COMPUTER ENGINEERING DEPARTMENT

ASSIGNMENT NO. 2

Subject: Natural Language Processing

COURSE: B.E Year: 2021-2022 Semester: VIII
DEPT: Computer Engineering
SUBJECT CODE: DLO8012 SUBMISSION DATE: 10/04/2022

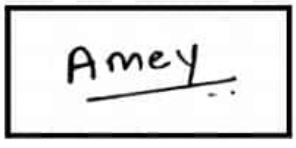
Roll No.: 50 Name: Amey Thakur

Class: BE-Comps B Date of Submission: 10/04/2022

NLP Assignment - 2

Sr. No.	Questions
1	What is CFG? How is it useful in NLP?
2	Explain the differences between HMM, Maximum entropy and CRF with an example?
3	How are generative and discriminative models used for sequence labelling in NLP?
4	Explain Homonymy, Polysemy, Synonymy, and Hyponymy with an example.
5	Write a short note on WorldNet and its relevance in NLP.
6	What is WSD? How is Word Sense Disambiguation (WSD) achieved through the dictionary-based approach?
7	With respect to Pragmatics, define the following terms (a) Discourse; (b) reference resolution; (c) reference phenomenon.
8	Write a short note on Syntactic and semantic constraints on co-reference.
9	How are the Machine translation and Question Answer (QA) systems performed in NLP? Describe in detail.
10	How are Text categorization and summarization performed in NLP? Explain with an example.
11	Describe different types of sentiment analysis performed in NLP.
12	What is Name Entity Recognition (NER)? Explain with an example.

Student Signature:



AMEY THAKUR B-50 Amey

Q1. What is CFG? How is it useful in NLP?

Ans:

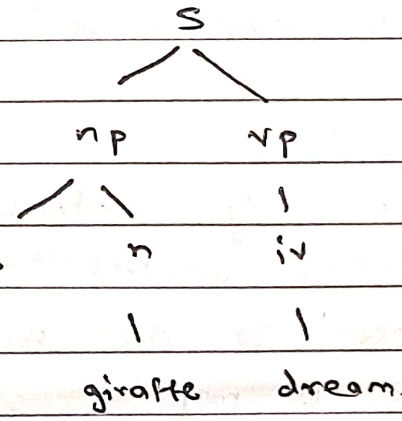
- A Context Free Grammar (CFG) is a list of rules that define the set of all well form sentences in a language. Each rule has a left hand side which identifies a syntactic category and a right hand side which defines its alternative component parts reading from left to right.
- Eg. the rule s → np vp means that "a sentence is defined as a noun phrase followed by a verb phrase".
- Following figure shows a simple CFG that describes the sentences from a small subset of English.
- A grammar and a parse tree for "the giraffe dreams".

s → np vp
np → det n
vp → tv np
→ iv

det → the
→ a
→ an

n → giraffe
→ apple

iv → dreams
tv → eats
→ dreams



AMEY THAKUR B-50 Amey

- A sentence in the language defined by a CFG is a series of words that can be derived by systematically applying the rules, beginning with a rule that has S on its left hand side. A parse of the sentence is a series of rule applications in which a syntactic category is replaced by the right hand side of a rule that has category on its left hand side and the final rule application yields the sentence itself.
- E.g. a parse of the sentence "the giraffe dream" is :

S ⇒ np vp ⇒ det n vp → the n vp
⇒ the giraffe vp ⇒ the giraffe iv
⇒ the giraffe dreams

A convenient way to describe a parse is to show its parse tree, which is simply graphical display of the parse.

- A CFG only defines a language. It does not say how to determine whether a given string belongs to the language it defines. To do this, a parser can be used whose task is to map a string of words to its parse tree. The parse tree of course remains the same.

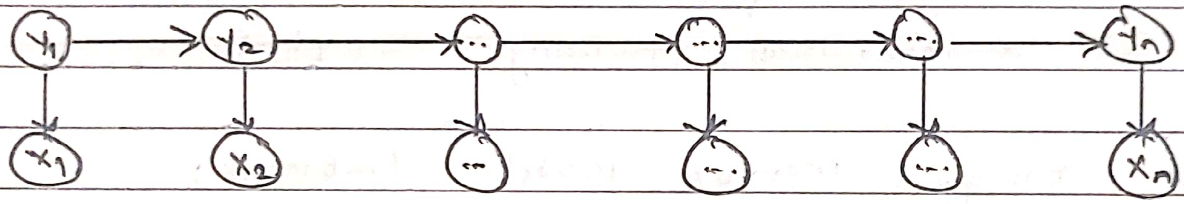
AMEY THAKUR B-50 Amey

Q2. Explain the differences between HMM, maximum entropy and CRF with an example.

