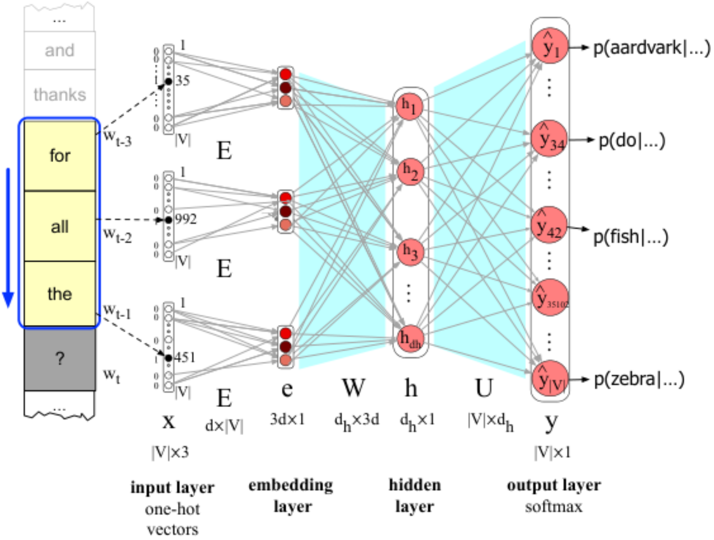
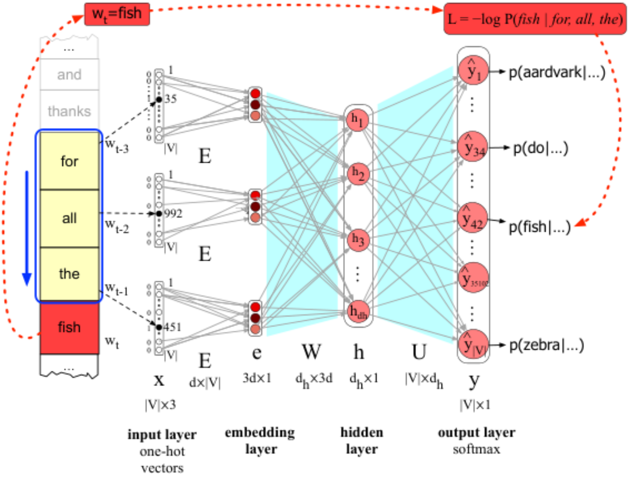
Forward Inference in a Feedforward Neural Language Model

