1. **List examples of real-world applications of NLP.**
   * Email platforms.
   * Voice-based assistants.
   * Modern search engines.
   * Machine translation.
2. **Explain the following NLP tasks:**
   * **Language modelling:** predicting the next word in a sentence.
   * **Text classification:** bucketing the text into a known set of categories.
   * **Information extraction:** extracting relevant information from text.
   * **Information retrieval:** finding documents relevant to a user query.
   * **Conversational agent:** building dialogue systems.
   * **Text summarization:** creating short summaries of longer documents.
   * **Question answering:** building systems that can auto answer questions
   * **Machine translation:** converting a piece of text from one language to another.
   * **Topic modelling:** uncovering the topical structure of a large collection of documents.
3. **What are the building blocks of language and their applications?**
   * **Phonemes:**
     + Speech to text.
     + Text to speech.
   * **Morphemes and Lexemes:**
     + Tokenization.
     + Word embeddings.
   * **Syntax:**
     + Parsing.
     + Entity extraction.
   * **Context:**
     + Summarization.
     + Topic Modeling.
4. **Why is NLP Challenging?**
   * Diversity across languages.
   * Common knowledge.
   * Ambiguity.
   * Creativity.
5. **How NLP, ML, and DL are related?**
   * **Artificial Intelligence (AI):** Branch of CS for building systems performing *tasks* that require human intelligence.
   * **Machine Learning (ML):** Branch of AI developing *algorithms* that can learn to perform tasks.
   * **Deep Learning (DL):** Branch of ML based on Artificial Neural Network Architectures
6. **Describe the heuristics-based NLP:**
   * Based on building rules for the task at hand.
7. **Explain briefly:**
   * **Naive Bayes:** classification algorithm that relies on Bayes’ theorem.
   * **Support Vector Machine (SVM):**
     + Classification algorithm**.**
     + **Strength:** robustness to noise.
     + **Weakness:** time taken to train.
   * **Hidden Markov Model (HMM):** statistical model that assumes there is an underlying unobservable process with hidden states generating data.
   * **Conditional Random Field (CRF):** classification algorithm for sequential data.
8. **What is the difference between:**
   * **Recurrent Neural Network (RNN):** remember what they have processed so far.
   * **LSTM NN:** type of RNN that let go irrelevant context thus perform better.
9. **How CNN can be used for text processing?**
   * **Convolutional NN:** by replacing each word in a sentence with its corresponding word vector.
10. **Describe the concept transfer learning.**
    * Knowledge gained while solving one problem is applied to a different but related problem.
11. **Give the architecture of autoencoder.**

**Diagram, schematic

Description automatically generated**

1. **List the key reason that makes DL not suitable for all NLP tasks.**
   * Overfitting on small datasets.
   * Common sense.
   * Domain adaptation.
   * Interpretable models.
   * On-device deployment.
2. **Explain the flow of conversation agents.**

**Diagram

Description automatically generated**

1. **What are the key stages of a generic pipeline for NLP system development?**
   * Data acquisition.
   * Text cleaning.
   * Pre-processing.
   * Feature engineering.
   * Modeling.
   * Evaluation.
   * Deployment.
   * Monitoring and model updating.
2. **How can we get data required for training an NLP technique?**
   * Use a public dataset.
   * Scrape data.
   * Product intervention.
   * Data augmentation.
3. **List the different data augmentation methods?**
   * Synonym replacement.
   * Replacing entities.
   * Back translation.
   * Bigram flipping.
   * Adding noise to data.
4. **Data can be collected from PDF files, HTML pages, and images, how this data can be cleaned based on their sources?**
   * Removing non-text info.
   * Converting text to required encoding format.
5. **Using dot (.) to segment sentences can cause problems, explain how?** 
   * As some abbreviations contains (.) such as (Dr.).
6. **What are the frequent steps in the data pre-processing phase?**
   * Removing Stop-word | Digits | Punctuation
   * Lowercasing.
   * Stemming & Lemmatization.
7. **With examples, explain the differences between segmentation and lemmatization.**
   * **Segmentation:** Dividing text into sentences at the appearance of full stops or question marks.
   * **Lemmatization:** Mapping different forms of a word to its base word (was -> be)
8. **What is the difference between code mixing and transliteration?**
   * **Code mixing**: phenomenon of switching between languages.
   * **Transliteration**: writing in specific language with other language spelling.
9. **Describe the concept coreference resolution.**
   * Finding all expressions referring to same entity in text.
10. **Explain the feature engineering for classical NLP versus DL-based NLP?**
    * **Classical NLP/ML Pipeline:** convert the raw data into a format that can be consumed by a machine.
    * **DL Pipeline:** raw data (after preprocessing) is directly fed to a model to learn from data and get better.
11. **How to combine heuristics directly or indirectly with the ML model?**
    * Create a feature from the heuristic.
    * Use it for your ML model before feeding it the data if it has high prediction for particular class.
12. **What is the difference between models ensembling and stacking?**
    * **Model Stacking:** feeding one model’s output as input for another model.
    * **Model Ensembling:** pool predictions from multiple models and make a final prediction.
13. **Which modeling technique can be used in the following cases of data:**
    * **Small data:** traditional ML solutions.
    * **Large data:** deep learning.
    * **Poor quality:** More data cleaning and pre-processing might be required.
    * **Good quality:** directly apply algorithms.
14. **What is the difference between:**
    * **Intrinsic evaluation** : The output compared against the corresponding label.
    * **Extrinsic evaluation :**  focuses on evaluating the model performance.
15. **What are the metrics that can be used in:**