Ans:

Hidden Markov Model (HMM)

- The word "Hidden" symbolizes the fact that only the symbol released by the system are observable, while the user cannot view the underlying random walk between states. Many in this field recognize HMM as a finite state machine.



Hidden Markov Model (HMM)

Advantages of HMM:

- HMM has a strong statistical foundation with efficient learning algorithms where learning can take place directly from raw sequence data. It allows consistent treatment of insertion and deletion penalties in the form of locally learnable method and can handle inputs of variable length. They are most flexible generalization of sequence profiles. It can also perform a wide variety of operations including multiple alignment, data mining and classification, structural analysis and pattern discovery. It is also easy to combine into libraries.

AMEY THAKUR B-50 Amey

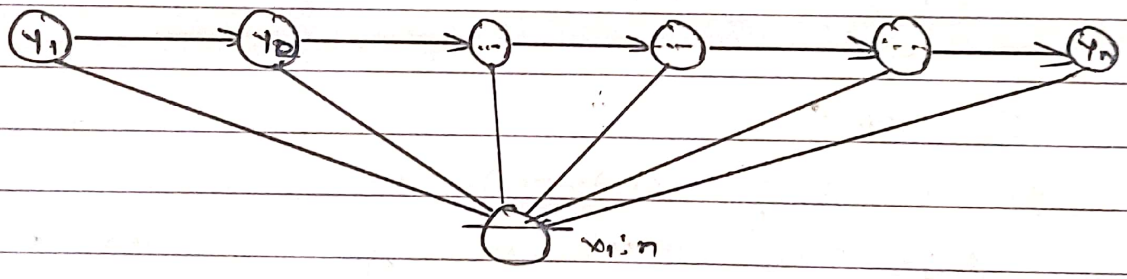
Disadvantages of HMM

- HMM is only dependent on every state and its corresponding observed object.

The sequence labelling in addition to having a relationship with individual words, also relates to such aspects as the observed sequence length, word context and others.

HMM acquires the joint distribution $P(Y, X)$ of the state and the observed sequence, while in the estimation issue we need a conditional probability $P(Y|X)$

Maximum Entropy Markov Model (MEMM):



Maximum Entropy

- The form of a CRF is heavily motivated by the principle of maximum entropy framework for estimating probability distributions from a set of training data.

- Entropy of a probability distribution is a measure of uncertainty and is maximized when the distribution in question is as uniform as possible.

AMEY THAKUR B-50

Amey

- The principle of maximum entropy asserts that the only probability distribution that can justifiable be constructed from incomplete information, such as finite training data is that which has maximum entropy subject to a set of constraints representing the information available.

CRF (Conditional Random Fields)

- The CRF is a conditional probabilistic model for sequence labelling just as structured perception is built on the perception classifier, conditional random fields are built on the logistic regression classifier.

- CRFs are a probabilistic framework for labelling and segmenting sequential data based on the conditional approach described in hidden markov model.

- A CRF is a form of undirected graphical model that defines a single log linear distribution over label sequences given a particular observation sequence.

- The primary advantage of CRF over HMM is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference.

AMEY THAKUR B-50 Amey

Q3. How are generative and discriminative models used for sequence labelling in NLP?

Ans:

(1) Many NLP tasks are sequence labelling tasks.
Input a sequence of tokens / w
output a sequence of corresponding labels

Eg. POS Tags, BIO encoding for NER solution.
Finding the most probable labels sequence for the given word sequence.

$t^* = \text{argmax}_t P(t|w)$
t is a vector matrix.

(2) Generative models
model the joint probability of labels and words

$t^* = \text{argmax}_t P(t/w) = \text{argmax}_t P(w/t) P(t)$

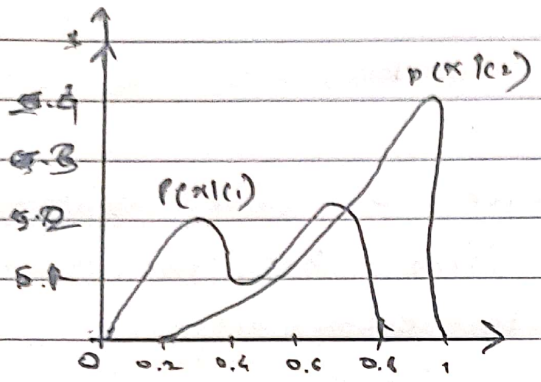
(3) Discriminative models
Directly model the conditional probability of labels given the words

