**Explain the different approaches for Machine Translation**

Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Spanish).

Like translation done by humans, MT does not simply substitute words but the application of linguistic knowledge, morphology, grammar, meaning; all this has to be taken into consideration. Generally, MT system is classified into various categories: Direct Based, Interlingua Based, Knowledge Based, Principle Based, Statistical Based, Rule Based, Neural Based Machine Translation Systems.

a. Direct Translation

It is one of the simplest machine translation approaches in which a direct word to word translation is done with the help of a bilingual dictionary.

b. Interlingua Based Translation.

In this approach, the source language is transformed into an interlingua (an abstract language-independent representation). The target language is then generated from the interlingua.

c. Knowledge Based MT

It requires complete understanding of the source text prior to the translation into the target text. It is implemented on the Interlingua architecture. It is supported by the linguistic semantic knowledge about the meanings of words and the relationship among them.

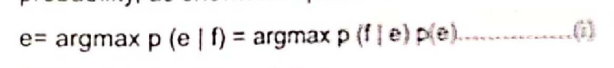
d. Principle Based MT

It is based on the Principles and Parameters of Chomsky’s Generative Grammar which employs the parsing method. In this, the parser generates a detailed syntactic structure which contains lexical, phrasal, grammatical information.

e. Statistical Based MT

Statistical Machine Translation is a type of machine translation where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. It uses computer algorithms that reference and analyse previously translated text and explore the millions of possibilities of how to order words and put smaller pieces of text together. The dataset consists of corpus of texts (e.g., United Nations Document) which have already been translated into multiple languages. The most common algorithm used is Bayes Theorem. In SMT, a document is translated according to the probability distribution function which is indicated by p(e|f). The highest probability that a certain word in the source language will be another word/phrase in the target language.

The following equation is used to calculate the probability



Systems Implementing STM :

* Tapta4UN is a statistical machine translation tool developed by the United Nations which is trained with documents from the UN to provide an output that is consistent with United Nations style and terminology.
* Google Translate, Microsoft Translator, SYSTRAN, Yandex used STM in their translation models till 2016 and then they all shifted to a new translation system.

           Advantages :

* Training for good corpora is automated and cheaper.
* It provides good quality when large and qualified corpora are available.

           Disadvantages :

* It has no knowledge of language or grammatical rules.
* It requires high CPU and disk space requirements.
* The translation is neither predictable nor consistent.

f. Rule Based Machine Translation.

Rules-Based Machine Translation (RBMT) systems were the first commercial machine translation systems and are based on linguistic rules that allow the words to be put in different places and to have different meanings depending on the context. RBMT technology applies to large collections of linguistic rules in three different phases: analysis, transfer, and generation. Rules are developed by human language experts and programmers who have deployed extensive efforts to understand and map the rules between two languages. RBMT relies on manually built translation lexicons, some of which can be edited and refined by users to improve the translation.

Process

* A parser analyses the grammatical input structure.
* The parser creates an intermediate representation of the message. This representation can have different levels of abstraction.
* The parser transfers that representation into the output language structure.

Strategies:

Direct approach - The source language input is translated word by word in the target language.

Transfer approach - The source language and the target language messages are transferred into intermediate representations. The difference with interlingual MT is that the transfer approach depends on the language pair involved.

Interlingual approach - The source language input is transformed into a semantic representation of the text, an interlingua. The interlingua is the basis for generating the target text.

For example:

*A girl gets an apple.*

source\_language = “english” , target\_language = “hindi”

           Minimally, to get a Hindi translation of this English sentence one needs:

* A dictionary that will map each English word to an appropriate Hindi word.
* Rules representing regular English sentence structure.
* Rules representing regular Hindi sentence structure.

  Systems Implementing RBMT:

          Japanese MT System & Eurotra.

          Advantages:

* Total control. Because all rules are hand-written, you can easily debug a rule-based system to see exactly where a given error enters the system, and why.
* Rules are usually written in a domain independent manner, so the vast majority of rules will "just work" in every domain, and only a few specific cases per domain may need rules written for them.

        Disadvantages:

* Building new dictionaries is expensive
* It is hard to deal with rule interactions in big systems with ambiguity.
* It lacks fluency.

g. NMT

Neural Machine Translation is a machine translation approach that applies a large artificial neural network towards predicting the likelihood of a sequence of words, often in the form of whole sentences. It uses a bidirectional recurrent neural network, also called an *encoder*, to process a source sentence into vectors for a second recurrent neural network, called the decoder, to predict words in the target language. This process not only proves to be compatible but it also achieves speed and accuracy for the same.

Systems Implementing NMT:

* One of the most popular translation machines in the world is Google Translate. The system uses Google Neural Machine Translation to increase its fluency and accuracy. The system not only applies a large data set for training its algorithms, its end-to-end design allows the system to learn over time and create better, more natural translations.
* Microsoft Translator, SYSTRAN, Yandex, Skype Translator & Baidu Translate.

