Applications (Preferably for Indian Regional Languages)

6

OBJECTIVES

After reading this chapter, the student will be able to understand:

- Machine translation
- Information retrieval
- Question answers system
- Categorization
- Summarization
- Sentiment analysis
- Named Entity Recognition

1. Machine Translation

Machine Translation (MT) is a standard name for computerized systems responsible for the production of translations from one natural language into another with or without human assistance. It is a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one language to another. Development of a fully-fledged bilingual machine translation (MT) system for any two natural languages with limited electronic resources and tools is challenging and demanding task. In order to achieve reasonable translation quality in open source tasks, corpus-based machine translation approaches require large number of parallel corpora that are not always available, especially for less resourced language pairs. On the other hand, the rule based machine translation process is extremely difficult and fails to analyze accurately a large corpus of unrestricted text. Even though there has been an effort towards building English to Indian Language and Indian Language to Indian Language translation system, we do not have an efficient translation system as of today.

Why should we interest in using computers for translation at all? The first most important reason is that there is too much that needs to be translated and human translator cannot cope. A second reason is that technical materials are too boring for human translators, they don't like translating them, so they will look for help from computers. Another reason is that, in corporation area, there is a major requirement that terminology is used consistently; Computers are consistent but human translator tend to seek variety; they do not like to repeat the same translation and it will be not good for technical translation. Computer based translation can increase the volume and speed of translation and corporate area like to have translations immediately, the next day or even the same day.

Machine Translation System Approaches

Like translation done by humans, MT does not simply substitute words nut 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, Rule based, Corpus based, Statistical based, Hybrid based, Example based, Knowledge based, Principle based and Online Interactive based systems .

1. **Direct Translation**

   Direct Machine Translation is one of the simplest machine translation approach in which a direct word to word translation is done with the help of a bilingual dictionary.

2. **Rule Based Translation**

   A Rule-Based Machine Translation (RBMT) system consists of a collection of various rules, called grammar rules, a bilingual lexicon or dictionary, and software programs to process the rules.

3. **Interlingua Based Translation**

   In this approach, the translation consists of two stages, where the source Language (SL) is first converted into the Interlingua (IL) form. The main advantage of Interlingua approach is that the analyzer and parser of SL is independent of the generator for the Target Language (TL) and this requires complete resolution of ambiguity in source language text.

4. **Statistical-based Approach**

   Statistical machine translation (SMT) is a data-oriented statistical framework which is based on the knowledge and statistical models which are extracted from bilingual corpora. In this MT, bilingual or multilingual corpora of the languages are required. In SMT, a document is translated according to the probability distribution function which is indicated by p(e|f). Finding the best translation is done by picking the highest probability, as shown in Equation 1.

   $$e = \text{argmax}\ p(e \mid f) = \text{argmax}\ p(f \mid e)\, p(e) \quad \ldots\ldots\ldots\ldots (i)$$

5. **Example-based translation**

   Basic idea of this MT is to reuse the examples of already existing translations. An example-based translation is uses a bilingual corpus as its main knowledge base and it is essentially translation by analogy.

6. **Knowledge-Based MT**

   Knowledge-Based Machine Translation (KBMT) requires complete understanding of the source text prior to the translation into the target text. KBMT is implemented on the Interlingua architecture. KBMT must be supported by world knowledge and by linguistic semantic knowledge about the meanings of words and their combinations.

7. **Principle-Based MT**

   Principle-Based Machine Translation (PBMT) Systems are based on the Principles & Parameters Theory of Chomsky's Generative Grammar and which employs parsing method.

   In this, the parser generates a detailed syntactic structure which contains lexical, phrasal, grammatical information.

8. **Online Interactive Systems**

   In this online interactive translation system, the user has authority to give suggestion for the correct translation. This approach is very useful, where the context of a word is not that much clear or unambiguous and where multiple possible meanings for a particular word.

9. **Hybrid-based Translation**

   By taking the advantage of statistical MT and rule-based MT methodologies, a new approach was developed, which is termed as "hybrid-based approach". The hybrid approach used in a number of different ways. Translations are performed in the first stage using a rule-based approach which is followed by adjusting or correcting the output using statistical information. Second way in which rules are used to pre-process the input data and for post-process the statistical output of a statistical-based translation system.

Rule Based Machine Translation System

A Rule based Machine Translation (RBMT) System consist of collection of Rules, which is called as grammar rules, a bilingual or multilingual lexicon and software programs to process the rules.

Building RBMT system entails a huge human effort to code all of the linguistic resources, such as source side part-of-speech taggers and syntactic parsers, bilingual dictionaries, source to target transliteration, target language morphological generator, and structural transfer, and so many reordering rules. A RBMT system is always extensible and maintainable. Rules play a major role in various stages of translation, such as syntactic processing, semantic interpretation and contextual processing of language. Generally rules are written with linguistic knowledge gathered from linguistics. Transfer based MT, Interlingua MT and direct based MT are three different approaches that come under the RBMT category.

In the case of English to Indian languages and Indian languages to Indian languages MT system, there have been fruitful attempts with all these approaches. The main idea behind these rule-based approaches is as follows

1. **Direct Translation:**

   In Direct Translation method, source language (SL) text is analysed structurally up to the morphological level and is designed for a specific source and target language pair. The performance of direct based MT systems depends on the quality and quantity of the source-target language dictionaries, morphological analysis, text processing software and word-by-word translation with major grammatical adjustments on word order and morphology.

2. **Interlingua Translation:**

   The next stage in the development of MT systems is the Interlingua approach, where translation is performed by first representing the SL text into an intermediary (semantic) form called Interlingua. In this approach, the analyzer of parser for the SL is independent on the generator of TL. But this have certain disadvantages that, difficulty in defining the Interlingua and Interlingua does not take advantage of similarities between the languages.

3. **Transfer Based Translation:**

   Because of the disadvantages of the Interlingua approach, a rule-based translation approach was discovered, called the transfer approach. On the basis of the structural difference between the source and target language, a transfer system can be broken down into three different stages: i) Analysis ii) Transfer and iii) Generation.

Example: English to Marathi Machine Translation System

The idea is to translate an Input document in English Natural Language to Marathi language document by going through various phases such as pre-processing, lexical phase , syntax phase , semantics phase and finally translating into target language using various mapping rules.

The translation process has 2 major stages involved in it as :

1. Decoding the meaning of source text and
2. Re-encoding this meaning in the target language.

Like translation done by humans, MT does not simply involve substituting words in one language for another, but the application of complex linguistics knowledge: morphology, syntax, semantics and understanding of concepts such as ambiguity.

Proposed Machine translation System consists of 3 major phases:

1. **Pre-Processing Phase:**

   This is the first phase of any machine translation system. This phase makes machine translation process easier and qualitative. The source language text may contain figures, flowcharts, etc that do not require any translation. So only textual portion from the document should be identified. Fixing up punctuation marks and blocking content which does not require translation are also done during pre-processing module. Pre-processing phase consists of main processes as Morphology analysis, Named Entity Recognition (NER) and syntactic analysis. The efficiency of the MT system is always dependent on the quality of pre-processing stage.