$t^* = \text{argmax}_t p(t|w) = \text{argmax}_t f(t, w)$

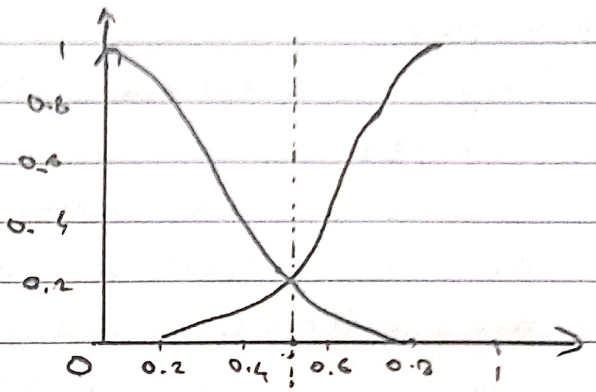
AMEY THAKUR B-50 Amey

Binary classification as example:

Generative models view



Discriminative models view



④ Generative models

- Specifying joint distribution
  Full probabilistic specification for all the random variables
- Dependence assumption has to be specified for P(w/t) and P(t).
- Flexible, can be used in unsupervised learning

⑤ Discriminative models

- Specifying conditional distribution only explain the target variable
- Arbitrary features can be incorporated for modelling P(t/w)
- Need labelled data, only suitable for (semi-supervised learning).

AMEY THAKUR B-50 Amey

Q4. Explain Homonymy, Polysemy, Synonymy and Hyponymy with an example.

Ans:

Homonymy

- They are lexemes that share a form but unrelated meanings
- Homonyms are words that have a same syntax or same spelling or same form but their meanings are different and unrelated to each other.
- Two or more words become homonyms if they either sound the same (homophones), have the same spelling (homographs) or if they both homophones and homographs but do not have related meaning
- Examples:
  1. bat (wooden stick thing) vs bat (flying mamal)
  2. bank (Financial institution) vs bank (riverside)

Polysemy

- Polysemy refers to words or phrases with different but related meanings.
- A word becomes polysemous if it can be used to express different meanings
- The difference between these meanings can be obvious or subtle
- It is sometimes difficult to determine whether a word is polysemous or not because the relations between words can be vague and unclear.
- But, examining the origins of the words can help to decide whether a word is polysemic or homonymous.

AMEY THAKUR B-50 Amey

- The following sentences contain some examples of polysemy.
  1. He drank a glass of milk.
  2. He forgot to milk the cow.
  3. He read the newspaper.

Synonymy

- A synonymy is a word or phrase that means exactly the same as another word or phrase.
- In other words, synonyms are words with similar meanings.
- For instance, words like delicious, yummy, succulent are synonymous of the objective tasty.
- Similarly, verbs like commence, initiate and begin are synonymous of the verb start.
- However, some synonyms do not have exactly the same meaning. There may be minute differences.
- Sometimes a word can be synonymous with another in one context or usage but not in another.
- Examples:
  1. Beautiful - Gorgeous.
  2. Purchase - Buy
  3. Use - Employ
  4. Rich - Wealthy
  5. Mistake - Error.

AMEY THAKUR B-50 Amey.

Hyponymy.

- In linguistics & lexicography, hyponym is a term used to designate a particular member of a broader class. for instance, daisy & rose are hyponyms are flowers. Also called a subtype or subordinate term. The adjective is hyponymic.
- Eg. - Pigeon, crow, eagle and seagull are hyponyms of birds. their hyponym itself is a hyponym of animal.

AMEY THAKUR B-50 Amey

Q5. Write a short note on WordNet & its relevance in NLP.

Ans:

WordNet

- It is a large lexical database of English nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets).
- WordNet is a big collection of words from the english language that are related to each other and are grouped in some way.
- It is also called as a lexical database.
- WordNet is a database of english words that are connected together by their semantic relationship
- It is like a superset dictionary with a graph structure.
- WordNet groups nouns, verbs, adjectives, etc. which are similar and groups are called synsets.
- In a WordNet a group of synset may belong to some other synset.
- For example, the synsets "stones" and "cement" belong to the synset "Building Materials" or the synset "stones"
- Every member of synset denotes the same concept but not all synset members are interchangeable in context.
- The membership of words in multiple synsets or concepts mirrors polysemy.

AMEY THAKUR B-50 Amey

Q6. What is WSD? How is WSD achieved through the dictionary based approach?

Ans:

Word Sense Disambiguation (WSD)

- WSD is a well known problem in NLP.
- WSD is used in identifying what the sense of word means in a sentence when the word has multiple meanings.
- When a single word has multiple meaning then for machine it is difficult to identify the correct meaning and to solve this challenging issue we can use the rule-based system or machine learning techniques.
- WSD is a natural classification problem. Given a word and its possible senses as defined by a dictionary, classify an occurrence of the word in the context into one or more of its sense classes.
- The feature of the context such as neighbouring words provide the evidence for classification.
- A famous example is to determine the sense of pen in the following phrase.

" Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy."