* **Classification:** F1 score
* **Measuring model quality**: AUC
* **Information retrieval:** MAP.
* **Machine translation**: METEOR
* **Summarization tasks:** ROUGE

1. **Describe phases of NLP pipeline.**
   * **Deploying**: deployed as a web service.
   * **Monitoring**: model performance is monitored after deploying to make sure output is correct.
   * **Updating**: gathering new data after deploying to iterate the model based on them.
2. **Explain how the NLP pipeline is different from a language to another?** 
   * Some is very similar to English and others require us to rethink how we approach the problem.
3. **List the four categories of text representation techniques.**
   * Distributed representations
   * Universal language representation
   * Basic vectorization approaches
   * Handcrafted features
4. **Describe the concept vector space models.**
   * Converting text data into vectors.
5. **Explain the difference between:**
   * **Distributional similarity:** whenmeaning is defined by context
   * **Distributional hypothesis:** hypothesizes words that occur in similar contexts have similar meanings.
   * **Distributional representation:**
     + Representation schemes obtained based on distribution of word.
   * **Distributed representation:** compress the dimensionality of DR.
6. **Describe the wording embedding concept with an example of its use.**
   * Learned representation for text where words that have the same meaning have a similar representation.
   * Example: “USA” and other cities in the USA
   * **What are the important aspects to keep in mind while using word embeddings?** 
     + All text representations are biased based on what they saw in training data.
     + Need to encode specific aspects of text.
     + Large-sized files.
7. **Explain with an example the two architectural variants of Word2vec.** 
   * **CBOW :** predict center word from context.
   * **SkipGram:** predict the context from the center word.
8. **How the OOV problem can be solved?** 
   * Create vectors that are initialized randomly.
   * Bringing in characters and other subword-level .
   * Using subword information (such as prefixes and suffixes).
9. **What is the difference between:** 
   * **Word2vec:** learned representations for words.
   * **fastText:** learned representations for character.
   * **Doc2vec**: learn the representations for texts of random lengths.
10. **How high-dimensional data can be represented visually?** 
    * T-distributed Stochastic Neighboring Embedding.
11. **With example explain the use of handcrafted feature representations.** 
    * Help teachers choose reading materials for students.
    * Identifying sources of comprehension difficulty in texts.
12. **What is the difference between:**
    * **Binary classification**: If the number of classes is two.
    * **Multiclass classification:** If the number of classes is more than two.
    * **Multilabel classification:** a document can have one or more classes
13. **Give some applications of text classification.**
    * Language identification.
    * Content classification.
    * Segregate fake news from real news
    * Customer support.
    * E-commerce.
14. **Classification can be done without the text classification pipeline, explain how?** 
    * Create lists of positive & negative words in English.
15. **Describe with an example the confusion matrix of a classifier.**
    * Shows how many prediction are correct/incorrect per class.
16. **List the potential reasons for poor classifier performance.**
    * Bad learning algorithm.
    * Bad pre-processing.
    * Bad feature extraction mechanism.
    * Not tuning the classifier’s parameters and hyperparameters.
17. **How to solve class imbalance problem of a dataset?** 
    * Resample training set.
    * Resample with different ratios.
    * Use right evaluation metrics.
    * Design your own models.
18. **What is the difference between:** 
    * **Generative classifier:** learns the probability of a text for each class + maximum probability.
    * **Discriminative classifier**: learn the probability distribution over all classes
19. **How to use word embeddings as features for text classification?** 
    * Using the sentiment-labeled sentences dataset from the UCI repository
20. **Which technique is better for text classification CNN or LSTM and why?** 
    * **LSTM:** hungry for data and the way-to-go specially when the size of dataset is large.
21. **How text classification models can be interpreted?** 
    * One such tool is Lime.
22. **How to solve no training and less training data problems?** 
    * **No Training Data:** create an annotated dataset.
    * **Less Training Data:** by active learning.
23. **Give some options to explore when no labels exist for a dataset.**
    * Active learning.
    * Use public datasets.
    * Use existing APIs or libraries
    * Utilize weak supervision
    * Learning from implicit & explicit feedback.