Figure 7.13. Shows a context of three preceding tokens



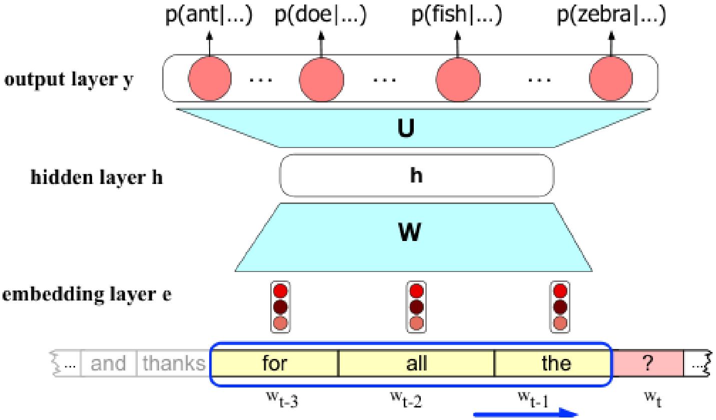
Learning Embeddings

Figure 7.18. Learn embeddings based on loss with respect to actual word



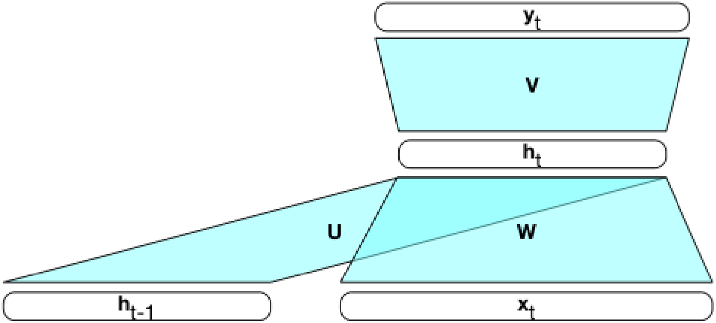
Forward Inference: Sliding Window

Figure 9.1



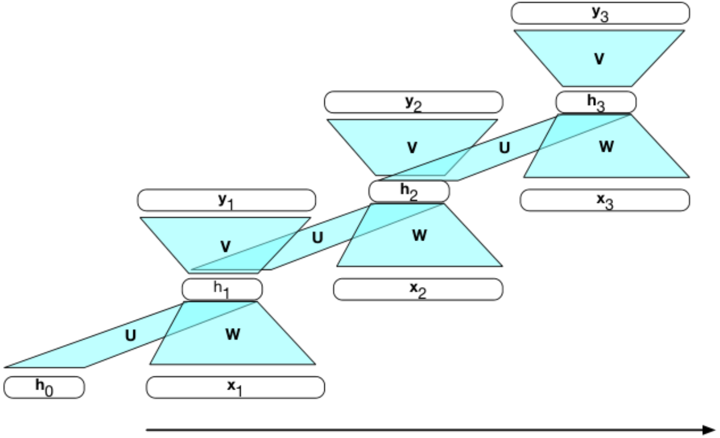
Recurrent Neural Network (RNN)