Advantages:

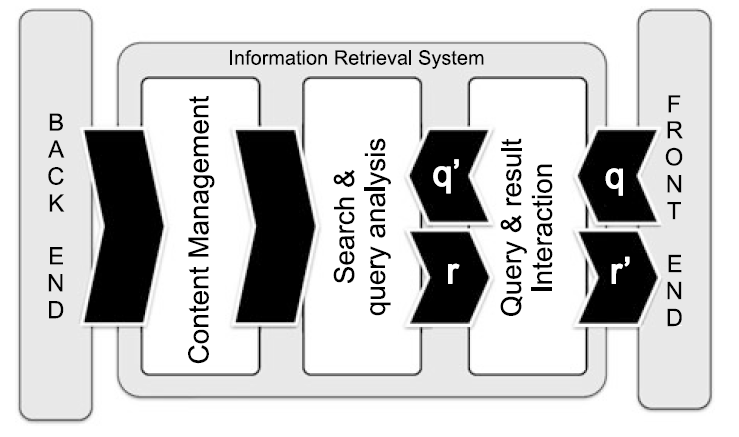
* The translation is more fluent compared to other methods.
* It enables better quality translations, in fact it contains 50% fewer word order errors, 17% fewer lexical errors and 19% fewer grammatical errors.

Disadvantages:

* It performs rather poorly for long sentences.

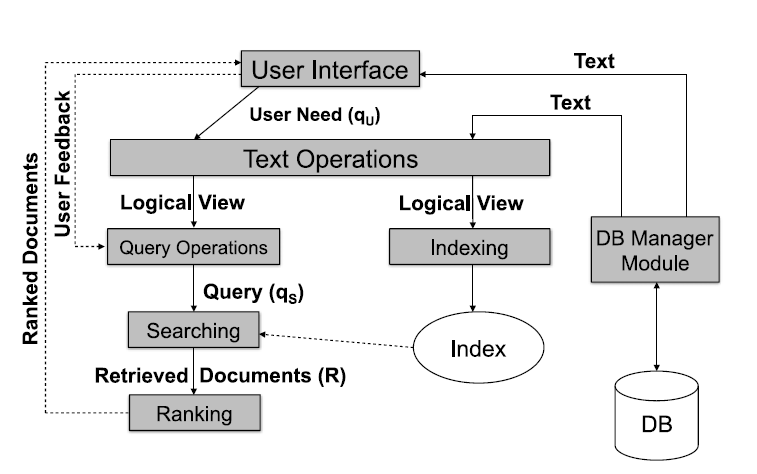
**Explain Information Retrieval System with Block Diagram and explain different steps in pre-processing.**

Information retrieval is the process of searching for relevant documents from unstructured large corpus that satisfy users information need. It is a tool that finds and selects from a collection of items/documents a subset that serves the user’s purpose. Traditionally, the information retrieval system techniques are based on keywords. For example, the search engine accepts keywords and in return they show a list of links to documents containing those keywords. Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents. The system searches over billions of documents. IR basically 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
* q is processed by a query interaction module that transforms it into a “machine-readable” query q’
* q‘ is  fed into a search and query analysis module.
* It has 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 expressed in natural language. The processing of textual queries typically performed by an IR engine are shown in figure.



1. The user need is specified via the user interface, in the form of a textual query qU (made of keywords).

2. 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 .

3. Then query is transformed into a system-level representation, qS .

4. The query qS is executed on top of a document source D to retrieve a set of relevant documents, R.

5. The set of retrieved documents R is then ordered: documents are ranked according to the estimated relevance.

6. The user then examines the set of ranked documents for useful information and provide feedback to the system.

Advantages :

* The retrieved information comes in several formats i.e. books, journals, PDFs, documents, format, etc.
* There is no geographical barrier to search for information from anywhere in the world.

Disadvantages :

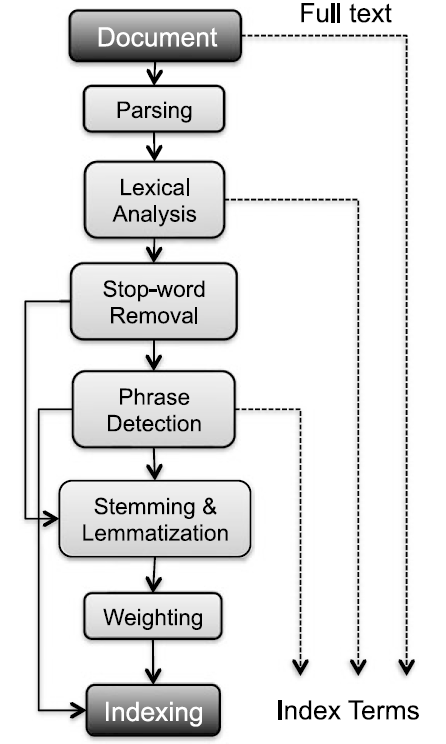
* Slow speed of internet and lack of networking facility can delay the retrieval system.
* Sometimes it gives irrelevant information.

**Text Processing in IR :**

* Not all words are equally effective for the representation of a document’s semantics.
* Noun words (words or noun phrase groups) are the most representative components.
* Based on this observation, IR system also pre-processes the text of the documents to determine the most “important” terms to be used as index terms
* When selecting candidate keywords, indexing must fulfil two different and potentially opposite goals:
  + Exhaustiveness: assign sufficient terms to a document
  + Specificity: the exclusion of generic terms that carry little semantics.
* 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.
* Generic terms have high term frequency, defined as the number of occurrences of the term in a document.
* Specific terms have higher discriminative power, due to their rare occurrences across collection documents: they have low document frequency.

The textual pre-processing phase typically performed by an IR engine, takes as input a document and yields its index terms as output.

The process of this extraction is given in the figure.



1. Document Parsing

* Documents come in all sorts of languages, character sets, and formats and the same document may contain multiple languages or formats.
* e.g., A French email with Portuguese PDF attachments.
* Document parsing deals with the recognition and “breaking down” of the document structure into individual components.
* In this pre-processing phase, unit documents are created;
* e.g., emails with attachments are split into one document representing the email and as many documents as there are attachments.