2. **Transfer and Generation Phase:**

   This is the second phase of machine translation system. This module composes the meaning representations and assigns them the linguistic inputs. The semantic analyser uses lexicon or dictionary and grammar to create context dependent meaning. The source of knowledge consist of meaning of words, meaning associated with grammatical structures, knowledge about the discourse context and common sense knowledge

3. **Post-Processing Phase:**

   This is the last and most important phase of any machine translation process. Once the text is translated the target text is to be reformatted after post-editing. Post editing is done to make sure that the quality of the translation is upto the mark. Post-editing is unavoidable especially for translation of crucial information.

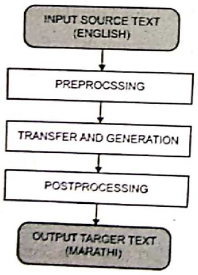


Figure 1: Proposed Architecture of English-Marathi Rule Based Machine Translation System

Algorithm of English to Marathi MT System:

Input: Digital document as input in English Natural language.

Output: Translated Document in Marathi Natural language.

1. Accept a digital document as input.
2. For each sentence in input document apply POS tagging.
3. Generate the parse tree for each sentence.
4. Apply NER on each sentence.
5. Identifies and names thematic relations between a verb and a noun in the sentence & Represents the tokens in the intermediate language based on target language reordering rules.
6. Use a bilingual dictionary to obtain appropriate translation & transliteration of the lemmas.
7. Obtain the proper form of words using inflections.
8. Represent sentences into target language.

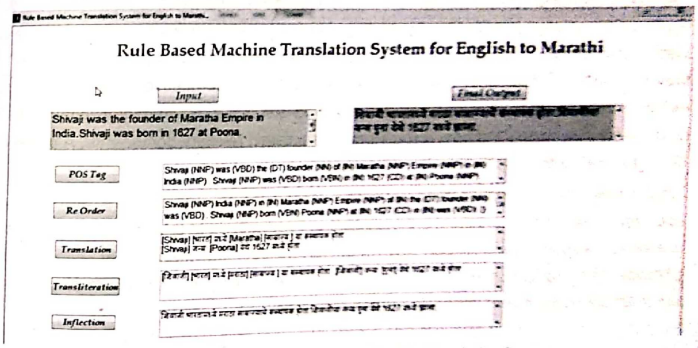


Figure 2: English-Marathi Rule Based Machine Translation System

Language technologies in general and machine translation system in particular are the instruments for narrowing the digital divide and overcoming language barriers to facilitate communication and commerce.

2. Information Retrieval

Information Retrieval remains one of the most challenging problems in NLP. Hundreds of millions of people engage in information retrieval everyday while using a web search engine or searching emails. Traditional database searching is becoming obsolete since most of the times the user is unclear what he or she is searching for. Information retrieval system is one that searches a collection of natural language documents with the goal of retrieving exactly the set of documents that pertain to a user's question"

A lot of readily available information is available through the World Wide Web which gets updated every time and is reached out to people all over the world. Thus, searching applications has changed tremendously from systems designed for specific applications belonging to well defined group to systems which are applicable for common people. However the huge and undefined structure of information present over networks have made it difficult for users to search and find relevant information. Many information retrieval techniques have been developed to deal with this problem.

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

Traditionally, the information retrieval system techniques are based on keyword. They use lists of keywords to describe the content of information but they do not say reveal semantic relationships between keywords nor consider the meaning of words and phrases. For example, the search engines accept keywords and in return they show a list of links to documents containing those keywords.

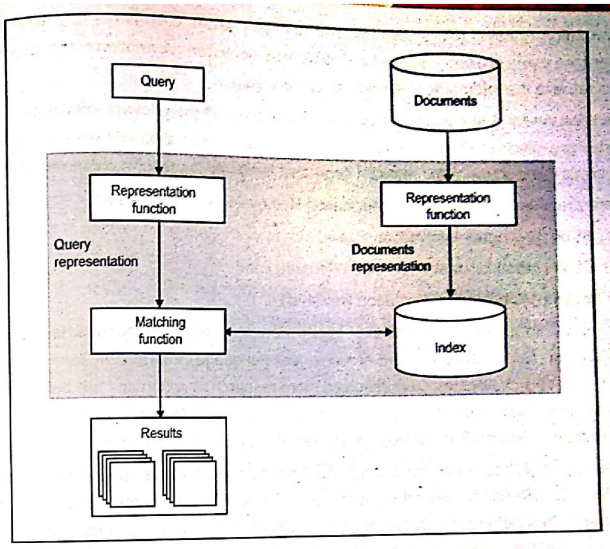


Figure 3: Basic IR system

Basic IR system involves following stages:

1. Indexing the collection of documents.
2. Transforming the query in the same way as the document content is represented.
3. Comparing the description of each document with that of the query.
4. Listing the results in order of relevancy.

In general all information Retrieval Systems consist of mainly two processes as:

1. **Indexing:** Indexing is the process of selecting terms to represent a text which involves tokenization of string, removing frequent words and stemming.
2. **Matching**: Matching is the process of computing a measure of similarity between two text representations. Relevance of a document is computed based on parameters like term frequency and inverse document frequency .

Information retrieval (IR) system aims to retrieve relevant documents to a user query where the query is a set of keywords. The ultimate goal of an information retrieval process is to retrieve the information relevant to a given request.

The criterion for a successful IR system is to retrieve all the relevant information items stored in a given system, and to reject the irrelevant information. However, the results will contain a mixture of the relevant items and irrelevant items for a given request.

There are many variants of the Information Retrieval systems such as :-

1. **BLIR** (Bi-Lingual Information Retrieval)
2. **CLIR** (Cross-Lingual Information Retrieval) and
3. **MLIR** (Multilingual Information Retrieval).

The ability to search and retrieve information in multiple languages is becoming challenging. So multilingual and cross-lingual (language) information retrieval (MLIR and CLIR) search engines have received more research attention and are being used to retrieve information on the Internet on a large scale. CLIR refers to searching, translating and retrieving information between a source language and target language.

Cross-lingual IR has become more important in recent years. Cross Language Information Retrieval (CLIR) can be described as the task of retrieving documents across languages. The basic idea behind the cross-lingual IR(CLIR) is to retrieve documents in a language different from the language used by the user to develop the query. This may be desirable even when the user is not a speaker of the language used in the retrieved documents. CLIR uses different translation approaches to translate queries to documents and indexes in other languages. Some CLIR systems use language resources such as bilingual dictionaries to translate the user's original query, while other systems use machine translation to translate the foreign-language documents beforehand, enabling them to be retrieved by the original query.

This task represents one extreme case of the vocabulary mismatch problem, i.e. The vocabulary of a user query and the vocabulary of relevant documents can differ substantially. The bag-of-words (BOW) model seriously suffers from the vocabulary mismatch problem because of different dimensions thus neglecting relations between different words in the same language as well as across languages. Therefore, the challenging task of retrieving documents to queries in other languages requires models other than traditional bag-of-words model.

