1. What are Sequence-to-sequence models?
2. What are the Problem with Vanilla RNNs?
3. What is Gradient clipping?
4. Explain Attention mechanism
5. Explain Conditional random fields (CRFs)
6. Explain self-attention
7. What is Bahdanau Attention?
8. What is a Language Model?
9. What is Multi-Head Attention?
10. What is Bilingual Evaluation Understudy (BLEU)?

Answer:

1. Sequence-to-sequence (Seq2Seq) models are a type of neural network architecture used for mapping an input sequence to an output sequence. They are commonly used in natural language processing (NLP) applications such as machine translation, summarization, and question answering.
2. Vanilla RNNs suffer from the vanishing gradient problem, which occurs when the gradients propagated through the network become very small and cause the network to learn slowly. This problem is especially pronounced when training on long sequences, as the gradients must be propagated through many time steps.
3. Gradient clipping is a technique used in machine learning to prevent gradients from becoming too large during training. It involves setting a maximum value for the gradients, and if the computed gradient exceeds this value, it is clipped to the maximum value.
4. Attention mechanism is a neural network component that allows the model to selectively focus on certain parts of the input sequence when making predictions. It works by assigning a weight to each input element based on its relevance to the current prediction. These weights are then used to compute a weighted sum of the input elements, which is passed through the rest of the network.
5. Conditional random fields (CRFs) are a type of probabilistic model used for sequence labeling tasks such as named entity recognition and part-of-speech tagging. They model the conditional probability of a label sequence given an input sequence and are trained to maximize the likelihood of the correct label sequence.
6. Self-attention is an attention mechanism where the attention weights are computed from the input sequence itself rather than a separate context vector. It allows the model to capture relationships between different elements of the input sequence.
7. Bahdanau Attention is a specific type of attention mechanism used in sequence-to-sequence models. It computes the attention weights based on a weighted combination of the decoder hidden state and the encoder outputs, allowing the decoder to selectively attend to different parts of the input sequence during decoding.
8. A Language Model is a type of machine learning model used for predicting the probability of a sequence of words. It is trained on a large corpus of text data and learns to predict the probability of the next word given the previous words in the sequence.
9. Multi-Head Attention is an extension of the attention mechanism where the model attends to multiple parts of the input sequence simultaneously, using multiple sets of attention weights. This allows the model to capture more complex relationships between different elements of the input sequence.
10. Bilingual Evaluation Understudy (BLEU) is a metric used to evaluate the quality of machine translations. It measures the similarity between a machine-generated translation and one or more reference translations based on the overlap of n-gram sequences. A higher BLEU score indicates a better quality translation.