2. Lexical Analysis

* After parsing, lexical analysis tokenizes a document, seen as an input stream, into words.
* Issues related to lexical analysis include the correct identification of accents, abbreviations, dates, and cases.
* The difficulty of this operation depends much on the language at hand.

3. Stop Word Removal

* Stop-word removal is the removal of high-frequency words.
* This process helps to identify relevant keywords to serve as index terms.

4. Phrase Detection

* This step captures text meaning.
* Phrase detection may be approached in several ways, including
  + rules
  + morphological analysis
  + syntactic analysis, and combinations thereof.
* A common approach to phrase detection relies on the use of thesaurus
* Thesaurus usually contain synonyms and antonyms.
* Thesaurus may be composed following different approaches.
  + Human-made thesaurus:

They are generally hierarchical, containing related terms, usage examples, and special cases.

Other formats are the associative one, where graphs are derived from underlying WordNet’s synonym sets or synsets.

* An alternative to the consultation of thesaurus is to use machine learning techniques.
* Key Extraction Algorithm (KEA) identifies candidate key-phrases using lexical methods, calculates feature values for each candidate, and uses a supervised ML algorithm to predict which candidates are good phrases based on a corpus of previously annotated documents.

5. Stemming and Lemmatization

Stemming and lemmatization aim at stripping down word suffixes in order to normalize the word.

Stemming is a technique used to extract the base form of the words by removing affixes from them.

Lemmatization aims to remove inflectional endings only and to return the base form of a word.

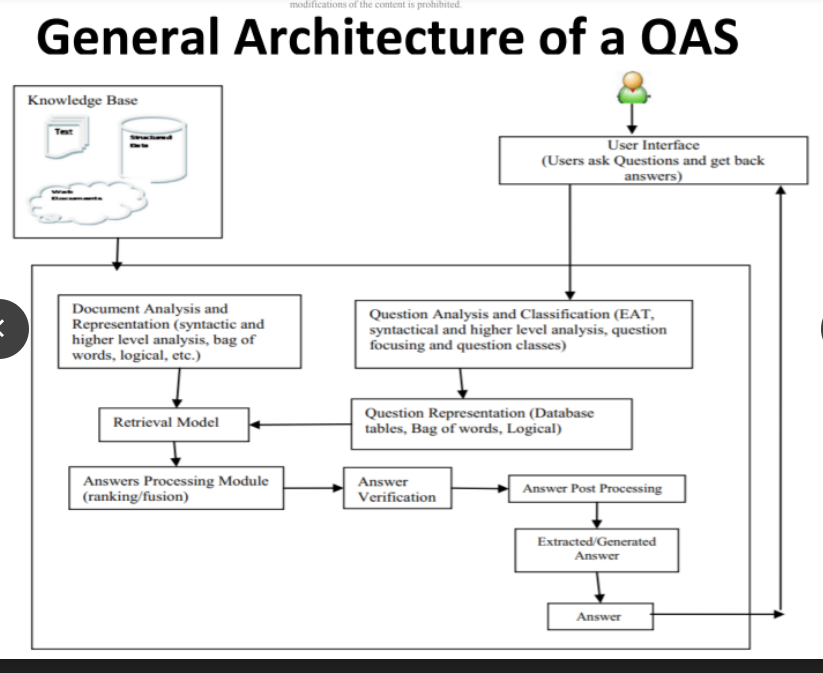
For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.

6. Weighting

* The final phase of text pre-processing deals with term weighting.
* The words in a text have different descriptive power; hence, index terms can be weighted differently to account for their significance within a document and/or a document collection.
* Such a weighting can be binary, e.g., assigning 0 for term absence and 1 for presence.

**3. Question Answering System.**

Question Answering System is concerned with automatically answering questions posed by humans in a [natural language](https://en.wikipedia.org/wiki/Natural_language). It retrieves and processes information from different data sources. The format of answers may also vary from simple text to multimedia.



A typical question answering system (QAS) consists of three main modules :

1. Question Analysis

2. Answer Retrieval

3. Answer Generation

1. Question Analysis

It takes the question as input from the user in natural language. It then aims to understand the purpose and meaning behind the question for e.g. location, date, person’s name, etc. The question is analysed in different ways.

Next a morpho-syntactic analysis of the words in the question is carried out by doing a POS tagging. After POS tagging, we find what information is the question looking for. A question class helps the system to classify the question type to provide a suitable answer.

To get the meaning of the question, the types of semantic in the question are classified.

For e.g. a question can seek for date, time, location or person. 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 questions description type must meet. This process is simple as taking out keywords from the question and find candidate answer sentences. These keywords may then be extended by using morphological and/or synonyms replacements or using query expansion techniques. They form representation of the question.

2. Answer Retrieval

It involves the following steps:

* Document Retrieval
* Document Processing
* Syntactic Analysis
* Semantic Analysis
* Relation Identification

2a. 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.

2b. Document Processing

* The retrieved documents are processed for extracting candidate answer set.
* This module is responsible for selecting the response based on the relevant fragments of the documents.

2c. 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.

2d. 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.

2e. Relation Identification

* The relations among different concepts are identified using the domain knowledge and the ontological information is obtained.

3. 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, Answer Ranking and Answer Generation.