3. Question Answering System

Humans are always in a quest to extract information related to some topic or entity. Question answering system helps user to find the precise answer to the question articulated in natural language. Question answering system provides explicit, concise and accurate answer to user questions rather than providing a set of relevant documents or web pages as answers as most of the information retrieval system does.

Any question answering system basically consists of three parts as question processing, answer retrieval and answer generation. In question processing users natural language questions are parsed to formulate questions in machine readable form using different approaches. Then in answer retrieval candidate answers are extracted based on intermediate representation of question. Finally in answer generation phase user understandable precise and accurate answer is generated and provided to user.

Any question answering system can be classified into two main types: close domain QAS and open domain QAS. In close domain QAS scope of user question is limited to a particular domain like medicine, movies, history and others. So if a QAS is created for history for history domain it will provide answers to questions related to history only. An open domain QAS mostly works like search engines like Google and all where it provides explicit answers to question belonging to any domain. So in open domain QAS the scope for question is global.

Question in any question answering system can be of varying types. Question can be factoid question for which answers are simple fact about the entity in question. Some questions can be of descriptive type where one needs to full detail about a person, place or any event. There can be simple yes/no type of question which simple provides answers as yes or no. A question can also be an instruction-based question where answers are provided as an instruction to accomplish any task. Question in a QAS can be of many other forms which provide precise answer in the same format as that of question provided.

Example: Question Answering System for Hindi and Marathi language

The Question Answering System for Hindi and Marathi language is as shown in Figure 6.3. Here input to the system is natural language Hindi or Marathi Query. Input query is first tokenized to generate individual token and then these tokens undergo word grouping where two or three corresponding words are merged together if they are related with each other by using the available word grouped list. POS tagging is performed on word

grouped tokenized query text to extract relevant part of speech associated with the query text. POS tagged query text then passes through chunking process where noun and verb grouped present in the query text are extracted. Based on the extracted chunked groups initially query triples are extracted using Subject, Object and Verb (SOV). Then next process is to generate onto triples by fetching relevant onto words from ontology. Finally ontology is traversed to fetch relevant answer based on generate onto triples, if onto triple matches with any onto set in ontology then corresponding answer is fetched and passed to answer generation process to present the answer as natural way as possible mostly in the form of natural language text. The system for Hindi Question Answering System and Marathi Question Answering System both follow the same flow and have the same basic architecture with exception in Marathi that it requires separate case extraction phase, because case marker are mainly attached to noun words as suffixes, before extraction of SOV from query text as Marathi is much more inflected language compared to Hindi.

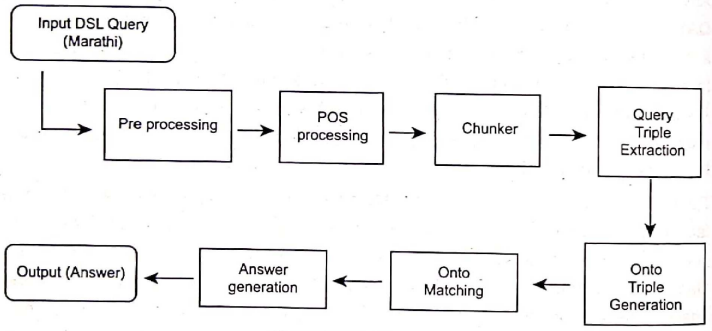


Figure 4: Basic QA system

The overall algorithm for proposed QA system is as follows:

Input: Natural language question in Hindi/Marathi language

Output: Answer in Hindi/Marathi language

Step 1: Tokenize the input question into word tokens.

Step 2: Grouped the correlated words into one merged word.

Step 3: Extract POS tag of each word in the tokenized list.

Step 4: Chunk the POS tagged words into noun and verb groups.

Step 5: Extract query triple from chunked grouped list.

Step 6: Generate onto triple.

Step 7: Traverse ontology to fetch answer.

Step 8: Formulate answer as natural language text.

Sample input and output for Hindi query:

Input Question:शिवाजी की माँ कौन थी?

Answer: शिवाजी की माँ जीजाबाई थी।

Sample input and output for Marathi query:

Input Question: शिवाजिची आई कोण होती?

Answer: शिवाजिची आई जीजाबाई होती.

The question answering system for Marathi natural language using concept of ontology as a formal representation of knowledge base for extracting answers. Ontology is used to express domain specific knowledge about semantic relations and restrictions in the given domains. The ontologies are developed with the help of domain experts and the query is analyzed both syntactically and semantically. The results obtained are accurate enough to satisfy the query raised by the user. The level of accuracy is enhanced since the query is analyzed semantically.

A question answering system is much more effective than traditional search engines as it provides accurate and precise answer to question rather than providing links to relevant documents or set of matching contents. Question answering system has become part of daily life of users, over a period of time many personal assistance software like Siri, Cortana, and Google Now are developed which provide precise and accurate answer to user question.

4. Sentiment Analysis

Sentiment Analysis (SA) is a natural language processing task that deals with finding orientation of opinion in a piece of text with respect to a topic. It deals with analyzing emotions, feelings, and the attitude of a speaker or a writer from a given piece of text. Sentiment Analysis involves capturing of user's behaviour, likes and dislikes of an individual from the text. The target of SA is to find opinions, identify the sentiments they express, and then classify their polarity.

The purpose of sentiment analysis is to determine the attitude or inclination of a communicator through the contextual polarity of their speaking or writing. Their attitude may be reflected in their own judgment, emotional state of the subject, or the state of any emotional communication they are using to affect a reader or listener. It is trying to determine a person's state of mind on the subject they are communicating about. This information can be mined from texts, tweets, blogs, social media, news articles, or comments.

There are different classification levels in SA: document-level, sentence-level and aspect-level. Document-level SA aims to classify an opinion of the whole document as expressing a positive or negative sentiment. Sentence-level SA aims to classify sentiment expressed in each sentence which involves identifying whether sentence is subjective or objective. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities which is done by identifying the entities and their aspects

Sentiment Analysis is a natural language processing task that deals with finding orientation of opinion in a piece of text with respect to a topic. It deals with analyzing emotions, feelings, and the attitude of a speaker or a writer from a given piece of text.

Sentiment Classification Techniques

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. 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, the 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 labelled as neutral. There are two main techniques for sentiment classification: machine learning based and lexicon based. Better performance can be obtained by combining these two methods.

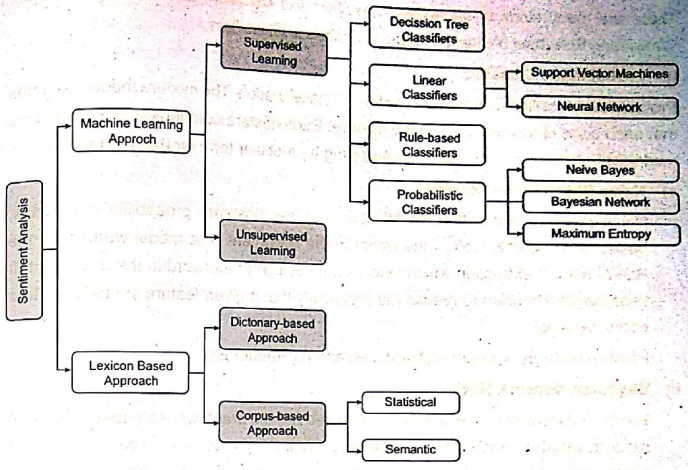


Figure 5: Sentiment Classification Techniques

Machine Learning Approach

The machine learning method uses several learning algorithms to determine the sentiment by training on a known dataset. The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification. In a machine learning based techniques, two sets of documents are needed: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to check how well the classifier performs.

