Question Answering System

- Question answering system helps user to find the precise answer to the question articulated in natural language.
- Q-A 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.
- Q-A System can be classified into two main types:
- **1. Closed Domain QAS:**
- In closed domain QAS scope of user question is limited to a particular domain like medicine, movies, history and others.
- If QAS is created for history domain it will provide answers to questions related to history only.

- **2. Open Domain QAS:**
- An open domain QAS mostly works like search engines like Google and all where it provide explicit answers to question belonging to any domain, So in open domain QAS the scope for question is global.

- Questions in any QAS can be varying types:
- Questions 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 details about a person, place or any event.
- There can be simple yes/no type of question which simple provides answers as yes or no.

- Example: QAS for Hindi and Marathi Language:
- Input: Natural Language question in Hindi/Marathi
- Output: Answer in Hindi/Marathi language
- Step 1: Tokenize the input into word tokens.
- Step 2: Grouped the correlated words into 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 out for Marathi Query:
- Input Question: शिवाजीची आई कोण होती?
- Answer: शिवाजीची आई जिजाबाई होती.

Basic QA System

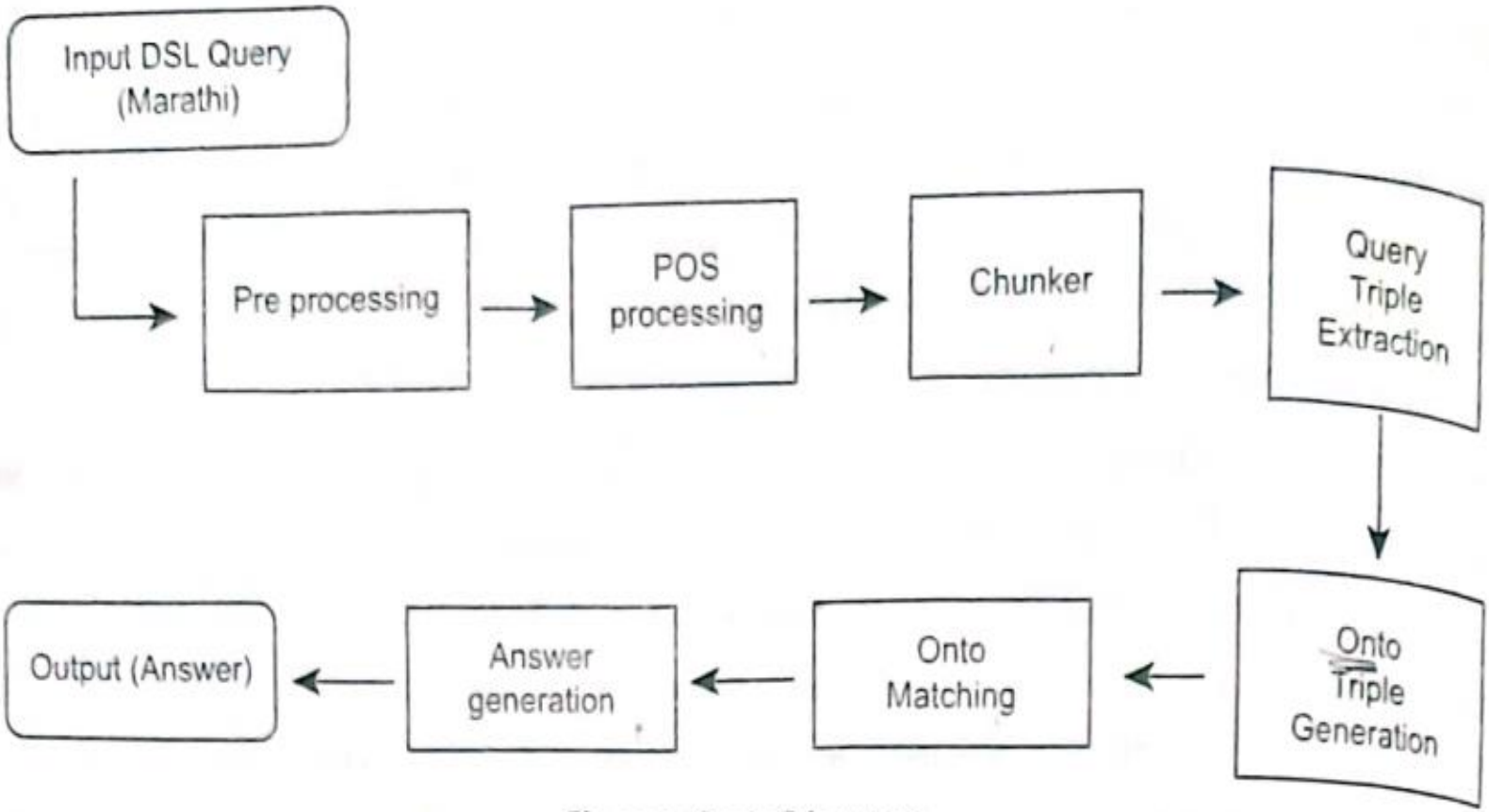


Figure 4: Basic QA system

- 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 passed through chunking process where noun and verb grouped present in the query text 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 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 text.