Filtering : The extracted sentences are filtered and the candidate answer set is produced. This is done by incorporating the information obtained from the question classification and document processing modules. The identified focus and frames are matched to get the candidate set.

Answer Ranking: The answer set is ranked based on the semantic similarity. Answers are ranked based on the similarity between the question frame and the answer frame.

• Example: The event E “John gave a balloon to the kid.” has the roles “AGENT verb/give THEME to RECIPIENT, the semantic frame is identified as

• “has\_possession(start(E),Agent,Theme)

• has\_possession (end(E),Recipient, Theme)

• transfer(during(E), Theme)”

matches exactly with the question frame: “Who give balloon to the kid?”.

**Explain the classification of approaches in Question Answering System.**

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

1. Application domains for which QASs are developed

2. Types of questions asked by the users

3. Types of analyses performed on users’ questions and source documents

4. Types of data consulted in data sources

5. Characteristics of data sources

6. Types of representations used for questions and their matching functions

7. Types of techniques used for retrieving answers

8. Forms of answers generated by QASs

1. Classification Based on Application Domain

The task of generating answers of questions is related to the type of questions asked.

* General Domain : General Information on a general topic
* Restricted Domain : Specific Information from a particular application domain

General Domain :

* It searches for answers with a large document collection.
* It exploits world knowledge for generating answers.
* The quality of answers delivered is not high.

Pros :

* It uses a general dictionary.
* No domain knowledge required for formulating questions
* Wikipedia/News text can be utilized as a source

Cons :

* The quality of answers is low.
* The answers satisfaction depends upon the users.
* Domain experts may not find this suitable.

Restricted Domain :

* It answers domain specific questions.
* Answers are searched within domain specific document collections.
* The question patterns is very limited and hence the systems can achieve good accuracy.
* The quality of answers is expected to be high.

There are various restricted domain QASs developed :

* Temporal, Geospatial. Medical, Patent, Community Based QAS.

Pros :

* It suits domain experts as they need specialized answers.
* The quality of the answers generated is high as it is more clear and specific.

Cons :

* The repository is limited

2. Classification Based on Types of Questions

Categories :

* Factoid Type Questions.
* List Type Questions.
* Hypothetical Type Questions.
* Confirmation Questions.
* Casual Questions.

Factoid Type Questions :

* The questions are simple and fact based. It requires answers in a single short phrase/sentence.
* The factoid type questions generally start with wh-word.
* Current QASs have a satisfactory performance in answering factoid type questions.

List Type Questions :

* The list type of questions require a list of entities/facts in answers. E.g. – List name of employees present in Google Company earning more than 80K?
* It considers such questions as a series of factoid question asked more than 10 times.
* The previous answers are ignored while firing next questions by QASs.

Hypothetical Type Questions :

* It asks for information related to any hypothetical event.
* It generally begins with ‘what would happen if’.
* It requires knowledge retrieval techniques for generating answers.
* There are no specific correct answers to these questions.

Confirmation Questions :

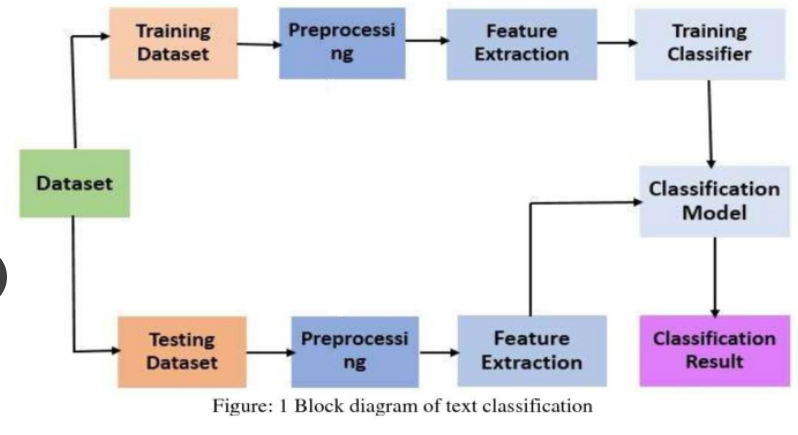
* It requires answers in the form of Yes or No.
* The system requires inference mechanism, world knowledge and common sense reasoning to generate answers.

Casual Questions

* It requires explanations about an entity.
* It requires advanced natural language processing techniques to analyze the text at pragmatic and discourse level for generating answers.

**Explain with diagram steps in Text Categorization/Classification.**

The goal of automatic text categorization is to assign a category to a document by analyzing or evaluating its text components. The dataset is split into training and testing for classification process.



Text Pre-Processing :

The main objective is to obtain key features/terms from stores text documents. Pre-Processing step is crucial in determining the quality of the classification. It is important to select the significant keywords that carry the meaning and discard the words that do not contribute for classification task. Pre-processing phase converts the original textual data into a data mining ready structure.

In general, text can be represented in two separate ways.

First is a bag-of-words, in which a document is represented as a set of words, together with their associated frequency in the document.

Second Method is to represent text directly as strings (a sequence of words).

1. Tokenization

- It is a pre-processing method which breaks a stream of text/string into words, phrases, symbols or other meaningful elements called tokens.

E.g. :

The sky looks so beautiful this evening.

The tokens are : { “The”, “sky”, “looks”, “so”, “beautiful”, “this”, “evening” }

2. Stop Work Removal

- It removes insignificant words that does not help during classification.

E.g. : “a”, “about”, “above”, “across” etc.