WordNet lists 5 senses for the word pen

- Pen - a writing implement with a point from which ink flows
- Pen - an enclosure for confining livestock
- Playpen, Pen - a portable enclosure in which babies may be left to play
- Pen - female swan.
- Penitentiary Pen - A correctional institution for those convicted of major crats.

AMEY THAKUR B-50 Amey.

There are four conventional approaches to WSD.

① Dictionary → knowledge based methods
- These primarily rely on dictionaries, and lexical knowledge bases without using any corpus evidence.

② Supervised Methods:
- These make use of sense annotated corpora to train from.

③ Semi-supervised / minimally supervised methods
- These make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process or a word aligned bilingual corpus.

④ Unsupervised Methods
- These eschew (almost) completely external information and work directly from raw unannotated corpora.

Dictionary and knowledge based methods.
- The lesk method is dictionary based methods. It is based on the hypothesis that words used together in text are associated with one another which the relation are often observed within the definitions of the words and their senses.
- Two or more words are disambiguated by finding the pair of dictionary senses with the best word overlap in their dictionary definitions
- For example, when disambiguating the words in pine cone the definition of the acceptable senses both include the words evergreen and tree a minimum of in one dictionary

AMEY THAKUR B-50

Amey

- An alternative to the utilization of the definition is to think about general word sense relatedness and to compute the semantic similarity of every pair of word senses supported a given lexical knowledgebase like wordNet.
- Graph based methods like spreading activation research of the first days of AI research are applied with some success.
- The use of selection preferences or selection restrictions also are useful.
- For example, knowing that one typically cooks food, one can disambiguate the word bass in I am cooking bass i.e. It's not a musical instrument.

AMEY THAKUR B-50 Amey.

Q7 With respect to pragmatics define following terms.
① Discourse, ② Reference solution, ③ Reference Phenomenon

Ans:

① Discourse

- Language does not normally consists of isolated, unrelated sentences but instead of collocated, related groups of sentences, such a group of sentences is known as a discourse.
- The flow of communication is only in one direction in a monologue, i.e from speaker to the hearer.
- A conversation with a friend about it, which would consist of a much freer interchange is called a dialogue.
- In this case, each participant periodically takes turns being a speaker and hearer.
- The computer system allow for HCI that has properties which distinguish it from normal human - human dialogue, in part due to the present day limitations on the ability of computer systems to participate in free, unconstrained conversation.

② Reference Solution

- It defined as the task of determining what entities are referred to by which linguistic expression.
- Terminologies
  a) Reference expression
     - Natural language expression that is used to perform reference.
  b) Referent
     - Entity that is referred

AMEY THAKUR B-50 Amey

c) Coreferr
- When two expressions are used to refer to same entity.

d) Antecedent
- Term has the license to use another term.

e) Anaphora & Anaphoric
- It may be defined as the reference to an entity that has been previously introduced into the sentence and the referring expression is called anaphoric.

f) Discourse Model
- The model that contains the representations of the entity that has been referred to in the discourse and the relationship they are engaged in.

(3) Reference Phenomenon
- The set of referential phenomena that natural languages provide is quite rich indeed.
- Indefinite noun phrases introduces entities that are new to the hearer into the discourse content.
- Definite noun phrases is used to refer to an entity that is identifiable to the hearer either because it has already been mentioned in the discourse content, it is contained in the hearer's set of beliefs about the world or the uniqueness of the object is implied by the description itself.
- Pronouns usually refers to entities that were introduced no further than one or two sentence back in the ongoing discourse.
- Demonstrative pronouns like this & that behave somewhat differently that simple definite pronouns like it.

AMEY THAKUR B-50 Amey.

- One anaphora blends the properties of definite & indefinite references.
- Inferrable are the referents where a regular expression does not refer to an entity that has been explicitly evoked in the text but instead one that is inferentially related to an evoked entity.
- Discontinuous sets: In some cases, references using plural referring expressions like they & them refer to set of entities that are evoked together.
- Making the reference problem even more complicated is the existence of generic reference.

AMEY THAKUR B-50 Amey

Q8.: Write a short note on: Syntactic & Semantic constraints on coreference

Ans:

- We need to find a way of filtering the set of possible referents using hard & fast constraints
- Such constraints include:
  a) Number agreement
  b) Person and case agreement
  c) Gender agreement
  d) Syntactic Constraints
  e) selectional restrictions

a) Number agreement

- Referring expressions & their referents must agree in number.

→ John has a new car. It is red.
John has three new cars. They are red.

→ John has a new car. They are red
John has three new cars. It is red.

b) Person & case agreement

- English distinguishes ~~strong~~ three forms of person First, second & third.

→ You & I have cars. We love them
John & Mary have cars. They love them.

→ John & Mary have cars. We love them
You & I have cars. They love them.