A) Supervised Learning:

The supervised learning methods depend on the existence of labeled training documents. Supervised learning process: two Steps; Learning (training): Learn a model using the training data testing: Test the model using unseen test data to assess the model accuracy.

There are many kinds of supervised classifiers like Decision Tree Classifiers, Linear Classifiers, Rule-based Classifiers and Probabilistic Classifiers.

I) Probabilistic Classifier

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.

a) Naive Bayes Classifier (NB)

Naive 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.

P(label | features)=(P(label)*P(features|label))/(P(features))

b) Bayesian Network (BN)

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 Maxent Classifier also 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 {(feature set, label)} pair to a vector.

II) Linear classifiers

Score (or probability) of a particular classification is based on a linear combination of features and their weights. A linear classifier determines which class a object belongs by making a classification decision based on the value of a linear combination of the characteristics. An object's characteristics are also known as feature values and are

typically presented to the machine in a vector called a feature vector. There are many kinds of linear classifiers; among them is Support Vector Machines (SVM) which is a form of classifiers that attempt to determine good linear separators between different classes.

a) Support Vector Machines Classifiers (SVM)

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. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. It does not depend on the probabilities.

b) Neural Network (NN)

Neural Network consists of many neurons where the neuron is its basic unit. A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand. This is the learning phase. Multilayer neural networks are used for non-linear boundaries. These multiple

layers are used to induce multiple piecewise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The outputs of the neurons in the earlier layers feed into the neurons in the later layers. The training process is more complex because the errors need to be back-propagated over different layers.

III) Decision tree classifier

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.

IV) 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. There are numbers of criteria in order to generate rules, the training phase construct all the rules depending on these criteria. The most two common criteria are support and confidence. The support is the absolute number of instances in the training data set which are relevant to the rule. The Confidence refers to the conditional probability that the right-hand side of the rule is satisfied if the left-hand side is satisfied. Both decision trees and decision rules tend to encode rules on the feature space, but the decision tree tends to achieve this goal with a hierarchical approach.

B) Unsupervised learning:

Unsupervised learning is that of trying to find hidden structure in unlabelled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, a large number of labelled training documents are used for supervised learning, as illustrated before. In text classification, it is sometimes difficult to create these labelled training documents, but it is easy to collect the

unlabelled documents. The unsupervised learning methods overcome these difficulties.

K-nearest neighbors (KNN) is a simple unsupervised machine learning algorithm. In this algorithm, the objects are classified based on the majority of its neighbor. The class assigned to the object is most among its k nearest neighbors. The KNN classification algorithm classifies the instances of objects based on their similarities to instances in the training data. In KNN, selection is based on majority voting or distance weighted voting. Consider the vector A and set of M labelled instances {ai, bi}. The classifier predicts the class label of A on the predefined N classes. The KNN classification algorithm finds the k nearest neighbors of A and determines the class label of A using majority vote. KNN classifier applies Euclidean distance as the distance metric.

Lexicon-based approach

The lexicon-based approach involves calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review. The semantic orientation is a measure of subjectivity and opinion in text. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data.

There are three methods to construct a sentiment lexicon: manual construction, corpus-based methods and dictionary-based methods. The manual construction of sentiment lexicon is a difficult and time-consuming task. In dictionary based techniques the idea is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet dictionary for their synonyms and antonyms. The newly found words are added to the seed list. The next iteration starts. The iterative process stops when no more new words are found. The dictionary based approach have a limitation is that it can't find opinion words with domain specific orientations. Corpus based techniques rely on syntactic patterns in large corpora. Corpus-based methods can produce opinion words with relatively high accuracy. Most of these corpus based methods need very large labeled training data. This approach has a major advantage that the dictionary-based approach does not have. It can help find domain specific opinion words and their orientations

Basic sentiment analysis of text documents follows a straightforward process:

- Break each text document down into its component parts (sentences, phrases, tokens and parts of speech)
- Identify each sentiment-bearing phrase and component
- Assign a sentiment score to each phrase and component (-1 to +1)
- Optional: Combine scores for multi-layered sentiment analysis

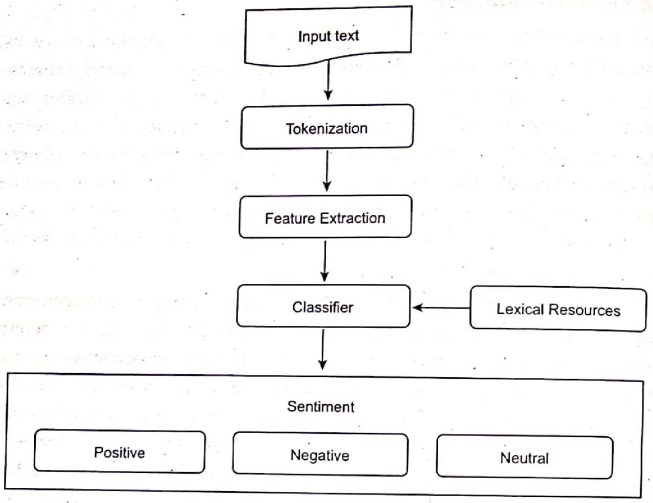


Figure 6: Basic Sentiment Analysis system

Example: Sentiment classification for Hindi language document using HindiSentiWoedNet

Problem Statement:

The proposed system extract sentiment associated with Hindi documents using Improved-HindiSentiWordNet (HSWN). The overall polarity of the document is calculated which can be positive, negative or neutral. The input to the system is a single text document in Hindi and output will be overall polarity associated with document.

The system consists of two stages:

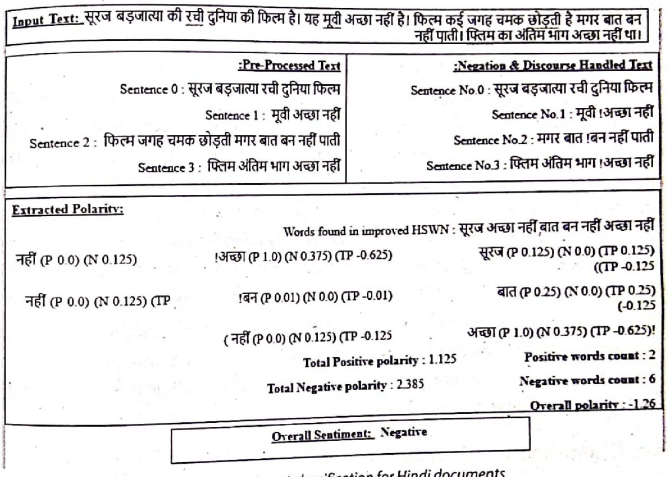
1. Improving HindiSentiWordNet (HSWN) .
2. Sentiment extraction

During the first stage , the existing HSWN is improved with the help of English SentiWordNet, where sentimental words which are not present in the HSWN are translated to English and then searched in English SentiWordNet to retrieve their polarity.

