rules are applied on the generated rich semantic graph to reduce it by merging, deleting, or consolidating the graph nodes.

The Text Generation Phase

The Rich Semantic Graph Generation module is responsible for generating a set of ranked RSGs for the input ranked semantic sub-graphs.

This phase aims to generate the abstractive summary from the reduced Rich Semantic Graph (RSG).

There are four modules namely the Text planning, the Sentence Planning, the Surface Realization, and the Evaluation modules.

These modules are performed by processes arranged as a pipeline, so the output of each process is the input of the next one as shown in figure 4.

1) The Text Planning module: It aims to select the appropriate content material to be expressed in the final text. This phase includes one process called “Content Determination”, which decides what information should be included in the generated text.

2) The Sentence Planning module: It specifies the sentence boundaries, and generates and orders intermediate paragraphs. The main objective of this phase is to improve the fluency or understandability of the text.

The sentence planning consists of four main processes:

1. Lexicalization Process: In this process, for each verb/noun object, its synonyms are selected by accessing the WordNet ontology to generate the target content.

2. Discourse Structuring Process: The main aim of this process is to build a structure that contains the selected object synonyms in the form of pseudo-sentences.

3. Aggregation Process: The main aim of this process is to decide how pseudo-sentences should be combined into semi-paragraphs.

4. Referring Expression Process: This process identifies and replaces the intended referent by its appropriate pronoun.

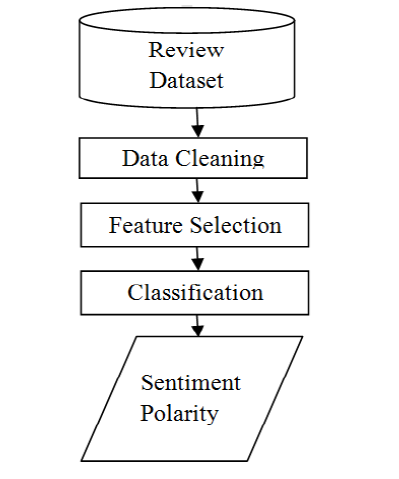
3) The Surface Realization module: This phase aims to transform the enhanced semi-paragraphs into paragraphs by correcting them grammatically (inflect words for tense, etc.) and adding the required punctuation (capitalization adding semicolon, etc).

4) The Evaluation module: The main objective of this phase is to evaluate and then rank the paragraphs according to two factors: coherence between paragraph sentences and the most frequently used paragraph word synonyms.

## 6.Sentiment analysis

1. Explain with diagram a general sentiment analysis system

Sentiment classification is a task under Sentiment Analysis (SA) that deals with automatically tagging text as positive, negative or neutral from the perspective of the speaker/writer with respect to a topic.



**Data Collection**

Data is collected from various social networking sites, blogging sites, and review sites

**Data Cleaning**

1. Removal of URL’s: Data extracted may contain some url’s which needs to be removed as they do not contain any sentiments.

2. Case conversion: All the text should be converted to either upper case or lower case i.e. there should be no difference between ‘paper’ and ‘PAPER’.

3. Removal of punctuation: Punctuation such as full stop, exclamatory sign, comma’s etc should be removed as they do not represent any emotions.

4. Removal of Hash tag: Hash tag word is preceded by a hash sign(#) and is generally used in social media for the identification of specific subjects.

**Feature selection** is a method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

Feature Selection methods can be divided into **lexicon-based methods** that need human annotation, and **statistical methods** which are automatic methods that are more frequently used.

Various feature selection methods are CountVectorizer ,TF-IDF(Term Frequency–Inverse Document Frequency), IG(Information Gain), MI(Mutual Information), Feature Vector, Unigram, Bigram and N-gram methods.

**Classification**

Sentiment Classification techniques can be roughly divided into

* **machine learning approach**- The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features.
* **lexicon based approach**- The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment erms. It is divided into **dictionary-based** approaches and **corpus-based** approaches which use statistical or semantic methods to find sentiment polarity.
* **hybrid approach**- The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

**Sentiment Polarity**

The final output obtained from Thus, a sentiment classifier tags the sentence ‘The movie is entertaining and totally worth your money!’ in a movie review as positive with respect to the movie.

On the other hand, a sentence ‘The movie is so boring that I was dozing away through the second half.’ is labeled as negative.

Finally, ‘The movie is directed by Nolan’ is labeled as neutral.