AMEY THAKUR B-50 Amey

c) Gender agreement
- Referents also must agree with the gender specified by the referring expression.
- English third person pronouns distinguish between male, female & non-personal

→ John has a car.
He is attractive (he = John, not the car)
It is attractive (it = car, not John).

d) Syntactic constraints.
- Reference relations may also be constrained by the syntactic relationship.
Reflexive pronoun co-refers with the subject of the most immediate clause that contains it, whereas a non-reflexive cannot co-refer with this subject.

→ John bought himself a new car [himself = John]
John bough him a new car [him ≠ John]

- The rule about reflexive pronouns applies only for the subjects of the most immediate clause.

→ John said that Bill bought him a new car
[him ≠ Bill]
John said that Bill bought himself a new car.
[himself = Bill]
He said that he bough John a new car.
[He ≠ John]

AMEY THAKUR B-50 Amey

e) Selectional Restriction:

- It is a restrictions that a verb imposes on its arguments, may be used for eliminating referents.

→ John parked his car in the garage
He had driven it for hours

- Selectional restrictions can be avoided in the case of metaphor.

→ John bought a new car. It drinks gasoline like you would not believe.

AMEY THAKUR B-50 Amey

Q9. How are the machine translation and Question Answer (QA) systems performed in NLP? Describe in detail.

Ans:

Machine Translation

- It is the process of converting the text in a source language to a required target language.
- The various types of machine translations are:
  1. Statistical Machine Translation.
  2. Rule-based Machine Translation
  3. Hybrid Machine Translation
  4. Neural Machine Translation

1. Statistical Machine Translation (SMT)
   - It is a machine translation paradigm where translations are made on the basis of statistical models, the parameters of which are derived on the basis of bilingual text corpus.
   - SMT is based on information theory which studies the quantification storage and communication of information

2. Rule-based Machine Translation (RBMT)
   - It relies on innumerous built in linguistic rules & millions of bilingual dictionaries for each language pair.
   - In RBMT, translations are built on various sophisticated rules, but it does provide the users with the freedom to make use of their own terminology by adding them to the translation process.

AMEY THAKUR B-50 Amey

③ Hybrid Machine Translation (HMT)

- It is a method of machine translation that incorporates the use of multiple different machine translation approaches within a single machine translation.
- The underlying motivation behind the use of HMT is the fact that failure of a single machine translation technique should not stop the system from achieving the required level of accuracy.

④ Neural Machine Translation (NMT)

- In NMT, we make use of a neural network model to learn a statistical model for machine translation.
- One of the key advantage of NMT over SMT is that in NMT, a single system can be trained directly on source as well as target text thereby removing the dependency on the pipeline of a specialized system as that is SMT.

Question - Answer (QA) system.

- QA is a computer science discipline within the fields of information retrieval & natural language processing which is concerned with building system that automatically answer questions posed by humans in a natural language.
- Effective answers mean answers relevant to the questions posed by the user. As sentences must effectively map semantics of the statement.

AMEY THAKUR B-50 Amey

- The process of the system is as follows
  1. Query Processing
  2. Document Retrieval
  3. Passage Retrival
  4. Answer Extraction.

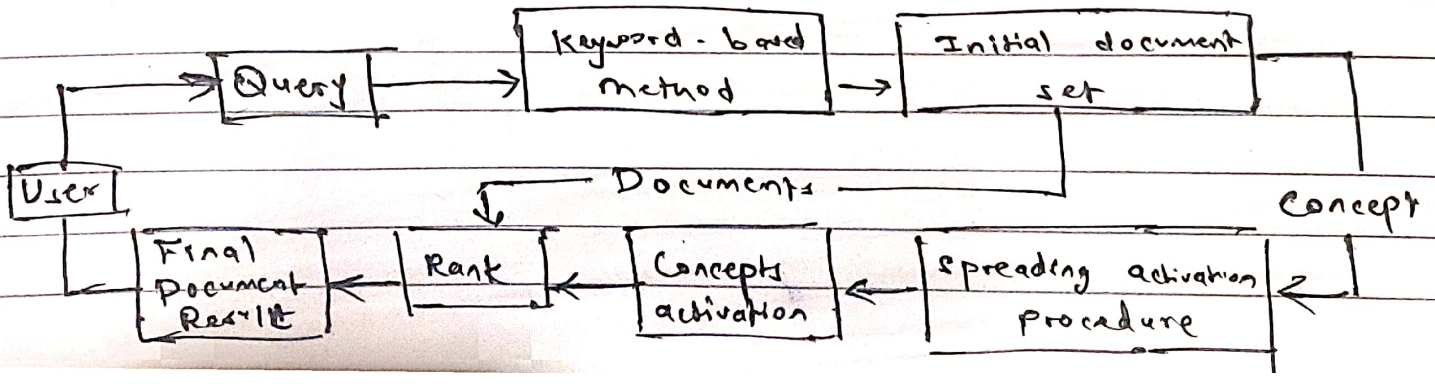
1. Query Processing
   - Classify questions into seven categories.
     - Who is/was/were/.... }
     - When is/did/are/.... }
     - Where is/are/were/.... }
   - a. Category application rules
   - b. Expected answer "Datatype".
     eg: Date, Person, Location, ...
   - c. eg: Where is the museum located }

Weight 1:
lots of non-answer could come back too

Weight 4:
If we get a match its probably right.