- Ontology:
- It is formal representation of knowledge base for extracting answers.
- It is used to express domain specific knowledge about semantic relations and restrictions in the given domain.
- The ontologies are developed with the help of domain experts and the query is analyzed both syntactically and semantically.
- The results obtained are accurate since the query is analyzed semantically

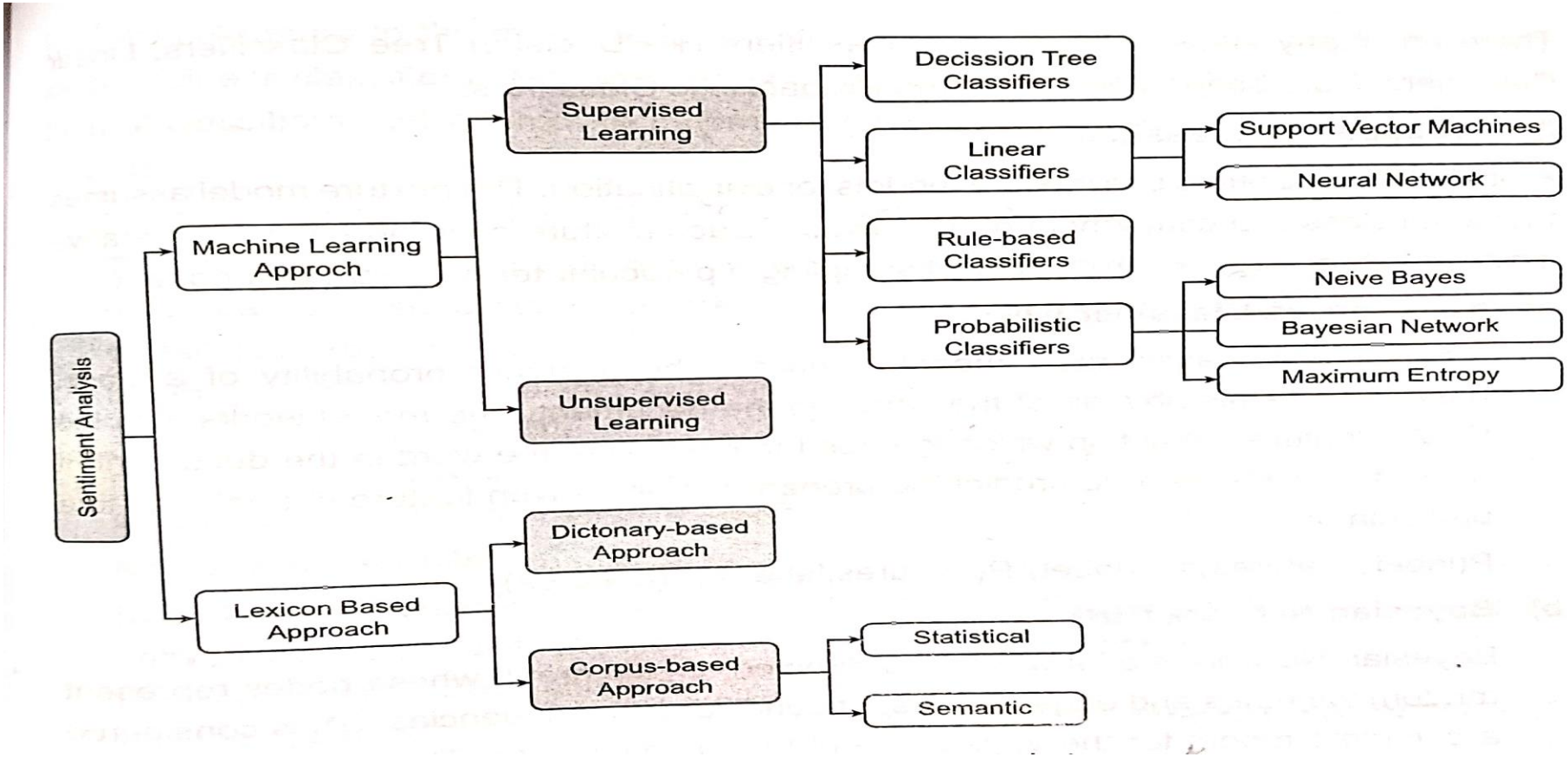
Sentiment Analysis

- Sentiment Analysis (SA) is 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, feeling, and the attitude of speaker or a writer from a given piece of text.
- SA involves capturing of user behavior, likes and dislikes of an individual text.
- Example:
- Suppose, there is a fast-food chain company and they sell a variety of different food items like burgers, pizza, sandwiches, milkshakes, etc. They have created a website to sell their food and now the customers can order any food item from their website and they can provide reviews as well, like whether they liked the food or hated it.