1. Explain the different feature selection methods in sentiment analysis

Feature Selection methods can be divided into **lexicon-based methods** that need human annotation, and **statistical methods** which are automatic methods that are more frequently used.

Statistical methods:

1. **Count Vector** : It is defined by the number of occurrences of features in review.

2. **TF-IDF** : It is defined by multiplying the value of frequency of word in review (TF) and frequency of word in whole corpus (IDF).**TF-IDFi = ti,j \* log (N/dfi)**. TF-IDFi is the weight of a term i. ti,j is the frequency of term i in sample j. N is the total number of samples in the corpus. dfi is the number of samples containing term i.

3**.Information gain** is the most widely used attribute selection measure in the area of sentiment analysis. It determines the relevant features to predict review by studying the presence or absence of features in a document.P(c|f) is the joint probability, class is c and feature is f and P(c) denotes the marginal probability.

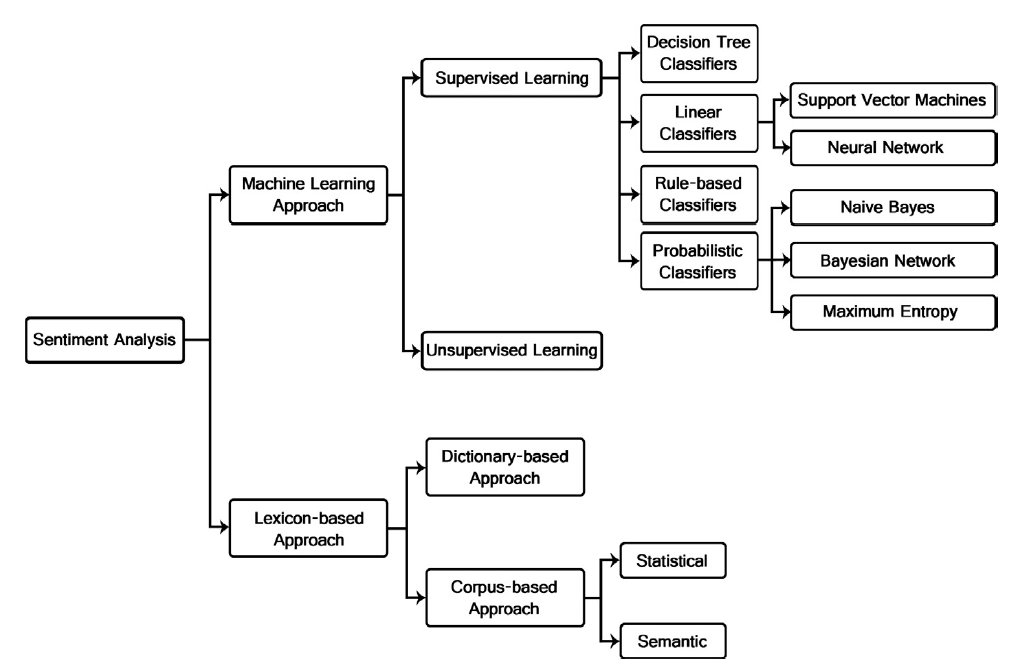
4.**Mutual Information**: MI is the process of selecting features that are not uniformly distributed across the sentiment classes because they are informative of their classes and we can see that MI gives more importance to only a few terms.

Where P(f,c) represents joint probability distribution function, P(f) and p(c) represent marginal probability distribution of f and c. c is positive and negative classes.

5.**Chi-square**: Chi-square measures observed count and expected count and analyzed how much deviation occurs between them.

W, X, Y, Z represents the frequencies, represent the presence or absence of feature in the sample. W is the count of samples in which feature f and c occurred together. N=W+X+Y+Z. f represents the feature and c represents the class.

1. Explain the classification of sentiment analysis system.



ML

The classification methods using ML approach can be roughly divided into **supervised and unsupervised learning methods**. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents.

***supervised***

The supervised learning methods depend on the existence of labeled training documents. There are many kinds of supervised classifiers.

