**NLP Assignment 5 Solution**

**1. What are Sequence-to-sequence models?**

**Ans:** Sequence-to-sequence (Seq2Seq) models are a type of neural network architecture that is used for tasks like machine translation, text summarization, and conversation models. The architecture consists of an encoder and a decoder, which work together to transform an input sequence into an output sequence. The encoder reads in the input sequence and encodes it into a vector representation, while the decoder takes the vector representation and decodes it into the output sequence. This architecture has been used to great success in a variety of tasks and is a powerful tool for natural language processing.

**2. What are the Problem with Vanilla RNNs?**

**Ans:** The primary problem with vanilla RNNs is that they are prone to the vanishing gradient problem, which is a phenomenon in which the gradients of the network weights can become so small during training that they are nearly impossible to update. This is because the gradients are repeatedly multiplied together over time and can quickly become much smaller than they were initially. Additionally, vanilla RNNs struggle to remember information over long time periods due to their limited capacity for long-term memory. Lastly, vanilla RNNs are difficult to train due to their sequential nature, as they require all of the previous data points to be input into the network before the current data point can be processed.

**3. What is Gradient clipping?**

**Ans:** Gradient clipping is a technique used to prevent the gradients in a neural network from becoming too large. It involves clipping the gradients to a predefined maximum value, which helps to prevent the gradients from exploding, which can lead to instability in training. This technique can be used to help the network converge faster and more accurately.

**4. Explain Attention mechanism**

**Ans:** A attention mechanism is a mechanism that allows a machine learning model to focus on a specific part of a given input. It is a mechanism that allows a model to focus on the most relevant parts of the input, while ignoring irrelevant parts. Attention mechanisms are used in many applications, such as natural language processing, computer vision, and reinforcement learning. The most common form of attention is the soft attention mechanism, where the model assigns weights to different parts of the input to emphasize the most important parts. By doing so, the model can better focus on the relevant information and make more accurate predictions.

**5. Explain Conditional random fields (CRFs)**

**Ans:** Conditional random fields (CRFs) are a type of discriminative probabilistic model often used for labeling or parsing structured data. CRFs are a type of graphical model, meaning that they use a graph-based structure to represent the relationships between variables in the model. CRFs are used for a variety of tasks, such as part-of-speech tagging, named entity recognition, and object recognition. Unlike other probabilistic models, such as hidden Markov models or naive Bayes classifiers, CRFs are able to take into account the relationships between variables in the model. This is done by defining a probability distribution over a set of output variables that depends on a set of input variables. This allows for more accurate predictions, as the model is able to learn how different variables interact with each other.

**6. Explain self-attention**

**Ans:** Self-attention is a type of neural network layer that allows for a more direct representation of relationships between input elements. It does this by computing the attention weights of each element with respect to all the other elements. This allows the model to better capture the context of the input elements and the relationships between them, leading to improved performance. Self-attention is used in a variety of models such as transformer networks, transformer-based language models, and vision transformer networks.

**7. What is Bahdanau Attention?**

**Ans:** Bahdanau Attention is a type of attention mechanism developed by Dzmitry Bahdanau and Kyunghyun Cho. It is a type of Neural Machine Translation (NMT) system that uses an encoder-decoder architecture with attention. The attention mechanism works by taking a set of input vectors and computing a context vector that is used to weigh each input vector. The weighted vectors are then combined to form an output vector. The output vector is used to decode the target language sentence. The attention mechanism helps the model to focus on specific parts of the input sentence and helps to improve the accuracy of the translation.

**8. What is a Language Model?**

**Ans:** A language model is a probability distribution over sequences of words. It is a type of artificial intelligence algorithm used to predict the next word or phrase in a sequence based on the words that have already been inputted. Language models are typically used in natural language processing (NLP) applications such as machine translation, speech recognition, and text generation.

**9. What is Multi-Head Attention?**

**Ans:** Multi-Head Attention is a type of attention mechanism used in neural network architectures such as transformers. It is used to capture multiple aspects of a sequence of data by applying attention to multiple different parts of the input sequence. This allows the model to learn a more complex representation of the data and to better capture complex patterns.

**10. What is Bilingual Evaluation Understudy (BLEU)**

**Ans:** BLEU (Bilingual Evaluation Understudy) is a method to evaluate the quality of machine translated texts. It was first proposed in 2002 by Kishore Papineni and colleagues as an alternative to existing methods of evaluating machine-generated translations. BLEU uses a modified version of precision to measure the quality of the translation. The score is calculated by comparing the predicted translation to a set of reference translations and computing the percentage of words in the predicted translation that also appear in the references.