In the second stage, sentiment is extracted by finding the overall polarity of the document; which can be positive, negative or neutral. Here during preprocessing tokens are extracted from sentences and stop words are removed. Rules are devised for handling negation and discourse relation which highly influenced the sentiments expressed in the document. Finally, overall sentiment orientation of the document is determined by aggregating the polarity values of all the sentimental words in the document.

Input: Accepts a single document in Hindi as input.

Output: Overall sentiment associated with document.



Input Text: सूरज बड़जात्या की रची दुनिया की फिल्म है। यह मूवी अच्छा नहीं है। फिल्म कई जगह चमक छोड़ती है मगर बात बन नहीं पाती। फ्तिम का अंतिम भाग अच्छा नहीं था।

:Pre-Processed Text	:Negation & Discourse Handled Text
Sentence 0 : सूरज बड़जात्या रची दुनिया फिल्म	Sentence No.0 : सूरज बड़जात्या रची दुनिया फिल्म
Sentence 1 : मूवी अच्छा नहीं	Sentence No.1 : मूवी !अच्छा नहीं
Sentence 2 : फिल्म जगह चमक छोड़ती मगर बात बन नहीं पाती	Sentence No.2 : मगर बात !बन नहीं पाती
Sentence 3 : फ्तिम अंतिम भाग अच्छा नहीं	Sentence No.3 : फ्तिम अंतिम भाग !अच्छा नहीं

Extracted Polarity:

Words found in improved HSWN : सूरज अच्छा नहीं बात बन नहीं अच्छा नहीं

नहीं (P 0.0) (N 0.125) !अच्छा (P 1.0) (N 0.375) (TP -0.625) सूरज (P 0.125) (N 0.0) (TP 0.125) ((TP -0.125

नहीं (P 0.0) (N 0.125) (TP !बन (P 0.01) (N 0.0) (TP -0.01) बात (P 0.25) (N 0.0) (TP 0.25) (-0.125

(नहीं (P 0.0) (N 0.125) (TP -0.125 अच्छा (P 1.0) (N 0.375) (TP -0.625)!

Total Positive polarity : 1.125 **Positive words count** : 2

Total Negative polarity : 2.385 **Negative words count** : 6

Overall polarity : -1.26

Overall Sentiment: Negative

Figure 7: Sentiment classification for Hindi documents

Sentiment Analysis has been quite popular and has lead to building better products, understanding user's opinion, executing and managing of business decisions. People rely and make decisions based on the reviews and opinions.

5. Text Categorization or Classification

Text categorization is the process of assigning tags or categories to text according to its content. It's one of the fundamental tasks in Natural Language Processing (NLP) with broad applications such as sentiment analysis, topic labelling, spam detection, and intent detection.

Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes.

Text classification is a process of dividing a given set of documents into one or more predefined classes. This classification of text is done automatically. Usually machine learning techniques are used for automatic text classification. There are mainly two types of techniques namely supervised and unsupervised learning methods. Supervised learning methods assign predefined class label to the testing documents using classification algorithms whereas in unsupervised learning methods grouping of testing documents are done using techniques like clustering. There is also a semi-supervised learning method where, parts of the documents are labelled by the external mechanism.

Text classifiers can be used to organize, structure, and categorize pretty much anything. For example, new articles can be organized by topics, support tickets can be organized by urgency, chat conversations can be organized by language, brand mentions can be organized by sentiment, and so on.

Text Classification Techniques

A growing number of machine learning approaches or more specifically supervised learning methods have been applied to text classification which include Decision tree(C 4.5), K-Nearest Neighbor (K-NN), Bayesian approaches (Naïve and non-Naïve approaches), Neural Networks (NN), Regression based methods, Vector-based methods etc. Several clustering techniques are also available for text classification like K-means, Suffix Tree Clustering, Label Induction Grouping (LINGO) algorithm, Semantic Online Hierarchical Clustering (SHOC) etc.

A) Supervised Learning Methods

The supervised learning techniques are explained below:

a) **Decision Tree**

Decision tree methods reconstruct the manual categorization of the training documents in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. When the tree has created, a new document can simply be categorized by placing it in the root node of the tree and let it run through the query structure until it reaches a certain leaf.

b) **K-Nearest Neighbor (KNN)**

KNN is a statistical approach for text classification where objects are classified by voting several labeled training examples with their smallest distance from each object. The KNN classification method is outstanding with its simplicity and is widely used techniques for text categorization.

c) **Neural Network**

Neural network is also called artificial neural network is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. Different neural network approaches have been applied to document categorization problems. While some of them use the simplest form of neural networks, known as perceptions, which consist only of an input and an output layer, others build more sophisticated neural networks with a hidden layer between the two others.

d) **Naïve Bayes (NB)**

A naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. A naïve Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. Depending on the precise nature of the probabilistic model, Naïve Bayes classifiers can be trained very efficiently in a supervised learning setting.

e) Vector Based Methods

The two types of vector-based methods: The centroid algorithm and support vector machines. From these two algorithms centroid is simplest.

Centroid Algorithm: During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category. This algorithm is also appropriate if number of categories is very large. Centroid algorithm computes similarity of test document with each centroid using cosine similarity measure. It assigns a document, class with whose centroid a document has greatest similarity.

Support Vector Machine (SVM): The main idea of SVM is to find a hyper-plane that best separates the documents and the margin, distance separating the border of subset and the nearest vector document, is large as possible. The nearest samples of the hyper-plane named support vectors are selected. The calculated hyper-plane permits to separate the space in two areas. To classify the new documents, calculate the area of the space and assign them the corresponding category.

B) Clustering Techniques

Clustering of documents is mainly used to minimize the amount of text by categorizing or grouping similar data items. This grouping is a common way for human processing information, and one of the good techniques for clustering helps to build different varieties which provide automated tools.

The following is a brief introduction to some of the clustering techniques:

a) K-means Algorithm

It is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is a positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data.

b) Label Induction Grouping (LINGO) Algorithm

Lingo algorithm is based on vector space model. First it extracts the user readable and frequent words/phrases from the input documents. Further by performing the Reduction of Original Term Document Matrix with Singular Value Decomposition (SVD) method to reduce the term document matrix, and then it find the labels of clusters and then assigns documents to that cluster labels based on the similarity value.

C) Ontology Based Classification

Traditional classification methods ignore relationship between words, they consider each term independent of each result. But, there exist a semantic relation between terms such as synonym, hyponymy etc. Therefore, for better classification results, there is need to understand the context of the text document. The ontology has different meaning for different users, in this classification task, Ontology stores words that are related to particular domain. Therefore, with the use of domain specific ontology, it becomes easy to classify the document even if the document does not contain the class name in it.