1. Probabilistic classifiers. Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers. Three of the most famous probabilistic classifiers are

a. Naıve Bayes Classifier (NB). The Naıve Bayes classifier is the simplest and most commonly used classifier. Naıve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

b. Bayesian Network (BN). The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. Therefore, a complete joint probability distribution (JPD) over all the variables, is specified for a model. In Text mining, the computation complexity of BN is very expensive; that is why, it is not frequently used

c. Maximum Entropy Classifier (ME). The Max ent Classifier (known as a conditional exponential classifier) converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set. This classifier is parameterized by a set of X{weights}, which is used to combine the joint features that are generated from a feature-set by an X{encoding}. In particular, the encoding maps each C{(featureset, label)} pair to a vector. The probability of each label is then computed using the following equation:

2. Linear classifiers : A linear classifier uses a classification algorithm that makes its classification based on a linear predictor function combining a set of weights with the feature vector.

a. Support Vector Machine

The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories.

b. Neural Network

Neural Network consists of many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector Xi which is the word frequencies in the ith document. There are a set of weights A which are associated with each neuron used in order to compute a function of its inputs f(). The linear function of the neural network is: pi = A. Xi. In a binary classification problem, it is assumed that the class label of Xi is denoted by yi and the sign of the predicted function pi yields the class label.

c. Decision tree classifiers.

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification.

d. Rule-based classifiers

In rule based classifiers, the data space is modeled with a set of rules. The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data.

***Unsupervised***

The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, large number of labeled training documents are used for supervised learning, as illustrated before. In text classification, it is sometimes difficult to create these labeled training documents, but it is easy to collect the unlabeled documents. The unsupervised learning methods overcome these difficulties. Many research works were presented in this field . One of the method divides the documents into sentences, and categorized each sentence using keyword lists of each category and sentence similarity measure.

Lexicon Based Approach

**Dictionary Based**

A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well known corpora WordNet or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors. The dictionary based approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations

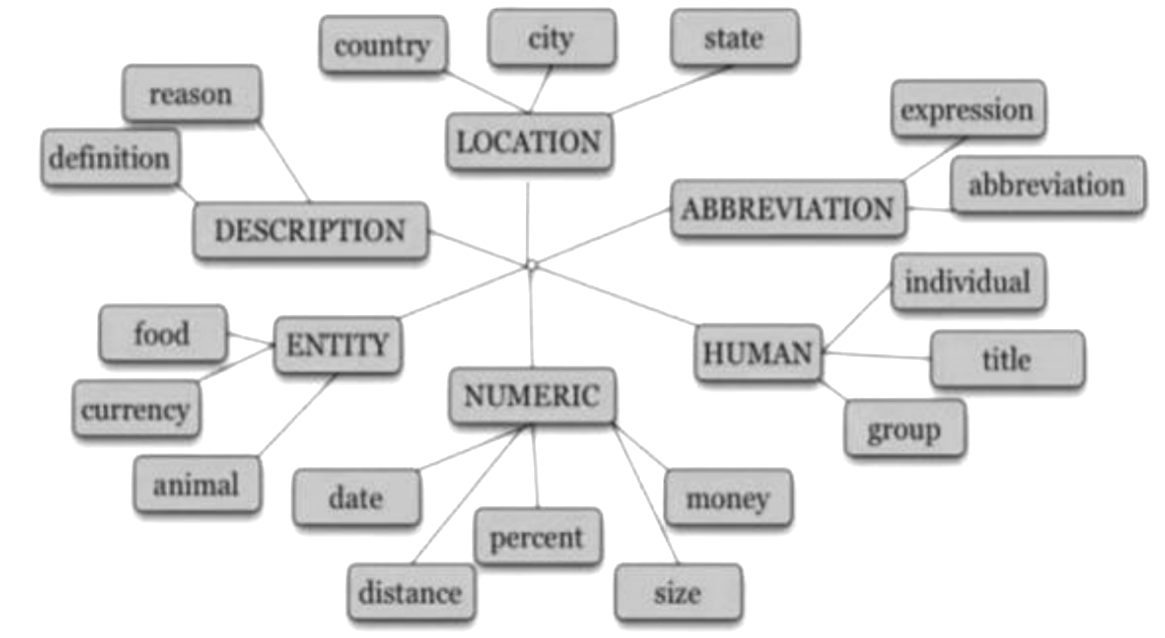
**Corpus Based**

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. One of these methods started with a list of seed opinion adjectives, and used them along with a set of linguistic constraints to identify additional adjective opinion words and their orientations.

