1. What are Sequence-to-sequence models?  
   A Seq2Seq model is a model that takes a sequence of items (words, letters, time series, etc) and outputs another sequence of items. In the case of Neural Machine Translation, the input is a series of words, and the output is the translated series of words.
2. What are the Problem with Vanilla RNNs?  
   A major issue with the vanilla RNN is that they suffers from vanishing/exploding gradients similarly to issues with deep feedforward networks. At each timestep, the hidden state ht is multiplied by W, at the last timestep, the value of ht is effectively multiplied by W
3. What is Gradient clipping?  
   Gradient clipping is a technique that tackles exploding gradients. The idea of gradient clipping is very simple: If the gradient gets too large, we rescale it to keep it small.
4. Explain Attention mechanism  
   The attention mechanism emerged as an improvement over the encoder decoder-based neural machine translation system in natural language processing (NLP). Later, this mechanism, or its variants, was used in other applications, including computer vision, speech processing, etc.
5. Explain Conditional random fields (CRFs)  
   Conditional random fields (CRFs) are a class of statistical modeling methods often applied in pattern recognition and machine learning and used for structured prediction. Whereas a classifier predicts a label for a single sample without considering "neighbouring" samples, a CRF can take context into account.
6. Explain self-attention  
   The self-attention decoder allows each position to attend each position up to and including that position.
7. What is Bahdanau Attention?  
   proposed an attention mechanism that learns to align and translate jointly. It is also known as Additive attention as it performs a linear combination of encoder states and the decoder states.
8. What is a Language Model?  
   A language model takes text input and outputs the next word or character. The input is often called a “context” or “history”, because it represents what has been written so far.
9. What is Multi-Head Attention?  
   Multi-head Attention is a module for attention mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension.
10. What is Bilingual Evaluation Understudy (BLEU)

Bilingual Evaluation Understudy (BLEU) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another.