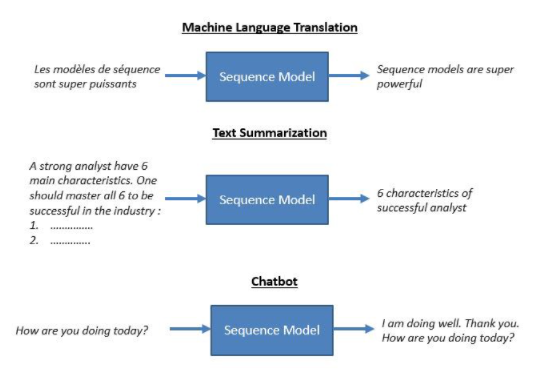
**1. What are Sequence-to-sequence models?**

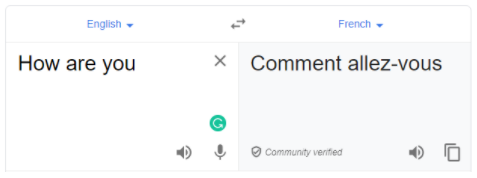
Ans: Sequence to Sequence (often abbreviated to seq2seq) models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.



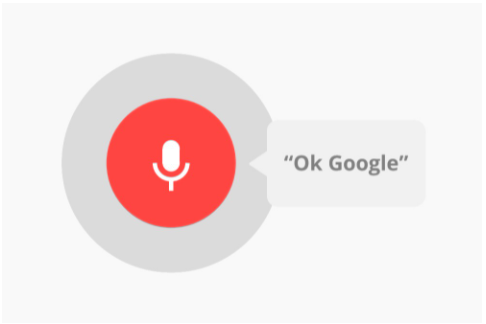
Use Cases of the Sequence to Sequence Models: Sequence to sequence models lies behind numerous systems that you face on a daily basis. For instance, seq2seq model powers applications like Google Translate, voice-enabled devices, and online chatbots.

The following are some of the applications:

Machine translation — a 2016 paper from Google shows how the seq2seq model’s translation quality “approaches or surpasses all currently published results”.



Speech recognition — another Google paper that compares the existing seq2seq models on the speech recognition task.



These are only some applications where seq2seq is seen as the best solution. This model can be used as a solution to any sequence-based problem, especially ones where the inputs and outputs have different sizes and categories.

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

Ans: RNNs suffer from the problem of vanishing gradients, which hampers learning of long data sequences. The gradients carry information used in the RNN parameter update and when the gradient becomes smaller and smaller, the parameter updates become insignificant which means no real learning is done.

**3. What is Gradient clipping?**

Ans: Gradient clipping is a technique to prevent exploding gradients in very deep networks, usually in recurrent neural networks. A neural network is a learning algorithm, also called neural network or neural net, that uses a network of functions to understand and translate data input into a specific output.

This type of learning algorithm is designed based on the way neurons function in the human brain. There are many ways to compute gradient clipping, but a common one is to rescale gradients so that their norm is at most a particular value. With gradient clipping, pre-determined gradient threshold be introduced, and then gradients norms that exceed this threshold are scaled down to match the norm. This prevents any gradient to have norm greater than the threshold and thus the gradients are clipped. There is an introduced bias in the resulting values from the gradient, but gradient clipping can keep things stable.

**4. Explain Attention mechanism**

Ans: The attention mechanism was introduced to improve the performance of the encoder-decoder model for machine translation. The idea behind the attention mechanism was to permit the decoder to utilize the most relevant parts of the input sequence in a flexible manner, by a weighted combination of all of the encoded input vectors, with the most relevant vectors being attributed the highest weights.

The attention mechanism was introduced by Bahdanau et al. (2014), to address the bottleneck problem that arises with the use of a fixed-length encoding vector, where the decoder would have limited access to the information provided by the input. This is thought to become especially problematic for long and/or complex sequences, where the dimensionality of their representation would be forced to be the same as for shorter or simpler sequences.

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

Ans: 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**

Ans: Self Attention, also called intra Attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. It has been shown to be very useful in machine reading, abstractive summarization, or image description generation.

**7. What is Bahdanau Attention?**

Ans: Conventional encoder-decoder architectures for machine translation encoded every source sentence into a fixed-length vector, irrespective of its length, from which the decoder would then generate a translation. This made it difficult for the neural network to cope with long sentences, essentially resulting in a performance bottleneck.

The Bahdanau attention was proposed to address the performance bottleneck of conventional encoder-decoder architectures, achieving significant improvements over the conventional approach.

The Bahdanau attention mechanism has inherited its name from the first author of the paper in which it was published.

It follows the work of Cho et al. (2014) and Sutskever et al. (2014), who had also employed an RNN encoder-decoder framework for neural machine translation, specifically by encoding a variable-length source sentence into a fixed-length vector. The latter would then be decoded into a variable-length target sentence.

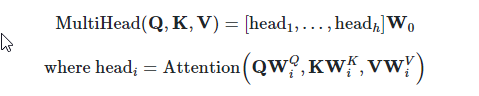
Bahdanau et al. (2014) argue that this encoding of a variable-length input into a fixed-length vector squashes the information of the source sentence, irrespective of its length, causing the performance of a basic encoder-decoder model to deteriorate rapidly with an increasing length of the input sentence. The approach they propose, on the other hand, replaces the fixed-length vector with a variable-length one, to improve the translation performance of the basic encoder-decoder model.

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

Ans: Language modeling (LM) is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. Language models analyze bodies of text data to provide a basis for their word predictions. They are used in natural language processing (NLP) applications, particularly ones that generate text as an output. Some of these applications include , machine translation and question answering.

9. What is Multi-Head Attention?

Ans: 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. Intuitively, multiple attention heads allows for attending to parts of the sequence differently (e.g. longer-term dependencies versus shorter-term dependencies).



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

Ans: BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations.

Although developed for translation, it can be used to evaluate text generated for a suite of natural language processing tasks.

The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence.

A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

The score was developed for evaluating the predictions made by automatic machine translation systems. It is not perfect, but does offer 5 compelling benefits:

It is quick and inexpensive to calculate.

It is easy to understand.

It is language independent.

It correlates highly with human evaluation.

It has been widely adopted.