3. Capitalization

The most common approach for dealing with inconsistent capitalization is to reduce every letter to lower case.

Exception: “US” (United States of America) to “us” (pronoun)

4. Slang and Abbreviation

A common method for dealing with these words is to convert them into formal language.

5. Noise Removal

It removes punctuation and special characters.

6. Spelling Correction

It is an optional pre-processing step.

Typos (typographical errors) are commonly present in texts and documents, especially in social media text datasets (e.g. Twitter).

7. Stemming

In NLP, one word could appear in different forms (i.e. singular and plural noun form) while the semantic meaning of each form is the same.

For example, the stem of the word “studying” is “study”.

8. Lemmatization

It is 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 :

Common Techniques are :

* Term Frequency-Inverse Document Frequency (TF-IDF)

The words like “of”, “as”, “the” appear frequently in an English corpus. By taking IDF, we can minimize the weighting of frequent terms while making infrequent terms have a higher impact.

* Term Frequency (TF)

It means how often a term appears in a document.

* Word2Vec

A Word2Vec model learns meaningful relations across words/products and encodes the relatedness into vector similarity.

* Global Vectors for Word Representation (GloVe).

It is an unsupervised learning algorithm for obtaining vector representation for words.

Classification Algorithms :

* Dimensionality Reduction

It refers to the task of reducing the number of features in a dataset.

* Naïve Bayes Classifier

It is based on the Bayes Theorem. In simple words, it is a way of finding a probability when we know certain other probabilities.

* Support Vector Machine

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

* K-nearest Neighbor

It works on the principle of closest training samples. The data points that are close to each other belong to one particular class.

* Artificial Neural Networks

It is a computational network based on biological neural networks that construct the structure of the human brain. ANN also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes.

* Centroid Based Classifier

Given a data point, the model simply assigns it the class of the training sample whose mean/centroid is closest to it.

* Clustering

It is the task of finding groups of similar documents in a collection of documents.

Applications:

* As marketing is becoming more targeted every day, automated classification of users into cohorts can make marketer’s life simple. Marketers can monitor and classify users based on how they talk about a product or brand online. The classifier can be trained to identify promoters or detractors. Thus, making brands to serve the cohorts better.
* Academia, law practitioners, social researchers, government, and non-profit organizations can also make use of text classification technology. As these organizations deal with a lot of unstructured text, handling the data would be much easier if it were standardized by categories/tags.

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

• Abstraction-based summarization

* **Extraction Based Summarization**

The extractive text summarization technique involves pulling key-phrases 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.

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.

As you can see above, the important words have been extracted and joined to create a summary — although sometimes the summary can be grammatically strange.

* **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.

Text Summarization Algorithm : (Extraction-Based Approach)

1. Introduce a method to extract the merited key-phrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the key-phrases.

2. Gather text documents with positively-labeled key-phrases. The key-phrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled key-phrases.

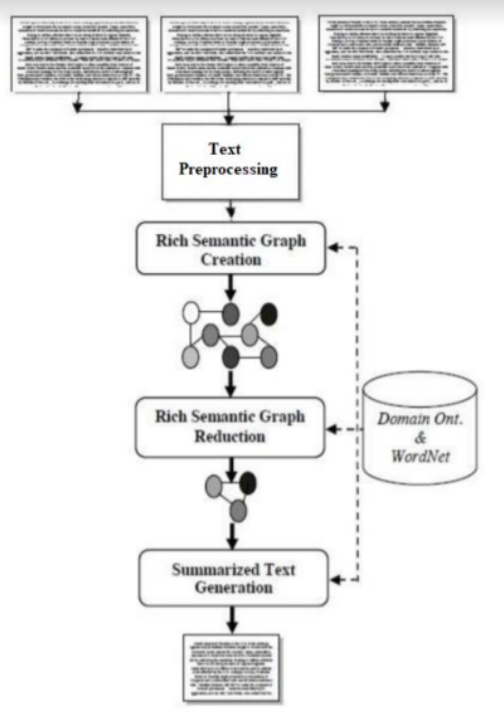
3. Train a binary machine learning classifier to make the text summarization.

Some of the features to include :

* Length of the key-phrase
* Frequency of the key-phrase
* The most recurring word in the key-phrase
* Number of characters in the key-phrase

Finally, in the test phrase, create all the key-phrase words and sentences and carry out classification for them.

**Text Summarization System**



The system uses the multiple documents in order to create abstractive summarization. At first, a semantic graph is generated for every sentence in the documents by pre-processing each sentence. Thereafter, the generated graph is reduced to a more reduced graph to generate abstractive summary. Heuristic rules have been used to generate an abstractive summary. The goal of the system is to condense the documents into a shorter version and preserve important contents.

Text Pre-Processing Module: It is responsible to accept the input text and converts it to pre-processed 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 element 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 express 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 main objective is to represent the input documents semantically using Rich Semantic Graph (RSG). It is able to capture the meaning of the words, sentences and paragraphs. This module is responsible to transform each pre-processed sentence to a set of ranked rich semantic subgraphs. It consists of three processes : Word Senses Instantiation, Concepts Validation and Semantic Sentences Ranking processes.

Word Senses Instantiation Process :

For each input pre-processed 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 pre-processed sentence, the sentence concepts instantiated are interconnected and validated to generate multiple rich semantic sub-graphs.

