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.

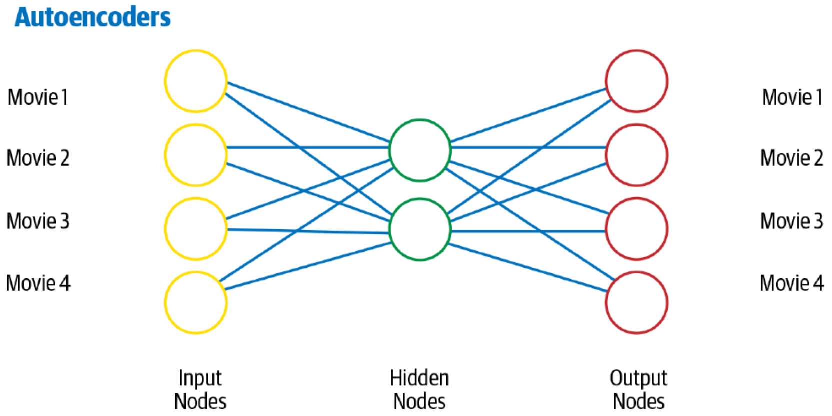
9- How CNN can be used for text processing?

- **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.

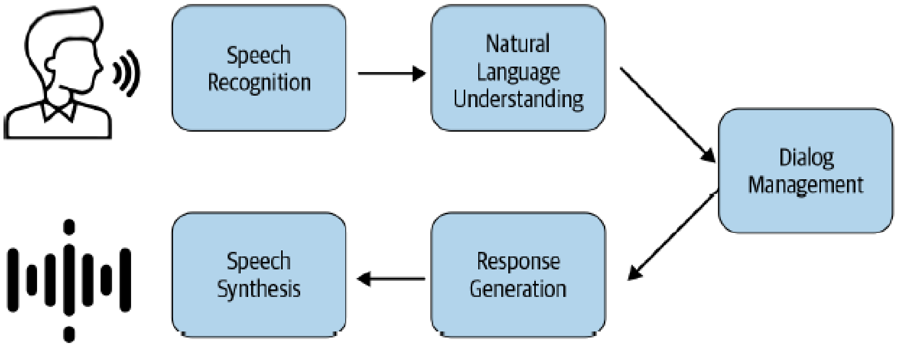
11- Give the architecture of autoencoder.



12- 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.

13- Explain the flow of conversation agents.



14- What are the key stages of a <u>generic pipeline</u> for NLP system development?

- Data acquisition.
- Text cleaning.
- Pre-processing.
- Feature engineering.
- Modeling.
- Evaluation.
- Deployment.
- Monitoring and model updating.

15- How can we get data required for training an NLP technique?

- Use a public dataset.
- Scrape data.
- Product intervention.
- Data augmentation.

16- List the different data augmentation methods?

- Synonym replacement.
- Replacing entities.
- Back translation.
- Bigram flipping.
- Adding noise to data.

17- 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.

18- Using dot (.) to segment sentences can cause problems, explain how?

- As some abbreviations contains (.) such as (Dr.).

19- What are the frequent steps in the data pre-processing phase?

- Removing <u>Stop-word | Digits | Punctuation</u>
- Lowercasing.
- Stemming & Lemmatization.

20- 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)

21- 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.

22- Describe the concept coreference resolution.

- Finding all expressions referring to same entity in text.

23- 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.

24- 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.

25- 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.

26- Which modeling technique can be used in the following cases of data:

- **Small data:** traditional ML solutions.
- **Large data:** deep learning.
- **Poor quality:** More <u>data cleaning and pre-processing</u> might be required.
- **Good quality:** directly apply algorithms.

27- What is the difference between:

- **Intrinsic evaluation** : The <u>output</u> compared against the corresponding <u>label</u>.
- **Extrinsic evaluation :** focuses on <u>evaluating the model performance</u>.

28- 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

29- 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.

30- 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.

31- Describe the NLP pipeline for ranking tickets in a ticketing system by Uber.?

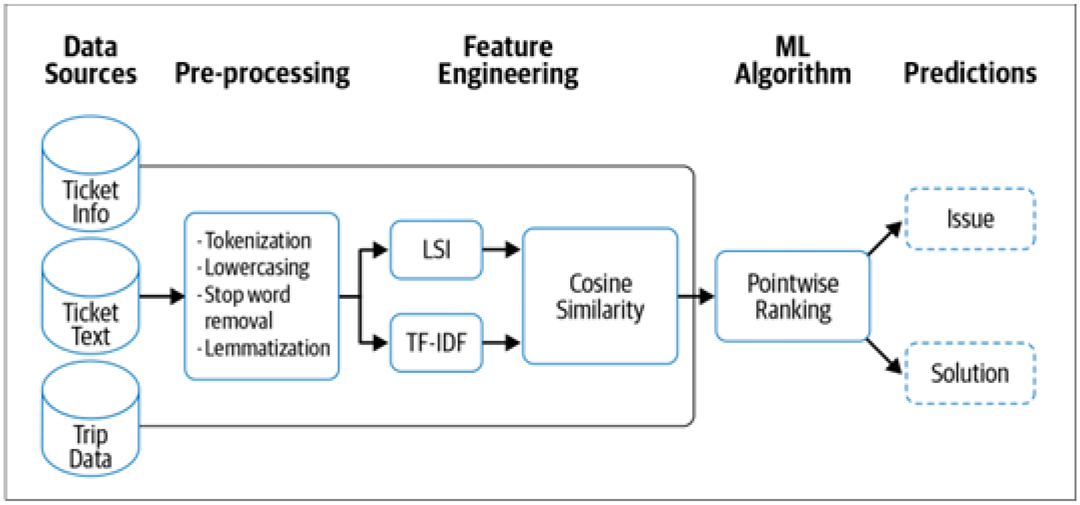


Figure 2-15. NLP pipeline for ranking tickets in a ticketing system by Uber [59]