Figure 9.3. The hidden state is incrementally built up



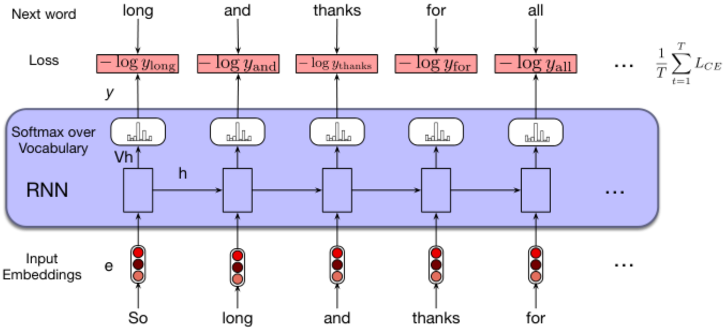
RNN Unrolled Over Time

Figure 9.5. Notice the long chain



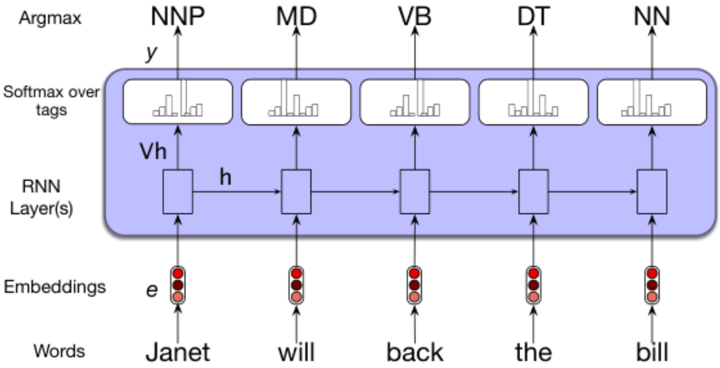
Training an RNN as a Language Model

Figure 9.6. Trains iteratively; uses correct token for subsequent steps



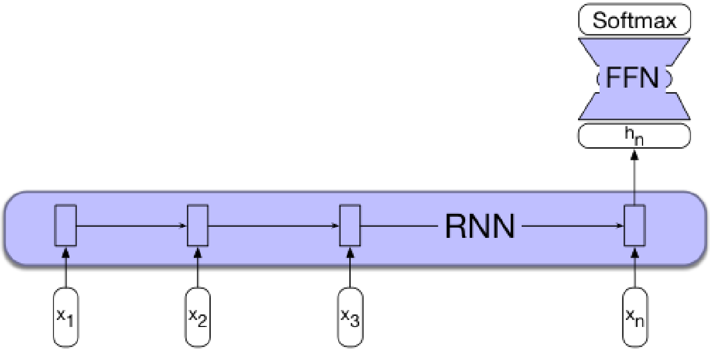
POS Tagging via an RNN

Figure 9.7. Example of sequence labeling



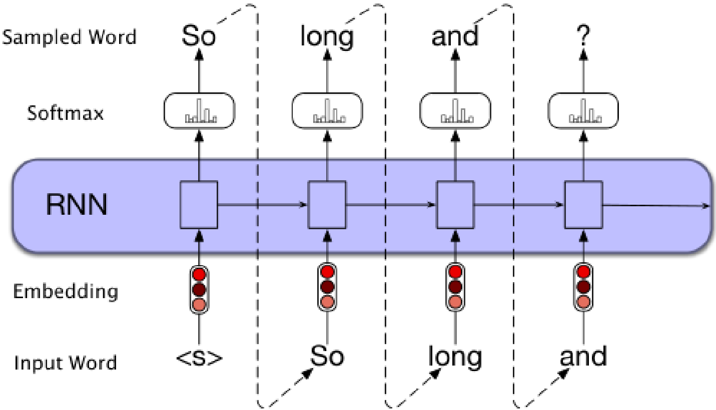
Sequence Classification

Figure 9.8. Uses the last hidden state to classify



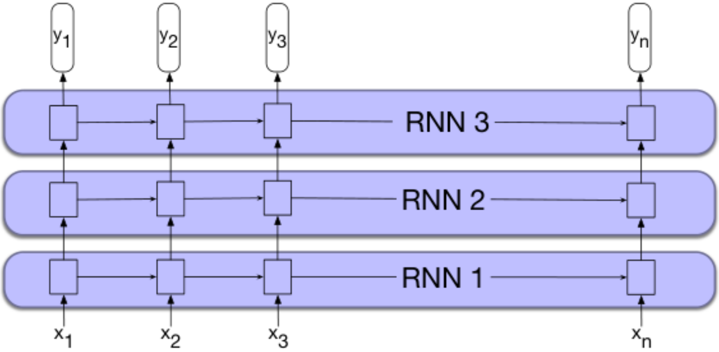
Autoregressive Generation with an RNN Language Model