+museum + located + "the museum is located"

2. Document Retrieval
   - Here, we will retriere relevant document by using the generated query
   - Users submit queries corresponding to their information used.



AMEY THAKUR B-50 Amey

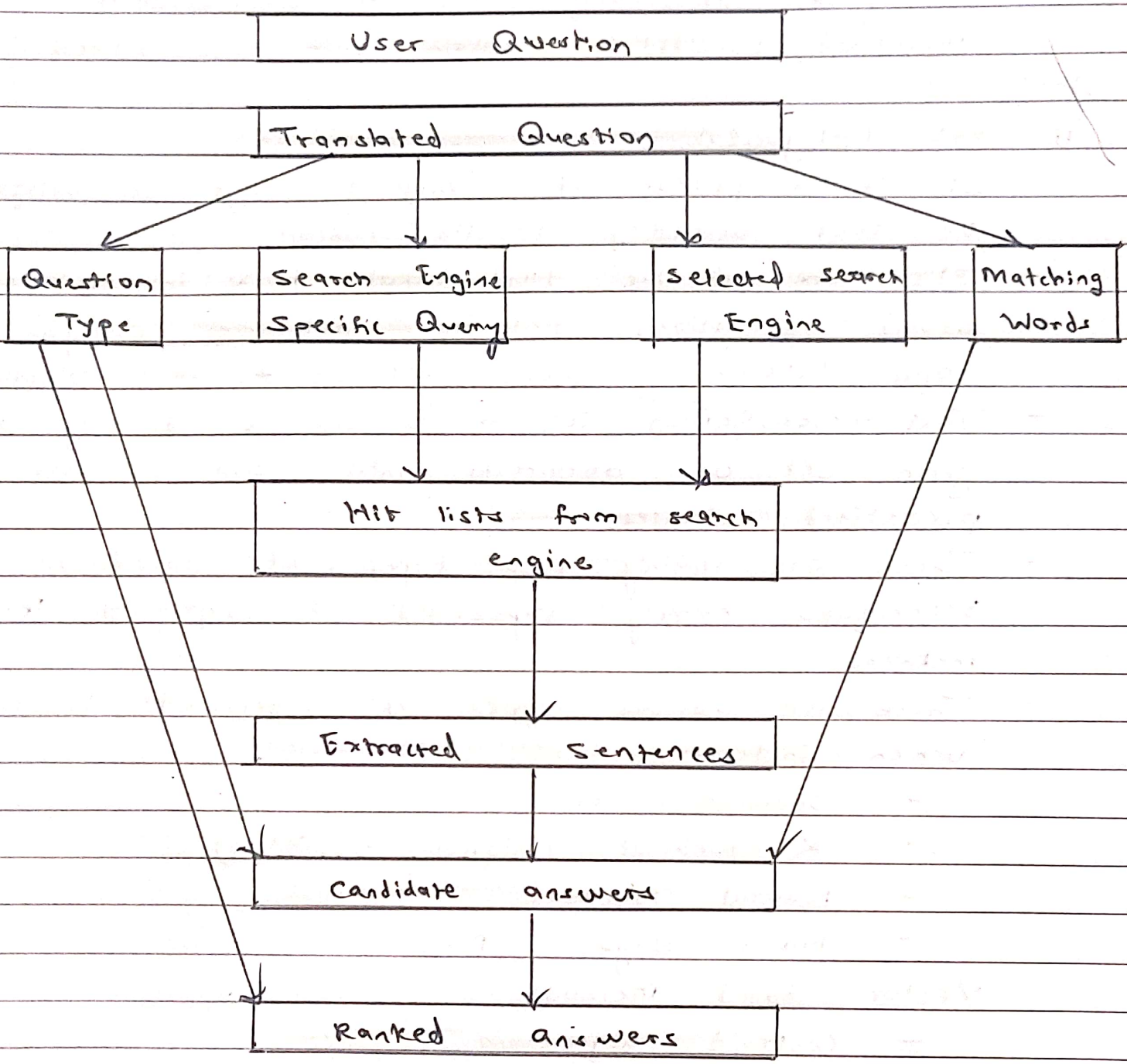
③ Passage Retrieval

- PR systems have been proposed for English language PR systems work on parts of texts so that can limit the relevance of a document to a query, besides detecting document portions that are likely to contain the required answers.
- The documents are separated into smaller units such as sentences & paragraphs that are likely to contain an answer are nominated.
- If the document is short, it is not needed but if it is long, passage selection is effective because long & uninformed which part of document the query matched.