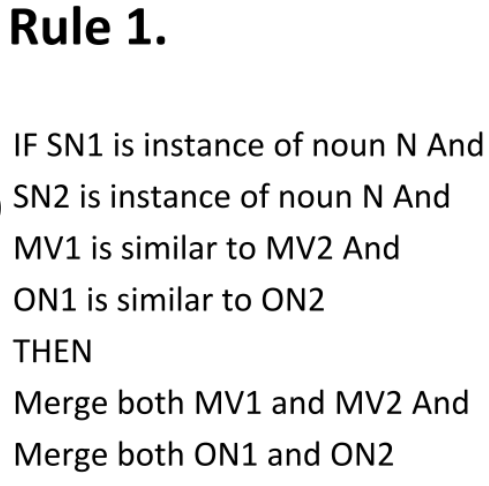
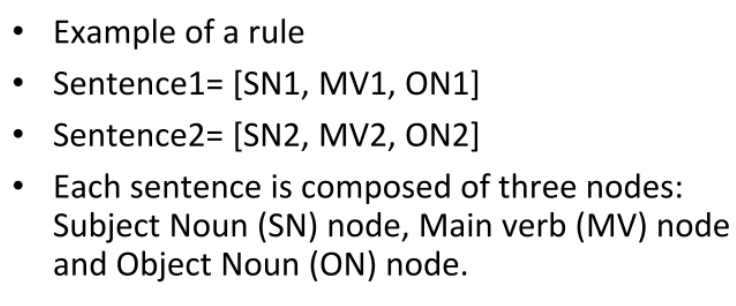
Sentences Ranking Process :

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).

Finally, the Rich Semantic Graph Generation module is responsible to generate the final rich semantic graphs of the whole input document from the highest-ranked rich semantic sub-graphs of the document sentences. The semantic sub-graphs of the input document will be merged to form the final rich semantic graph.

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.



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.

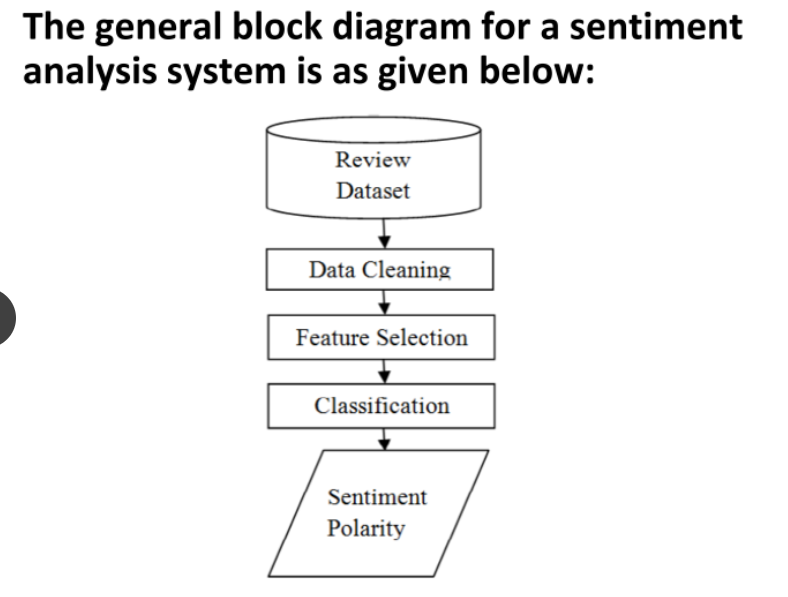
3) The Surface Realization module: This phase aims to transform the enhanced semi-paragraphs into paragraphs by correcting them grammatically (inflect words for tense, etc.) and adding the required punctuation (capitalization adding semicolon, etc).

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.

Sentiment Analysis System :

It is an automated system that deals with the textual data. It determines the context or meaning from the text data. It figures out whether the text is a positive or of a negative sentiment.

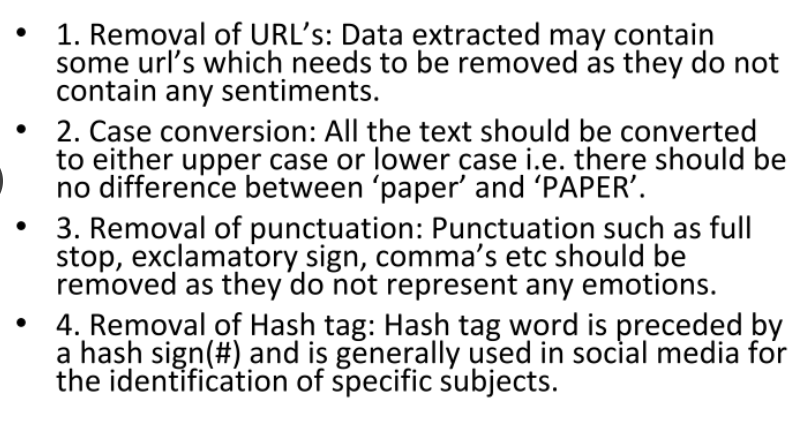
Thus, a sentiment classifier tags the sentence ‘The movie is entertaining and totally worth your money!’ in a movie review as positive with respect to the movie. On the other hand, a sentence ‘The movie is so boring that I was dozing away through the second half.’ is labelled as negative. Finally, ‘The movie is directed by Nolan’ is labeled as neutral.