- User Review 1: I love this cheese sandwich, it’s so delicious.
- User Review 2: This chicken burger has a very bad taste.
- User Review 3: I ordered this pizza today.
- So, as we can see that out of these above 3 reviews,
- The first review is definitely a **positive** one and it signifies that the customer was really happy with the sandwich.
- The second review is **negative**, and hence the company needs to look into their burger department.
- And, the third one doesn’t signify whether that customer is happy or not, and hence we can consider this as a **neutral** statement.
- By looking at the above reviews, the company can now conclude, that it needs to focus more on the production and promotion of their sandwiches as well as improve the quality of their burgers if they want to increase their overall sales.

- There are different classification levels in SA:
- 1. Document level:
- Document level aims to classify an opinion of the whole document as expressing a positive ,negative or neutral sentiment.
- 2. Sentence Level SA:
- Sentence level SA aims to classify sentiment expressed in each sentence which involves identifying whether sentence is subjective or objective.
- 3. Aspect Level SA:
- 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 Classification is a task under Sentiment Analysis that deals with automatically tagging text as positive, negative and neutral from the perspective of the speaker/writer with respect to topic.



Machine Learning Approach

- The machine learning method uses several learning algorithms to determine the sentiment by training on a known dataset.
- In a machine learning based techniques two sets of documents are needed: training and test set.
- A training set is used by an classifier to learn the different characteristics of documents, and 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:
- 1. Learning(training): Learn a model using training data.
- 2. Testing: Test a model using unseen test data to assess the modal accuracy.

1. Decision Tree Classifier

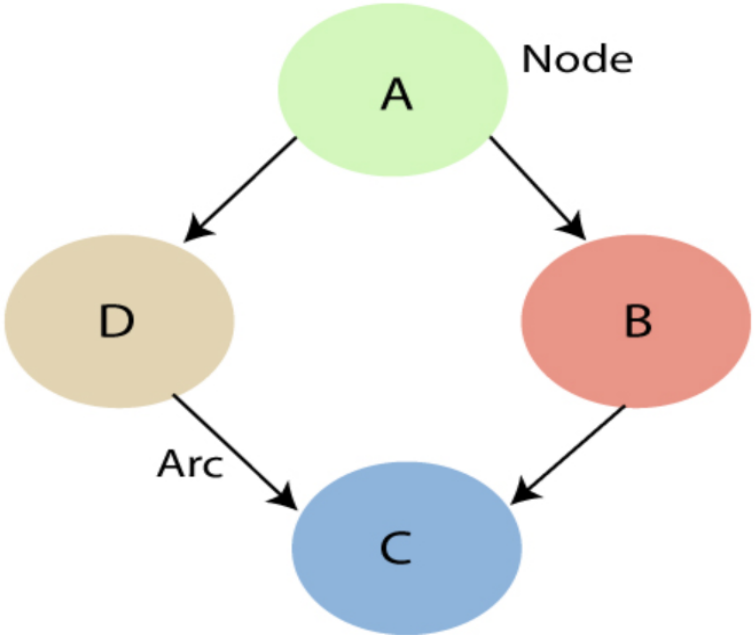
- Decision Tree classifier provides a hierarchical decomposing of the training data space in which a condition on the attribute value is used to divide the data.
- The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification.
- 2. Linear Classifiers:
- A linear classifier determines which class a object belongs by making a classification decision based on the value of linear combination of the characteristics.
- An objects characteristics are known as feature values and are presented to the machine in a vector called a feature vector.

- a) Support Vector Machine:
- The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes.
- b) Neural Network:
- 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 input takes an input, applies a 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.

- 3. 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 while the right- hand side is the class label.
- Training phase construct all the rules.
- Ex: Decision Tree

- 4.Probabilistic Classifiers:
- It uses probabilities model for classification.
- a) Naïve bayes Classifiers:
- This model computes conditional probability of a class, based on the distribution of the words in the document.
- This model works with BOWs feature extraction which ignores the position of the word in the document.
- It uses bayes Theorem to predict the probability.

- b) Bayesian Network (BN):
- A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph.
- Bayesian networks are probabilistic, because these networks are built from a **probability distribution**.



- c) Maximum Entropy Classifier:
- Maxent Classifier also known as a conditional exponential classifiers converts labeled feature sets to vectors using encoding.
- This encoded vector is then used to calculate weights for each feature that can be then combines to determine the most likely label for feature set.

- B) Unsupervised Learning:
- The main purpose of text classification is to classify documents into a certain number of predefined categories.
- By using unsupervised learning it is difficult to labeled documents.