Figure 9.9



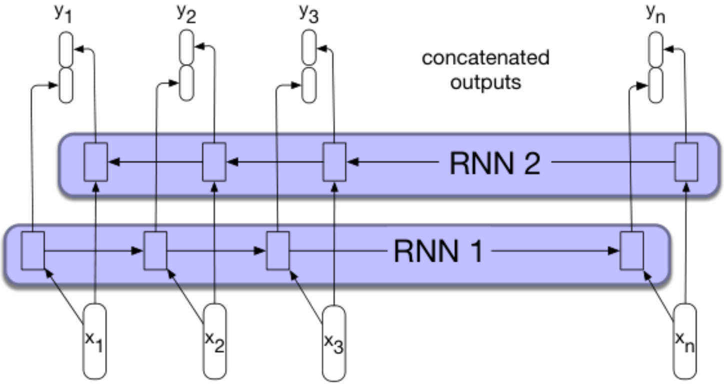
Stacked RNNs

Figure 9.10. Each layer captures a distinct level of abstraction



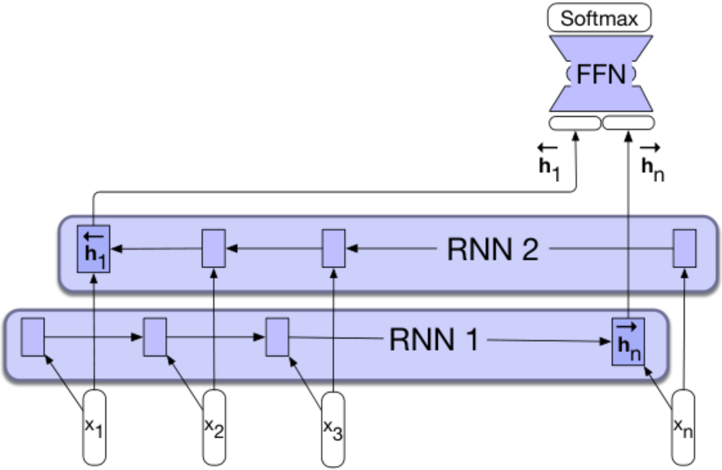
Bidirectional RNN

Figure 9.11. Each output is a concatenation of the forward and backward outputs



Bidirectional RNN for Sequence Classification

Figure 9.12. Uses the last hidden states of forward and backward components

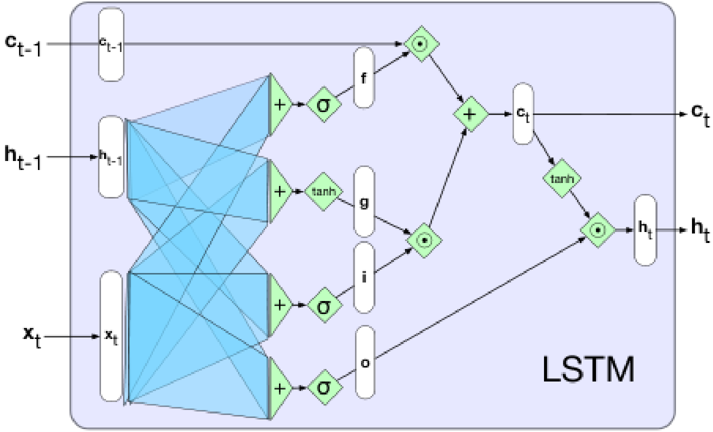


Long Short-Term Memory (LSTM) Unit, Computationally

Figure 9.13.

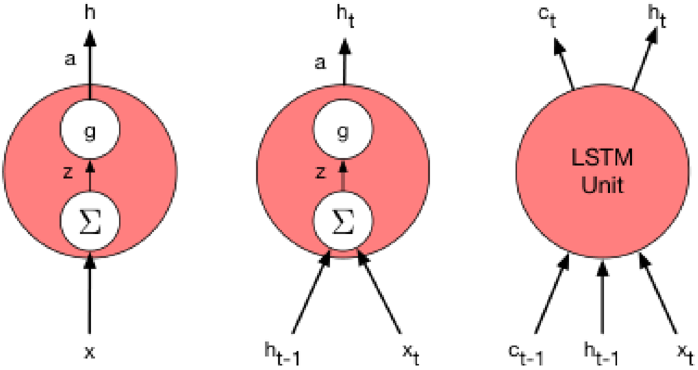
Inputs: current token, previous hidden, previous context