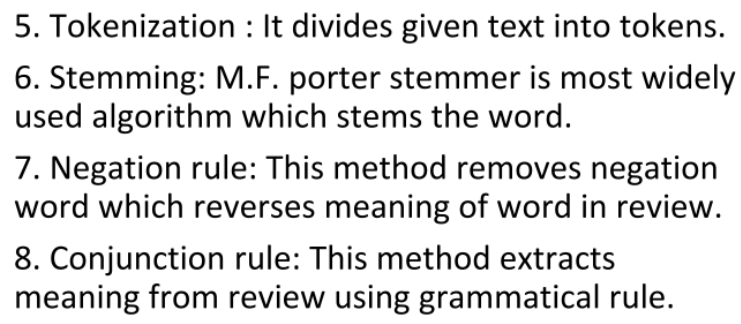


Data Collection :

Data is collected from various social networking sites, blogs, news, etc.

Data Cleaning :





Feature Extraction and Selection :

Some of the feature extraction techniques are :

1. Terms Presence and Frequency : It is based on individual word/n-gram and frequency counts.

2. Part of Speech (POS) : It extracts adjective nouns from data.

3. Opinion Words and Phrase : It is based on words which represents opinions such as good/bad, like/hate

4. Negation: It may reverse the meaning of words. For example : “not good” is equal to “bad”.

Some of the feature selection techniques are :

1. CountVectorizer : It is used to convert a collection of text documents to a vector of token counts.

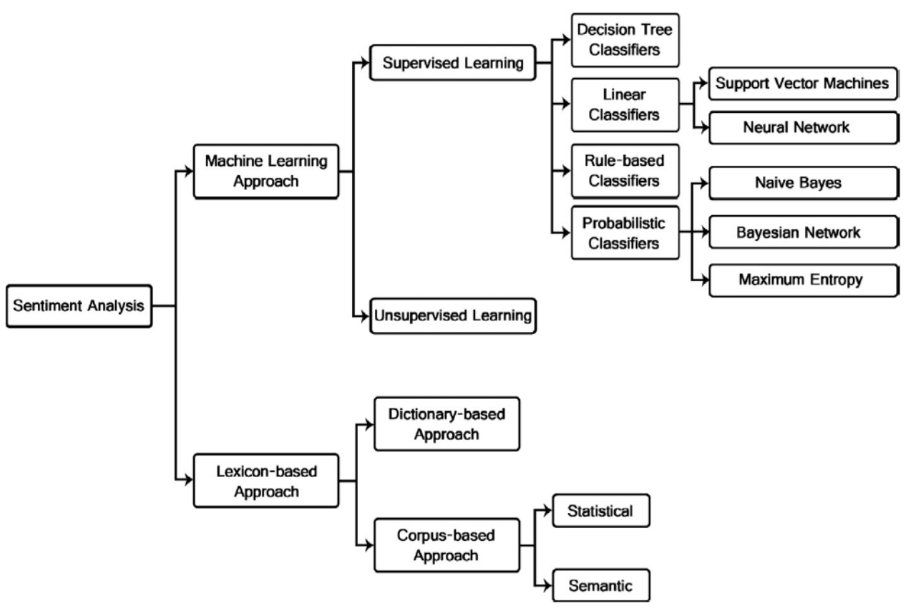
2. TF-IDF : It quantifies the importance of words in a document amongst a collection of documents

3. Chi-Square : It measures observed count and expected count and analyses how much deviation occurs between them.

4. Information Gain : The most widely used selection technique. It determines the relevant features to predict, by studying the presence/absence of features in a document.

5. Mutual Information : It is the process of selecting features which are not uniformly distributed in the sentiment classes.

Classification of Sentiment Analysis :



* *Add ‘hybrid based approach’*

Sentiment Analysis techniques is roughly divided into *machine learning approach, lexicon based approach and hybrid approach.*

Machine Learning Approach :

The machine learning approach applies algorithms and uses linguistic features. The ML approach is further classified into Supervised and Unsupervised learning. Supervised Learning makes use of labeled documents for training and Unsupervised Learning processes unlabelled or raw data. The supervised learning method is further classified into various supervised classifiers.

* Decision Tree Classifiers: It needs to be trained on some annotated data. Given a set of documents, together with their labels, the algorithm will calculate how much each word correlated with a particular label. For ex : It might find that the word “excellent” often appears in documents labeled as positive, whereas the word “terrible” mostly appears in negative documents.
* Linear Classifiers: It classifies data into labels based on a linear combination of input features. Therefore, these classifiers separate data using a line or plant or a hyperplane. They can only be used to classify data that is linearly separable. It consists of *Support Vector Machine (SVM)* and *Neural Networks*

SVM: It determines linear separators in the search space which can best separate the different classes. Text data are ideally suited for SVM because it is generally organized into a linearly separable category.

Neural Networks: Neural networks are made of units called neurons. The inputs to the neurons are word frequencies in the document. LSTM and CNN are the common algorithms used for classification. LSTM outperforms CNN in comparison for sentiment analysis classification.

* Rule Based Classifiers: It classifies from a set of rules created. The conditions are on the present terms and not on absent terms.
* Probabilistic Classifiers: It is able to predict, given an observation of an input, a probability distribution over a set of classes, rather than only outputting the most likely class that the observation should belong to. The following are the three most popular prob-classifiers :

Naïve Bayes: It depends on the principle of Bayes Theorem. It computes the posterior probability of a class, based on the distribution of words in the document.

Bayesian Network: It is a Directed Acyclic Graph (DAG) whose nodes represent random variables. It assumes that the features are independent of each other.

Maximum Entropy Classifier: Unline BN, it does not assume that the features are independent of each other. It is a conditional probability model.