- K-nearest Neighbors (KKN) algorithm:
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- KNN applies Euclidean distance as the distance metric

Lexicon Based Approach

- Sentiment lexicon contains list of words and expressions used to express peoples subjective feelings and opinions.
- For example; start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if document has more positive word lexicons, it is positive, otherwise it is negative.
- There are two methods to construct a sentiment lexicon;
- 1. Dictionary Based :
- In dictionary based techniques the idea is to collect a small set of opinion words manually, 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 limitation is that it can not find opinion words with domain specific orientations.

- 2. Corpus Based:
- Corpus based techniques rely on syntactic patterns in large corpus.
- Corpus based methods can produce opinion words with relatively high accuracy.
- Most of these Corpus based methods need very large labeled training data. So it can help to find domain specific opinion words and their orientations.

Basic Sentiment Analysis System

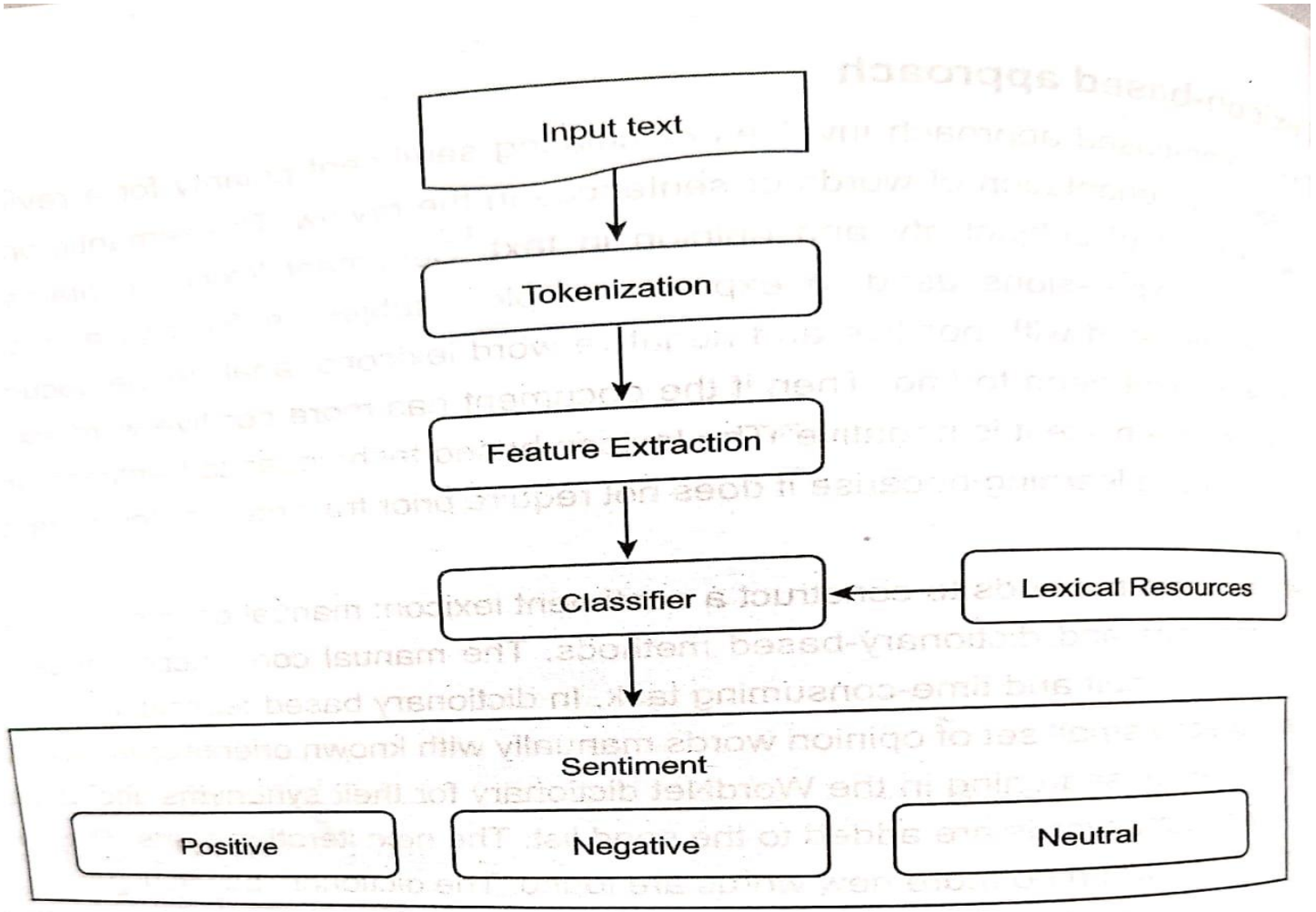


Figure 6: Basic Sentiment Analysis system