WHAT IS CLAIMED IS:

1. A neural machine translation system implemented by one or more computers, the neural machine translation system configured to receive an input sequence of input tokens representing a first sequence of words in a first natural language and to generate an output sequence of output tokens representing a second sequence of words that is a translation of the first sequence into a second natural language, the neural machine translation system comprising:

an encoder neural network comprising:

an input forward long short-term memory (LSTM) layer configured to process each input token in the input sequence in a forward order to generate a respective forward representation of each input token,

an input backward LSTM layer configured to process each input token in the input sequence in a backward order to generate a respective backward representation of each input token,

a combining layer configured to, for each input token in the input sequence, combine the forward representation of the input token and the backward representation of the input token to generate a combined representation of the input token, and

a plurality of hidden LSTM layers configured to process each combined representation in the forward order to generate a respective encoded representation of each of the input tokens in the input sequence; and

a decoder subsystem configured to receive the respective encoded representation of each of the input tokens in the input sequence and to process the encoded representations to generate the output sequence, wherein a bottom layer of the encoder neural network is a bi-directional LSTM layer that combines the input forward LSTM layer, the input backward LSTM layer, and the combining layer that process the input sequence in both the forward order and the backward order, and wherein the bottom layer that is bi-directional is followed by the plurality of hidden LSTM layers that process the input sequence received from the bottom layer in the forward order as uni-directional LSTM layers.

2. The neural machine translation system of claim 1, wherein the hidden LSTM layers are arranged in a stack one after the other, and wherein the plurality of hidden LSTM layers comprises at least seven LSTM layers.

3. The neural machine translation system of claim 2, wherein at least two of the plurality of hidden LSTM layers are connected with a residual connection.

4. The neural machine translation system of any one of claims 1-3, wherein the decoder subsystem comprises:

a decoder neural network, the decoder neural network comprising:

a plurality of LSTM layers arranged in a stack one on top of the other and configured to, for each of a plurality of positions in the output sequence:

receive an attention context vector and the output token at a preceding position in the output sequence, and

process the attention context vector and the token at the preceding position in the output sequence to generate an LSTM output for the position, and

a softmax output layer, the softmax output layer configured to, for each of the plurality of positions, receive the LSTM output for the position and to generate a respective score for each output token in a vocabulary of output tokens.

5. The neural machine translation system of claim 4, wherein the plurality of LSTM layers comprises at least eight LSTM layers.

6. The neural machine translation system of any one of claims 4 or 5, wherein at least two of the plurality of LSTM layers are connected with a residual connection.

7. The neural machine translation system of claim 6, wherein a first LSTM layer in the stack is configured to, for each of the plurality of positions:

receive a first layer input comprising the attention context vector, a layer output generated by a LSTM layer directly below the first LSTM layer in the stack, and a layer input to the LSTM layer directly below the first LSTM layer in the stack; and

process the first layer input in accordance with a current hidden state of the first LSTM layer to generate a first layer output and to update the current hidden state.

8. The neural machine translation system of any one of claims 4-7, wherein the decoder system further comprises:

an attention subsystem, wherein the attention subsystem is configured to, for each of the plurality of positions:

receive an attention input comprising a bottom layer output generated by a bottom LSTM layer in the stack of LSTM layers in the decoder neural network for the preceding position; and

process the attention input to generate the attention context vector for the time step.

9. The neural machine translation system of claim 8, wherein processing the bottom layer output and the respective encoded representation of each of the input tokens in the input sequence to generate the attention context vector for the time step comprises, for each of the plurality of time steps:

processing the bottom layer output and the encoded representations through one or more neural network layers to generate a respective score for each of the input tokens;

generating a respective weight for each of the input tokens from the respective scores; and

determining a weighted sum of the encoded representations in accordance with the weights.

10. The neural machine translation system of any one of claims 1-9, wherein the input and output tokens are wordpieces.

11. The neural machine translation system of any one of claims 1-10, wherein the encoder neural network and the decoder subsystem are configured to perform quantized computations during generation of the output sequence.

12. A method of generating an output sequence of output tokens that represents a second sequence of words that is a translation of a first sequence of words into a second natural language, the method comprising:

obtaining a plurality of candidate output sequences from a neural machine translation system;

maintaining at most a predetermined number of the candidate output sequences in a beam for further consideration by the neural machine translation system, comprising:

assigning a respective score to each candidate output sequence;

ranking the candidate output sequences based on the respective scores;

removing from the beam all candidate output sequences other than the predetermined number of highest-ranked candidate output sequences according to the ranking;

determining whether any of the candidate output sequences in the beam have been identified as a finalized candidate output sequence by the neural machine translation system; and

when at least one of the candidate output sequences has been identified as a finalized candidate output sequence:

decreasing a number of additional iterations that need to be performed by removing from the beam all candidate output sequences that have a score that is more than a threshold value lower than the score for a highest-ranked finalized candidate output sequence.

13. The method of claim 12, wherein determining whether any of the candidate output sequences have been identified as a finalized candidate output sequence by the neural machine translation system comprises:

determining whether any of the candidate output sequence ends with a predetermined end-of-sentence output token.

14. The method of any one of claims 12-13, wherein assigning a respective score to each candidate output sequence comprises:

determining a log likelihood of the candidate output sequence according to output scores generated by the neural machine translation system; and

normalizing the log likelihood based on a number of output tokens in the candidate output sequence to generate a length normalized log likelihood.

15. The method of claim 14, wherein assigning a respective score to each candidate output sequence further comprises:

modifying the length normalized log likelihood using a coverage penalty to favor candidate output sequences that fully cover the words in the first sequence.

16. The method of any one of claims 12-15, further comprising:

when all of the candidate output sequences in the beam have been identified as a finalized candidate output sequence by the neural machine translation system, selecting a highest-scoring candidate output sequence as the output sequence.

17. One or more computer storage media encoded with instructions that, when executed by one or more computers, cause the one or more computers to implement the system of any one of claims 1-11.

18. One or more computer storage media encoded with instructions that, when executed by one or more computers, cause the one or more computers to perform the operations of the respective method of any one of claims 12-16.

19. A system comprising one or more computers and one or more storage devices storing instructions that, when executed by the one or more computers, cause the one or more computers to perform the operations of the respective method of any one of claims 12-16.