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)

Answers:

1. Sequence-to-sequence (Seq2Seq) models are a class of neural network architectures designed for mapping an input sequence to an output sequence, where the input and output sequences may have different lengths and structures. This is achieved by using an encoder to convert the input sequence into a fixed-length vector representation and then using a decoder to generate the output sequence based on this representation. Seq2Seq models have been successfully applied to a range of tasks such as machine translation, speech recognition, and text summarization.
2. The problem with vanilla RNNs is that they suffer from vanishing and exploding gradients, which make it difficult to learn long-term dependencies in sequential data. When the gradients become too small, the model cannot effectively update its parameters and learn from the data. When the gradients become too large, the model can become unstable and diverge during training.
3. Gradient clipping is a technique used to prevent the exploding gradient problem by scaling the gradients if their norm exceeds a certain threshold. This helps to prevent the gradients from becoming too large and causing the model to diverge during training.
4. Attention mechanism is a technique used to improve the performance of Seq2Seq models by allowing the decoder to selectively focus on different parts of the input sequence when generating each output token. This is achieved by computing a set of attention weights that indicate the relative importance of each input token for generating the current output token. The attention weights are then used to compute a weighted sum of the input sequence, which is used as an additional input to the decoder.
5. Conditional random fields (CRFs) are a type of probabilistic graphical model commonly used for sequence labeling tasks such as named entity recognition and part-of-speech tagging. CRFs model the conditional probability distribution of the output sequence given the input sequence and can take into account global constraints such as label dependencies and consistency.
6. Self-attention is an attention mechanism used in transformer-based architectures, which computes attention weights based solely on the input sequence itself, rather than a separate context vector. This allows the model to selectively attend to different parts of the input sequence and capture long-range dependencies.

7. Bahdanau Attention is a specific type of attention mechanism used in Seq2Seq models, which computes attention weights based on a learned alignment score between the decoder hidden state and each encoder hidden state. This allows the decoder to selectively attend to different parts of the input sequence when generating each output token.
8. A language model is a type of statistical model that predicts the probability of a sequence of words or tokens. Language models can be trained on large amounts of text data and used for a variety of natural language processing tasks such as text generation, machine translation, and speech recognition.
9. Multi-Head Attention is an extension of the attention mechanism used in transformer-based architectures, which computes multiple sets of attention weights using different linear projections of the input sequence. This allows the model to attend to different aspects of the input sequence and capture different types of information.
10. Bilingual Evaluation Understudy (BLEU) is a metric used for evaluating the quality of machine translation systems. BLEU compares the output of the system to one or more reference translations and computes a score based on the overlap between the n-grams (contiguous sequences of n words) in the output and reference translations. BLEU scores range from 0 to 1, with higher scores indicating better translation quality.