Example : Automatic Text Categorization of Marathi Language Documents

The designed system takes input as a set of Marathi Language text documents. These documents undergo preprocessing steps which include input validation, tokenization, stop word removal, stemming and morphological analysis. Then the features are extracted from preprocessed tokens. Then at last supervised learning methods and ontology based classification are applied to get output as classified Marathi documents as per class labels. The supervised learning methods considered here are Naïve Bayes (NB), Modified K-Nearest Neighbor (MKNN) and Support Vector Machine (SVM),

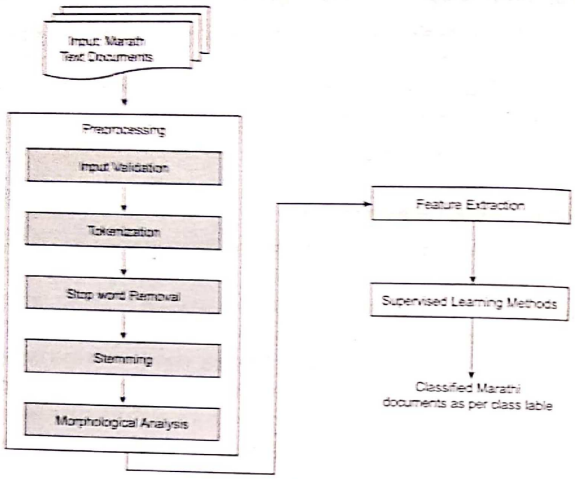


Figure 8: Text Categorization system for Marathi Documents

The input to the system would be a set of Marathi documents. The documents would be of different domains like Politics, Entertainment, Literature, Sports, History and Technology.

Sample Document 1:

आधीच ठरलेल्या इशाऱ्याप्रमाणे भेटीच्या वेळी तीन तोफांचे बार प्रतापगडावरून काढण्यात आले, आणि खानाच्या छावणीच्या जवळपासच्या झाडाझुडुपांमध्ये दडून बसलेल्या मावळ्यांनी हल्ला करून खानाच्या सैन्याची दाणादाण उडविली. खानाचा मुलगा फाजलखान आणि इतर काही सरदार लपूनछपून वाईच्या मुख्य छावणीपर्यंत आले. इथे खानाचा जनाना होता. ते पाठलागावर असलेल्या नेताजीच्या सैन्यापासून वाचण्यासाठी खजिना, हत्ती व इतर जड सामान टाकून विजापूरला जनान्यासकट पळाले.

शिवाजीराजांना जनतेत मिळालेला आदर आणि प्रेम अनेक शतकांनंतरही टिकून आहे त्यामागचे त्यांची सहिष्णू वृत्ती हे फार महत्त्वाचे कारण आहे. अफझलखानाच्या मृत्यूनंतर त्यांनी त्याच्या शवाचे अंत्यसंस्कार इस्लामी पद्धतीने करून त्याची एक कबर प्रतापगडावर बांधली आणि त्या कबरीच्या कायम देखभालीची व्यवस्था केली.

अफझलखानच्या मृत्यूनंतर शिवाजीराजांनी दोरोजी नावाच्या सरदाराला कोकणपट्ट्यातील आणखी किल्ले आणि प्रदेश जिंकण्यास पाठवले. स्वतः राजे सातारा प्रांतात घुसून कोल्हापुरापर्यंत गेले व त्यांनी पन्हाळा जिंकून घेतला.

Sample Document 2:

तेंडुलकर हा नियमितपणे गोलंदाजी करत नसला तरी त्याने १३२ कसोटी सामन्यांमध्ये ३७ बळी आणि ३६५ एकदिवसीय सामन्यांमध्ये १४२ बळींची कामगिरी केली आहे. ज्यावेळेस महत्त्वाचे गोलंदाज अपयशी ठरत असतात त्यावेळेस सचिनला गोलंदाजी देण्यात येते. आणि बऱ्याच वेळेस तो बळी मिळविण्यात यशस्वी ठरतो. जरी त्याची गोलंदाजीची सरासरी ५० च्या वर असली, त्याला 'जम बसलेली फलंदाजांची जोडी फोडण्याचा हातगुण असणारा गोलंदाज' समजण्यात येते.(End of text)

Sample Document 3:

लता मंगेशकर (जन्म: सप्टेंबर २८, इ.स. १९२९) भारताच्या महान गायिका आहेत. त्या भारताच्या हिंदी चित्रपटसृष्टीच्या ख्यातनाम गायक-गायिकांपैकी एक आहेत. हिंदी संगीतविश्वात त्यांना 'लता दीदी' म्हणून ओळखले जाते. लता मंगेशकरांच्या कारकिर्दीची सुरूवात इ.स. १९४२मध्ये झाली आणि ती कारकीर्द सहा दशकांपेक्षा अधिक काळ टिकून आहे. त्यांनी ९८० पेक्षा अधिक हिंदी चित्रपटांची गाणी गायली असून, विसाहून अधिक प्रादेशिक भारतीय भाषांमध्ये (प्रामुख्याने मराठी) गायन केले आहे. लता मंगेशकरांचे कुटुंब संगीतासाठी प्रसिद्ध असून, सुप्रसिद्ध गायिका आशा भोसले, उषा मंगेशकर, मीना मंगेशकर आणि ख्यातनाम संगीतकार-गायक हृदयनाथ मंगेशकर ही त्यांची सख्खी भावंडे आहेत. लता मंगेशकरांचे वडील मास्टर दीनानाथ मंगेशकर हे मराठी नाट्य-संगीताचे प्रसिद्ध गायक होते.

भारताचा सर्वोच्च पुरस्कार 'भारतरत्न' प्राप्त होणाऱ्या गायक-गायिकांमध्ये लता दीदी या दुसऱ्या गायिका आहे.

The Output of text categorization system is as follows :

Categories in which the documents are classified:

Sample Document 1 : शिवाजीराजा

Sample Document 2 : गोलंदाजी

Sample Document 3 : गायिका

The tremendous development in the Information Technology has led to the availability of large number of documents on the Internet. Classifying such documents according to class wise or topic wise manually is time consuming and complex task. Also, managing and retrieval of such documents is a difficult task. To overcome these difficulties, automatic text classification systems can be used.

6. Text Summarization

Summarization means to reduce the size of the document without changing its meaning. It is one of the most researched areas among the Natural Language Processing (NLP) community. A good summary should cover the most vital information of the original document or a cluster of documents, while being coherent, non-redundant and grammatically readable. Automatic text summarization is the data science problem of creating a short, accurate, and fluent summary from a longer document. Summarization methods are greatly needed to consume the ever-growing amount of text data available online. In essence, summarization is meant to help us consume relevant information faster. Furthermore, applying text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area.

Summarization techniques are categorized into extractive and abstractive techniques on the basis of whether the exact sentences are considered as they appear in the original text

Types of Text Summarization:

A. 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.

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

B. Abstraction-based summarization :

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method.

The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text — just like humans do.

Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that's why the use of extraction is still popular.