Lexicon Based Approach:

This approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach : Dictionary & Corpus Based Approach.

Dictionary Based Approach: A small set of opinion words is collected manually with known orientations. Them, this set is grown by looking for their synonyms and antonyms in the well known corpora WordNet/Thesaurus. The added words are then put into list and next iteration is started right after. This iterative process stops when no words are found and errors are corrected later if any. It has a limitation with domain and context specific orientations.

Corpus Based Approach: It solves the problem faced by Dictionary based approach i.e. finding opinion words with context specific orientations. Its method 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. The final result is to generate two sets of words : positive or negative.

Hybrid Based Approach :

This approach combines both ML and Lexicon approach. It is the most modern, efficient and widely used approach for sentiment analysis.

Advantages of SA :

* For any organization or a company, the attitude of the customers toward their products or them is very necessary. Sentiment Analysis helps in analyzing the emotions from the reviews present in the companies products which helps the company to create better services and user experience. All of which adds up to increased sales and revenue.

Disadvantages of SA :

* SAs would find it difficult to analyze sentences that are sarcastic, exaggerated or jokes. Failing to recognize them can skew the results.

Applications of SA :

* Social Media Monitoring
* Product Analysis
* Brand Monitoring
* Analysis of Customer Feedback

**Named Entity Recognition :**

Named Entity Recognition (NER) is an AI technique that automatically identifies key information in a text, like names of people, places, companies, etc. It is also known as entity identification, entity chunking and entity extraction.

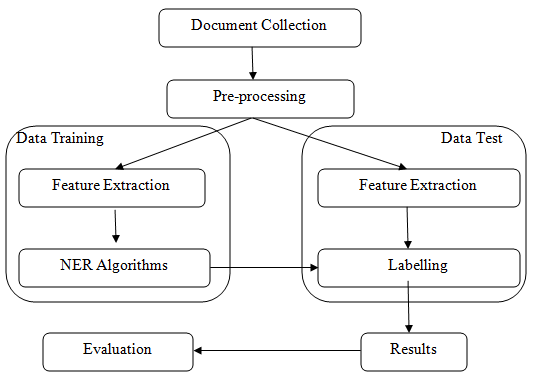
It basically detects the entities from the text and classifies them into different categories. For example, “Mark Zuckerburg is the CEO of Facebook”, in this sentence the entites are “mark”, “ceo” and “facebook”. These entities belong to the categories “person”, “title” and “company” respectively.

It is solved by using methods such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF).

The following are the different type of categories :

Person, Location, Organization, Event, Product, Skill, Address, PhoneNumber, Email, URL, IP.

Block Diagram :



Steps in NER:

1. Document Collection: Documents of different formats such as .pdf, .html, .docx, etc are collected from all sourced. These documents are then passed as input to the system.

2. Pre-Processing: It includes any type of preprocessing which might be required before sending the data for further processing.

- Validation: It checks if the language of the input text is same as the language of the system implemented at hand. It also checks for syntactic structures but ignores the semantic correctness of it.

- Tokenization: The aim of this step is to create tokens for every word, phrases present in the sentence.

- Stop Word Removal: The words in the documents that appear frequently and do not contribute to the context and do not help the system in identifying entites are removed. For example : “on”, “it”, “the”, etc.

- Stemming:

It is the process of keeping the root word and removing its suffix word. For example :

Shielding is reduced to Shield.

* Morphological Analysis

It is the study of the way words are built up from smaller meaning-bearing units, morphemes. For example, the word “fox” consists of a single morpheme (the morpheme fox) while the word cats consists of two: the morpheme cat and the morpheme -s.

3. Data Training :

The system is trained in this step by using an algorithm. The output of this stage will be given to the testing stage.

* Feature Extraction : A small subset from the sentence is extracted and then a feature set is applied to the NER Algos.
* NER Algos : Various NLP Algos are applied including Rule Based, ML and Hybrid Approach.

4. Data Testing

- Feature Extraction : The extracted features are tagged

- Labeling (tagging) : The entities are tagged using an algorithm.

5. Result

- The output of all the above stages will go through an evaluation stage.

6. Evaluation

- The accuracy level is calculated with Precision, Recall and F1-measure metric.

**Classification Approaches for NER :**

1. Rule Based Approach

The rules are written manually by the researchers for the system and for any particular language. This system parses the source text and produce an intermediate representation which may be a parse tree/some abstract representation. It is classified into List Lookup and Linguistic Approach.

List Lookup : A large corpus is built for all named entities and their classes. List lookup is performed to identify named entities.

Linguistic Approach : One needs a deep knowledge of the grammar of any specific language. The understanding of language leads to more accurate rules due to which classification becomes easy.

2. ML Approach

In this the system looks for patterns and relationships in text. Three types of ML models that are used for NER are : Supervised, Semi-Supervised and Unsupervised.

Supervised Learning : Uses only labelled data to generate a model.

Semi-Supervised Learning : It combines both labelled data and useful evidence from unlabelled data.

Unsupervised Learning : It is designed to be able to learn without or with very few labelled data.

3. Hidden Markov Model (HMM)

It is a statistical model. It computes the likelihood of a sequence of words by employing a Markov Chain, in which the likelihood of the next word is based on the current word.

4. Maximum Entropy Markov Model (MEMM)

It is a graphical model for sequence labelling that combined features of Hidden Markov Models (HMMs) and Maximum Entropy (MaxEnt) models.

5. Conditional Random Fields

It predicts a label for a single sample without considering neighbouring samples.