④ Answer Extraction

- It processes the top ranked passages for extracting the final answer to the user's question.
- Typically, named entity recognition technique is used to find the candidate answers that match the questions named entity.
- Answer extraction extracts answer from passage. Here, the question & passage are input to the answer extraction model & the model outputs the answer offset with the score.

AMEY THAKUR B-50 Amey



Answer Extraction

AMEY THAKUR B-50 Amey.

Q10. How is text categorization & summarization performed in NLP? Explain with an example.

Ans:

① Text Categorization

- It is a process of assigning tags or categories to text according to its content.
- It's one of the fundamental tasks in NLP with broad applications such as sentimental analysis, topic labelling, spam detection & intent detection.
- Text classification is the process of dividing a given set of documents into one or more predefined classes.
- There are mainly two types of classification techniques namely supervised & unsupervised learning methods.
- There are various types of supervised learning which include:
  - Decision Tree
  - K Nearest Neighbour (KNN)
  - Neural Network
  - Naive Bayes (NB)

Vector Based Methods
  - Centroid algorithm
  - Support Vector Machine (SVM)

Clustering algorithm
  - K-means algorithm
  - Label Induction Grouping (LINGO) algorithm.

AMEY THAKOR B-50 Amey

Example: Fraud & Online Abuse Detection
- Filtering fraud & abuse is also possible thanks to text classification.
- These classifiers are used to detect bullying & trolling on social networks & unwanted content on the internet.
- Using a topic classifier, for example, you can track specific topics that are mentioned on the internet or train it to recognize specific language & categorize it as abusive.

(2) Text Summarization.
- It is the process of generating short, fluent and accurate summary of a respectively longer text document.
- The main idea behind automatic text summarization is to be able to find a short subset of the most essential information from the entire set & present it in a human readable format.
- Types of summarization
  (1) Extractive Summarization
  (2) Abstractive Summarization

AMEY THAKUR B-50 Amey

① Extractive summarization

- These methods rely on extracting several parts such as phrases & sentences from a piece of text & stack them together to get a summary
- Eg: Before summarization
  - John & Joseph took a time to attend the night party in the city while in the party, John collapsed & was ~~re~~ rushed into the hospital.

After summarization

- John & Joseph ~~attended~~ attended party John was ~~rushed~~. rushed.

② Attractive summarization

- Here, summary of the texts can be different from original text which is contrast to the extraction based summarization where only existing sentences which are present are used.
- Eg: Before summarization
  - John & Joseph took a time to attend the night party in the city while in the party John collapsed & was rushed into the hospital

After summarization

- John was hospitalized after attending the party.

AMEY THAKUR B-50 Amey

Q11. Describe the different types of sentiment analysis performed in NLP.

Ans:

- Sentiment analysis is the process of detecting positive or negative sentiment in text. Since customers express their thoughts & feeling more openly than ever, sentiment analysis is becoming an essential tool to monitor & understand the sentiment.
- Automatically analyzing customer feedback such as opinions in survey, response & social media conversation allows brands to learn what makes customers happy or frustrated so that they can tailor products & services to meet the customer needs.
- There are different algorithms depending on how much data needs to be analyzed & how accurate the model needs to be.