Here is an example:

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

Text summarization algorithm :

Usually, text summarization in NLP is treated as a supervised machine learning problem (where future outcomes are predicted based on provided data).

Typically, here is how using the extraction-based approach to summarize texts can work:

1. Introduce a method to extract the merited keyphrases from the source document. For example, you can use part-of-speech tagging, word sequences, or other linguistic patterns to identify the keyphrases.
2. Gather text documents with positively-labeled keyphrases. The keyphrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled keyphrases.
3. Train a binary machine learning classifier to make the text summarization. Some of the features you can use include:
   - Length of the keyphrase
   - Frequency of the keyphrase
   - The most recurring word in the keyphrase
   - Number of characters in the keyphrase
4. Finally, in the test phrase, create all the keyphrase words and sentences and carry out classification for them.

Example : Abstractive text summarisation for MARATHI documents

The idea is to summarize an input Marathi document by creating semantic graph called rich semantic graph(RSG) for the original document, reducing the generated semantic graph, and further generating the final abstract summary.

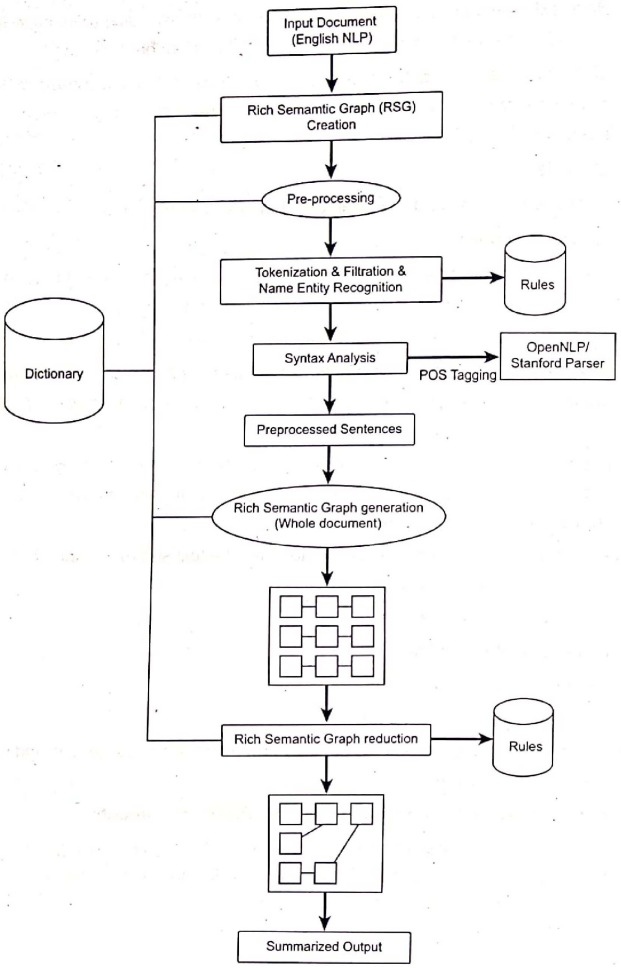


Figure 9: Text Summarization system for Marathi Documents

This approach consists of the following phases:

a. **Marathi text document as input.**

b. **Rich Semantic Graph (RSG) creation pha:**

In Rich semantic graph creation step, analysis of the input text document is done, finds the sentence and produces tokens for the complete document. For every word it creates POS tags and detects the words into predefined categories such as person's name, location and organization. After this it generates the graph for every sentence and then it concatenates rich semantic sub-graphs. At last the sub-graphs are mixed together to show the complete document correctly.

c. **Rich Semantic Graph (RSG) reduction phase:**

Rich semantic graph reduction phase targets to reduce the obtained rich semantic graph of the source document to more reduced graph. Here a set of rules are applied on the obtained rich semantic graph to reduce it by merging, consolidating or deleting the graph nodes.

d. **Summary generation from reduced RSG:**

The summarized text generation phase targets to obtain the abstractive text summar from the reduced Rich Semantic Graph (RSG). To reach the target, this phase accesses the domain ontology; it has the data required in the same domain of RSG to obtain the final output. In addition, the Word Net ontology is used to obtain multiple texts according to the word synonyms. The obtained multiple texts are accessed and ranked, the most ranked text is considered.

Input : Single text document in Marathi Language.

मुरलीधर देविदास आमटे उर्फ बाबा आमटे हे एक थोर मराठी समाजसेवक होते.बाबा आमटेंच्या जन्म डिसेंबर २३ १९१४ रोजी महाराष्ट्रातील वर्धा जिल्ह्यात झाला.बाबा आमटेंना भारत सरकार कडून १९७१ मध्ये पद्मश्री पुरस्कार प्राप्त झाला आहे. तसेच बाबा आमटेंना भारत सरकार कडून १९८६ मध्ये पद्मविभूषण पुरस्कार प्राप्त झाला आहे. Baba Amte was a very good human being.

Output : Reduced Maeningful summary.

मुरलीधर देविदास आमटे उर्फ बाबा आमटे हे एक थोर मराठी समाजसेवक होते. त्यांचा जन्म डिसेंबर २३ १९१४ रोजी महाराष्ट्रातील वर्धा जिल्हात झाला. त्यांना भारत सरकार कडून १९७१ मध्ये पद्मश्री पुरस्कार आणि १९८६ मध्ये पद्मविभूषण पुरस्कार प्राप्त झाला आहे

7. Named Entity Recognition

The term "named entity", now widely used in Natural Language Processing, was first introduced for Sixth Message Understanding Conference (MUC-6) in 1995. During that time the MUC was focused on the information extraction were structured information of company and defence activities has been extracted from the unstructured text, from sources such as newspaper articles. During this task people felt a need for a system that can identify the names that includes names of persons, organisations, locations, and many different entities also numeric expressions which includes date, time, currency, numbers, percentages etc. For named entity all phrases in the text were supposed to be marked as person, location, organization, time or quantity. So identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)". Natural languages are the languages which have naturally

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text. NER involves identification of proper names in texts, and classification into a set of predefined categories of interest.

For example: Mahatma Gandhi is very famous in India as "Bapu" or The full name of him is Mohandas Karamchand Gandhi. He was a great freedom fighter who led India as a leader of the nationalism against British rule. He was born on the 2nd of October in 1869 in Porbandar, Gujarat. He was a leader of Congress, as it was the only national party in India.

Tag Names Tagged Entities

Person	*Mahatma Gandhi, Bapu, Mohandas Karamchand Gandhi*
Location	*India, Porbandar, Gujarat*
Organisation	*Congress*
Date	*2nd, October, 1869*

There are many NER applications which include Information Extraction, Question Answering, Information Retrieval, Automatic Summarization, Machine Translation etc. The Named Entities can be known to us, if we perform computations on the natural language. The task of extracting and retrieving important information and details become easier and faster only if the Nes are already known. In defining IE tasks, people noticed

that it is essential to recognize information units such as names including person, organization, and location names, and numeric expressions including time, date, money, and percentages. Identifying references to these entities in text was acknowledged as one of IE's important sub-tasks and was called **"Named Entity Recognition (NER).**

NER Entity Classes (Tagset)

Named Entities (NEs) are noun phrases in natural language text. Natural language text is a sequence of sentences, where sentence in turn is a sequence of words and punctuation combined to add semantic to the text. Further, a word is a character sequence. In NER the aim is to distinguish between character sequence that represent noun phrases and character sequence that represents normal text. A proper noun in the text that is used to refer to a person, company, location etc. can be tagged as Named Entity (NE).