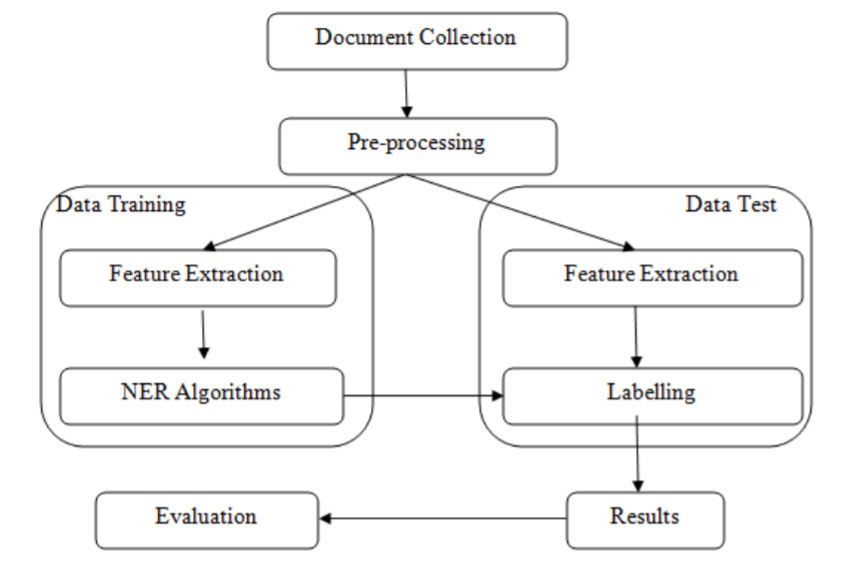
## 7.NER

1. Explain what is Named Entity Recognition. Give the different named entity categories.

* Entities are the who (and some of the what) of text analytics. On the most basic level, an entity in text is simply a proper noun such as a person, place, or product: John Coltrane, Coca Cola, and Indiana are all entities.
* Named Entity Recognition is a process where all the named entities which are the proper nouns are identified and classified into their predefined appropriate class.
* Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Thus it is the task of finding names such as organizations, persons, locations, etc. in text.



1. Explain with diagram the different steps in a Named Entity Recognition System.



**Step 1: Document collection**

Documents of varied formats such as .pdf, .html, .docx etc. from all sources will be collected. These documents will be inputs for the system.

**Step 2: Pre-processing**

Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure.

**Step 2.1: Validation of input document**

Validation is to check whether the given input text is in language for which the system is implemented. It also checks whether the input is syntactically correct, but does not check the semantic correctness.

**Step 2.2: Tokenization**

The aim of the tokenization is the exploration of the words in a Sentence where every word, symbol, special character in the sentence is considered as a token.

**Step 2.3: Stop word removal**

In stop word removal, words that occur very frequently and does not contribute much to the context and content, and also have no impact on their existence are removed.

**Step 2.4: Stemming**

Trimming or cutting out the extraneous words to the stem is called stemming. Here inflections are removed using stemming algorithms.

**Step 2.5: Morphological analysis**

Morphological analysis is the procedure to find out the root word. It is applied to recognize the inner structure of the word.

**Step 3: Data Training:**

This step is required to train the system. Training is done based on the feature extraction and the algorithm used. The output of this stage will be given to the testing stage.

**Step 3.1: Feature extraction** – In this process a small subset from the sentence is extracted and then a feature set is applied to the NER algorithms.

**Step 3.2: NER algorithms** – Various NER NLP algorithms include rule based, machine learning and hybrid approaches.

**Step 4: Data Testing**

**Step 4.1: Feature extraction** – This process is the same as explained in the training data stage with the test data. The extracted features are then tagged.

**Step 4.2: Labeling (tagging)** – In this process the entities are tagged using any of the algorithm.

**Step 5: Result** – the output of all the above stages will be then go through the evaluation stage using evaluation parameters.

**Step 6: Evaluation** – The accuracy level of NER can done by Precision (P), Recall (R) and F1-measure metrics.

1. Explain the different classification approaches for Named Entity Recognition.

[Lec38.pptx](https://docs.google.com/presentation/d/1Roxbh09J9vqBKhb2t_Bqs8PmV7IXG5bU/edit#slide=id.g11e633baac4_0_906)

NER task is done by the following rule based approaches. It is also called hand crafted rules or linguistic approach. In this approach the rules are written manually by the researchers for the system and for any particular language. Rule based systems parse the source text and produce an intermediate representation which may be a parse tree or some abstract representation. Rule based are further classified into

* list lookup approach
* linguistic approach

**List lookup Approach** : In List Lookup a large corpus which is called as a bag of words are built for all the named entities and their classes. List lookup is performed to identify named entities. This list is also called as gazetteer.

**Linguistic Approach**: In linguistic approach one should have a deep knowledge of the grammar of any specific language. The understanding and the knowledge of the language leads to more accurate rules so that the named entities will be identified and classified very easily.