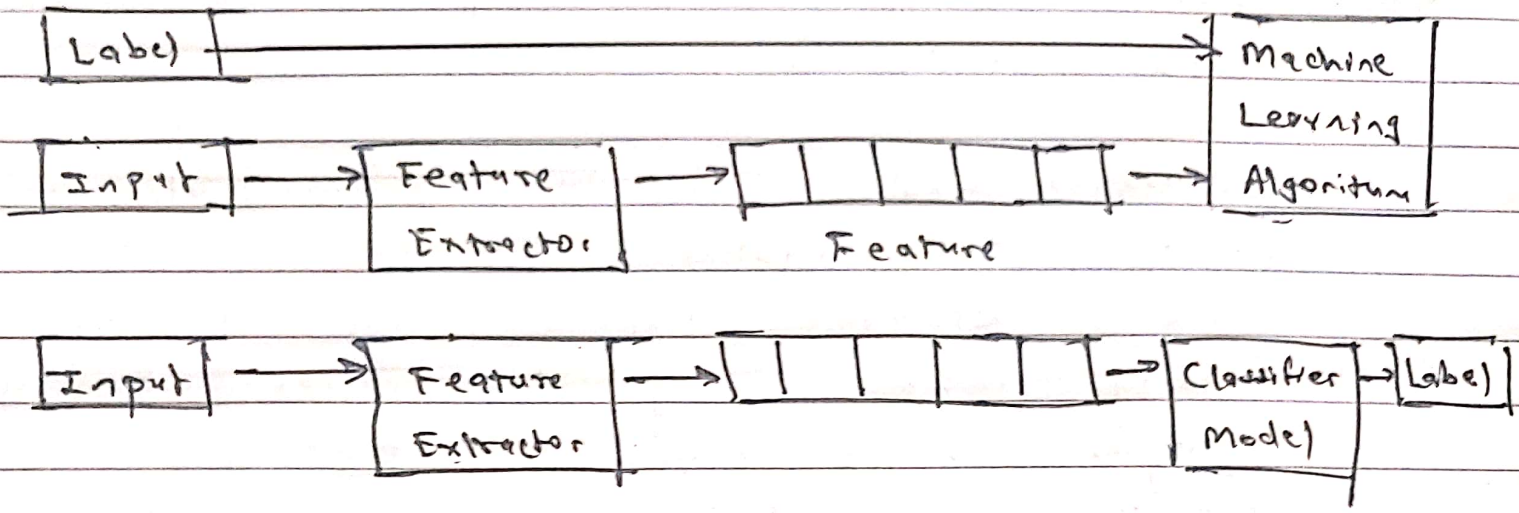
Types:

① Rule - based
- These systems automatically perform sentiment analysis based on a set of manually crafted rules.

② Automatic
- Systems only rely on machine learning technique to use learn from data.

③ Hybrid.
- Systems combine both rule-based & automatic approaches.

AMEY THAKUR B-50 Amey.



Training & classification sentiment analysis

AMEY THAKUR B-50 Amey

Q 12 What is NER? Explain with an example.

Ans:

- Named Entity Recognition (NER) also called Entity identification is a NLP technique that automatically identifies the named into predefined categories.

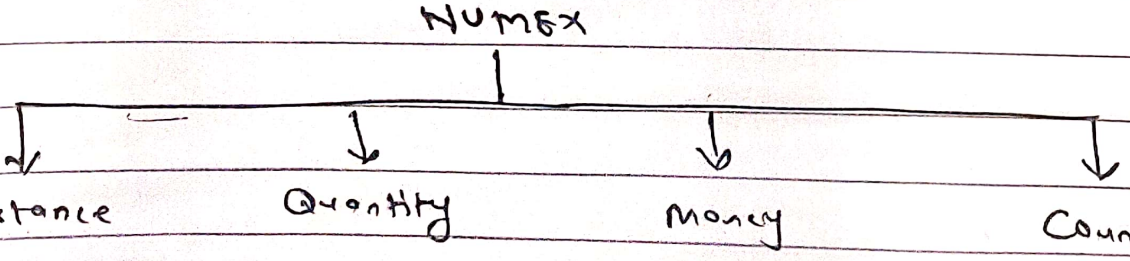
- Entities can be names of people, organization, location, times, quantities, values, percentage and more

Eg: We work founder Adam lives his New York
[organization] [Person] [location]

- The named entity hierarchy is divided into three major classes.: Entity, None, Time, and Numerical expressions

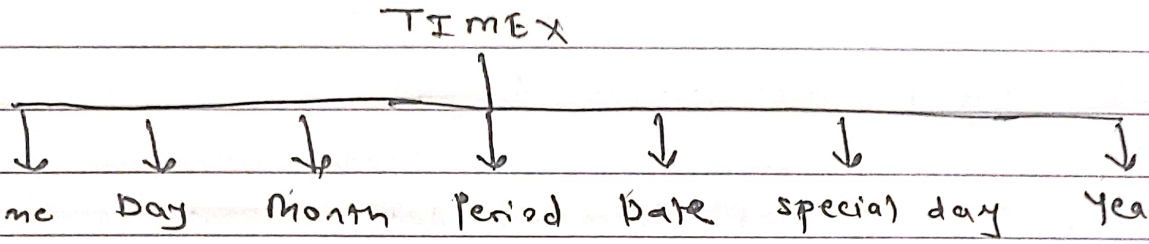
① Entity Name Types
- Person are entities limited to humans individual refers to names of each individual person.
- Location entities are limited to geographical entities such as oceans, ~~countries~~ countries, etc.
- Organization entities are limited to corporations agencies, other groups of people defined by an established organizational structure.

② Numerical Expression



AMEY THAKUR B-50 Amey

③ Time Expressions



Eg. "Mark Zukerberg is one of the founders of Facebook, a company from United States".

Person: Mark Zukerberg
Organization: Facebook (company)
Location: United States