The main types of NEs are shown in fig below.

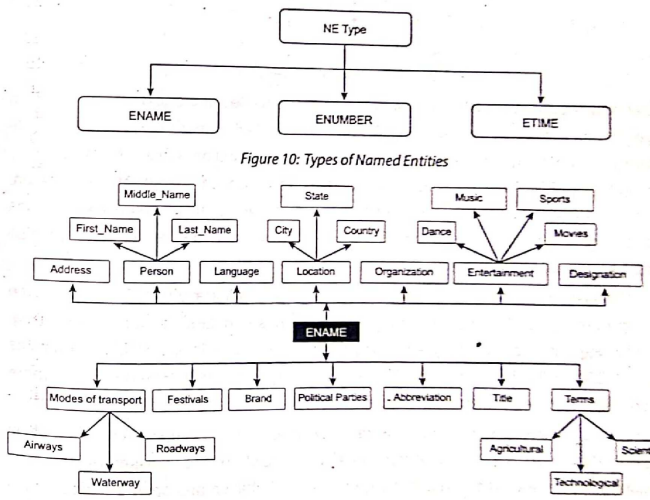


Figure 10: Types of Named Entities

Figure 11: Expansion of NAME Tagset

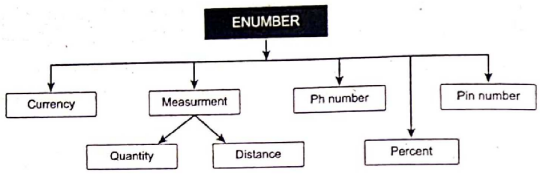


Figure 12: Expansion of NUMBER Tagset

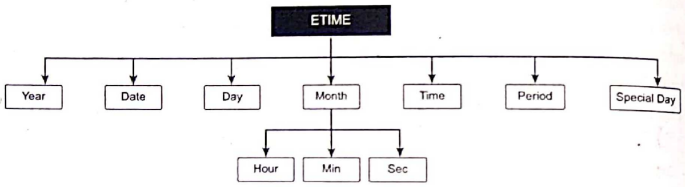


Figure 13: Expansion of TIME Tag set

Proper nouns play vital role in discovery of semantics hidden in the text. Identification of proper nouns in the raw text and classification of identified proper nouns in appropriate category is an important sub task of information extraction called Named Entity Recognition (NER). Any entity such as person, organization, geographical location, government agency, event that has a name is treated as a named entity. Expressions such as money, percentage, phone numbers, date, time, URLs, addresses are also treated as named entities even if they are not exactly like traditional proper nouns.

The NER system extracts name entity present in the given Marathi text. To extract name entity from the given text we are using two approaches in natural language processing. One of the approaches is machine learning where we are going to use HMM to predict the correct NER tag for the given word. The second approach is Rule Based system where handcrafted rules are written which helps is identifying the correct NER tag of the word.

The input to the system consists of Marathi text related to news domain which consist of text related to politics, sports, entertainment etc. Input text is processed to create an annotated NER data set by linguistic experts. In HMM based approach the input set is divided into two parts: training and testing.

In the training phase of HMM the annotated text is tokenized and the parameters for Viterbi algorithm like transition, emission etc are computed. In the testing phase correct

NER tag is predicted using Viterbi algorithm. In the Rule Based approach first the input text is tokenized and POS tagged. By using the NER tag set, correct NER tags are assigned to words in the sentence. The words which are left to be tagged are then tagged by using handcrafted rules. Finally the output of the system is NER tag text.

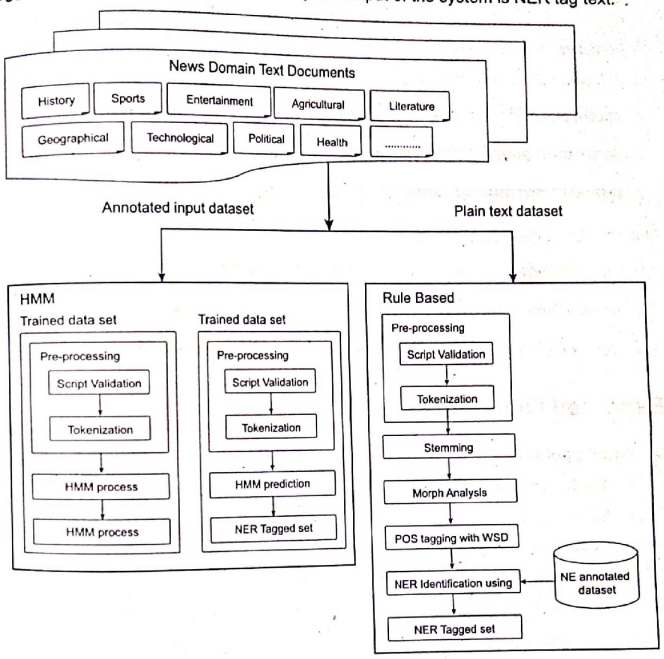


Figure 14: Named Entity Recognition Framework for Marathi Language

Supervised machine learning algorithm consists of train and test phase. In training phase prediction model is generated on the basis of various input parameters and state. In testing phase prediction model is used to predict the labels for the input. HMM is used as supervised machine learning algorithm to predict the NER tagged for Marathi text.

The input to the HMM module consists of annotated dataset which is further divided into test and train dataset. The training phase of HMM module gets the input as annotated dataset which is tokenized and then HMM algorithm is applied which results in creation of HMM model. In the testing phase the input can be the test dataset sentences or random user input which are tokenized and provided to HMM model which with the help of viterbi algorithm is used to find the correct tag for the user input.

Input: Marathi Plain Text (test data set).

1. सुबश्रीने आज ब्लँकेट विणण्याचे उपक्रम चालू केले.
2. ऑगस्ट पासून जानेवारी पर्यंत हे चालू होता.
3. सुबश्रीने जानेवारी पासून सुरू असलेला रेकॉर्ड आज पूर्ण केला.

Output: NER tagged Data.

1. <NER सुबश्रीने/Name> आज ब्लँकेट विणण्याचे उपक्रम चालू केले.
2. <NER ऑगस्ट/Month> पासून <NER जानेवारी/Month> पर्यंत हे चालू होता.
3. <NER सुबश्रीने/Name> <NER जानेवारी/Month> पासून सुरू असलेला रेकॉर्ड आज पूर्ण केला.

Expected Questions

Q. Write a note on :

1. Machine translation.
2. Information retrieval vs Information Extraction.
3. Question answering system.
4. Text Categorization.
5. Text Summarization.
6. sentiment analysis.
7. Named Entity Recognition.

Q. List various applications of NLP and discuss any 2 applications in detail.