NLP UNIT-V

1. Discourse cohension:- Refers to the various linguistic devices that contributes to the smooth and logical flow of a text, ensuring that ideas are connected and understood by the reader.

Reference:- This involves using pronouns & other words to refer back to something mentioned earlier.

Ex. John lost his wallet. He reported it to the police. The officer promised to investigate.

conjunction:- words like and, but, however etc. link ideas and show the relationship between them.

Ex. Although it was raining, they decided to go for a hike. However, they took umbrellas, just in case.

Parallelism:- Using similar grammatical structures for elements in a sentence or series of sentences.

Ex. She enjoys reading, swimming, and hiking in her free time.

Repetition: Repeating words & phrases for emphasis & to reinforce a point.

Ex. The new policy aims to reduce waste. By reducing waste, we contribute to a cleaner environment.

→ Discourse structure

Discourse structure involves organizing ideas in a coherent manner, ensuring a logical progression from one point to the next.

Logical order :- presenting ideas in a sequence that makes sense.

Ex. First, gather the ingredients, Next, mix them together. Finally bake the mixture.

Comparison :- highlighting/ highlighting similarities & diff. between ideas.

Ex. Similarly, like her sister, Mary enjoys painting.

Summarizing :- providing a brief overview of the main points

Eg. In conclusion, to sum up, overall, in summary.

cause and effect:- Explaining the relationship b/w actions and their outcomes.

Ex. Due to heavy rain, the event was cancelled. As a result, participants were disappointed.

chronological order:- presenting information in the order in which events occured

Ex. In the morning, I walk,
In the afternoon, "
In the evening "

→ N gram models

N gram models are a type of probabilistic language model widely used in NLP. The "N" in N gram represents the no. of consecutive items (words or tokens) considered as a unit.

1. Unigram (1-gram)
consider each word a token in isolation, treating them as independent entities.

Ex. I love programming.

I

love

programming.

2. Bigram (2-gram)

consider pair of consecutive words of tokens

Ex. I love programming.

3. Trigram (3-gram).

consider triplets of consecutive words of tokens.

Ex: I love programming.

4. N-gram in General.

An N-gram model considers seq. of N consecutive words & tokens.

Ex. N=4

How N-gram models work

1. Probability Estimation.

'N-gram models estimate the probability of the next word in a seq. given the preceding N-1 words.

Eg. P (Programming)

Represents the probability of the word programming.

2. chain rule of probability.

The probability of a seq. of words can be decomposed using the chain rule of probability.

3. language modeling!-

N-gram models are often used for language modeling, helping predict likely word sequences in a given content

Application

Speech recognition,
machine translation
text generation.

Evaluating language models

1. Perplexity :- It measures how well a model can predict the next word in a sequence of words.
2. Human evaluation :- Human evaluation involves having humans assess the quality of generated text or the performance of a language mode on a specific task.
3. Task - Specific evaluation :- involves evaluating a language model on a specific task, such as machine translation, text summarization & sentiment analysis.
4. Diversity and novelty evaluation :- involves evaluating the diversity and novelty of the text generated by a language model. is used to evaluate the creativity and originality of generated text.

→ Parameter estimation

Parameter estimation is the process of estimating the values of the parameters of a statistical model from the data.

1) Maximum likelihood estimation :- which involves finding the set of parameters that maximizes the probability of the observed data

(*) Bayesian estimation :- which involves finding the posterior distribution of the parameters given the data.

Bayes, uses a hierarchical model to estimate the parameters.

Steps 1 :-

1. Preprocessing the data
2. Select a model architecture
3. define objective function
4. Select optimization algorithm.

The optimization algorithm is used to find the values of the parameters that minimize the objective function.

1. Maximum-Likelihood estimation and smoothing

It is commonly used method for estimating the parameters of a statistical model based on observed data.

Smoothing method is laplace smoothing also known as add-one smoothing. This method involves adding a small constant value to the count of each event, which ensures that the probability estimate is never zero.

Kneser-Ney smothing which estimates the probability of a word based on its frequency in the training corpus and the no. of unique contents in which it appears.

(2) Bayesian parameter estimation

Estimating the parameters of a statistical model in

→ language ~~vetue~~ Model Evaluation

→ Parameter estimation

→ Types of language models

→ language - specific modeling problem

→ multilingual & crosslingual language modelling

Multilingual vs crosslingual language modeling

Scope of training data :-

	Multilingual	Cross lingual
1) Scope of Training data	Trained on data from multiple languages simultaneously, aiming to understand and generate text in all included languages	Primarily trained on data from one & more languages but designed to transfer knowledge across languages, allowing to perform tasks in languages not explicitly seen during training.
2) Language specificity	capable of handling multiple languages with varying proficiency levels, treating them as a single unified model.	Focused on leveraging knowledge learned from one language to understand & generate text in another language, typically through transfer learning technique.
3) Data representation	often relies on shared representations across languages, enabling the model to generalize linguistic patterns across different languages.	Emphasizes aligning representations between languages, allowing information learned in one language to be applied effectively to another language.
4) Task adaptability	suited for a wide range of tasks (eg. translation, sentiment analysis, text generation) across multiple languages without task specific fine tuning.	Requires task specific adaptation & fine tuning to perform well on tasks in languages not part of the models training data.
5) Use cases	useful for applications requiring language-agnostic text processing such as search-engines, social media analysis and customer support.	Beneficial for tasks like cross-lingual information retrieval, low resource language processing, zero-shot translation where understanding multiple languages is essential for effective performance.

→ Language-specific modeling problems

Language-specific modeling problems refer to challenges that arise when developing NLP models tailored to a particular language. Some of these challenges.

① data availability:- Availability of large and high-quality datasets in some languages may be limited compared to widely spoken languages like English. This scarcity of data can hinder the development of effective language models for less-resourced languages.

② linguistic complexity:- Languages vary in terms of their grammatical structure, syntax, morphology and semantics. Building models that accurately capture these linguistic nuances requires language specific expertise and careful consideration of language-specific phenomena.

③ NER:- Named entities such as names of people, Organizations, locations etc, can vary significantly across languages due to cultural differences, naming conventions, and transliteration issues. Developing robust NER models that perform well across languages requires language specific training data and annotation guidelines.

④ machine translation:- language-specific challenges in machine translation include handling language-specific idioms, expressions, the word order variations. Translating between languages with vastly different linguistic properties poses additional difficulties, required specialized techniques and resources for each language pair.

5. Sentiment analysis:- Sentiment analysis models trained on one language may not generalize well to other languages due to differences in sentiment expressions, cultural contexts and linguistic nuances.