Outputs: new hidden, new context



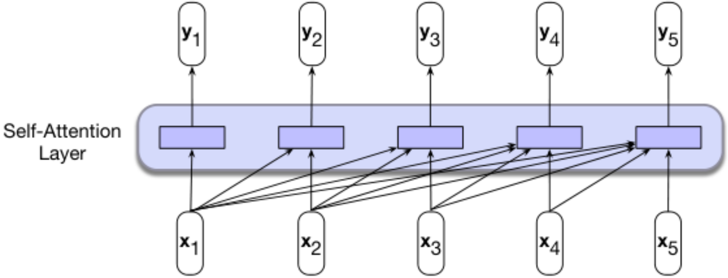
Comparing Neural Units

Figure 9.14. Feedforward neuron; RNN unit; LSTM unit



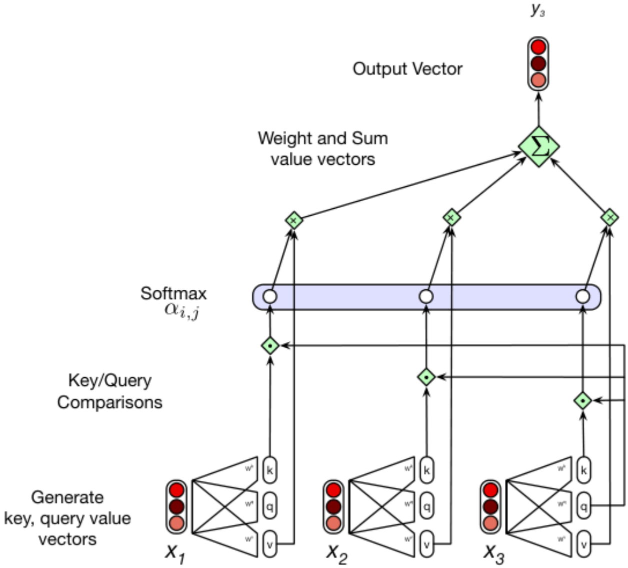
Self-Attention: Information Flow

Figure 9.15. Each unit attends to all previous tokens
Unlike in RNNs, there is no flow between the units



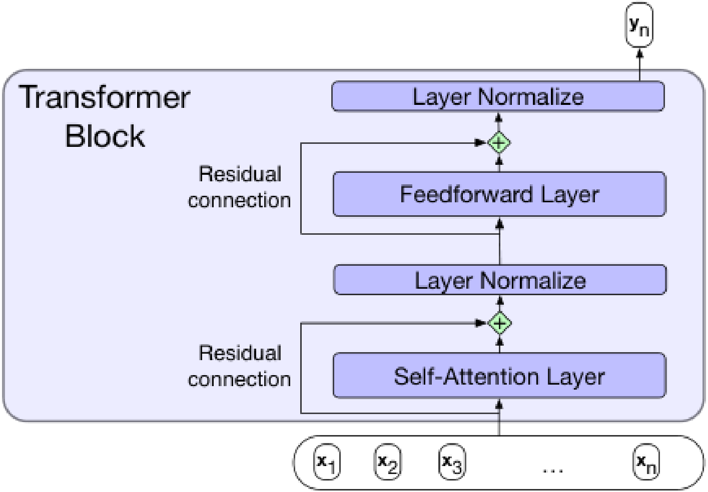
Query-Key-Value Paradigm for Self-Attention

Figure 9.16



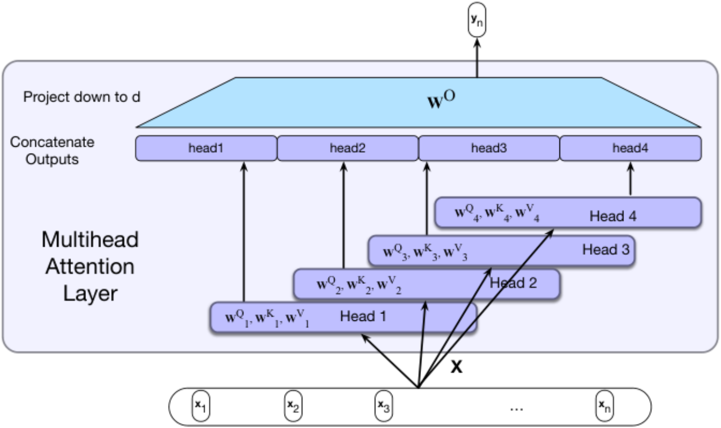
Transformer Block

Figure 9.18. Residual connections are ways to bypass complex layers that improve learning



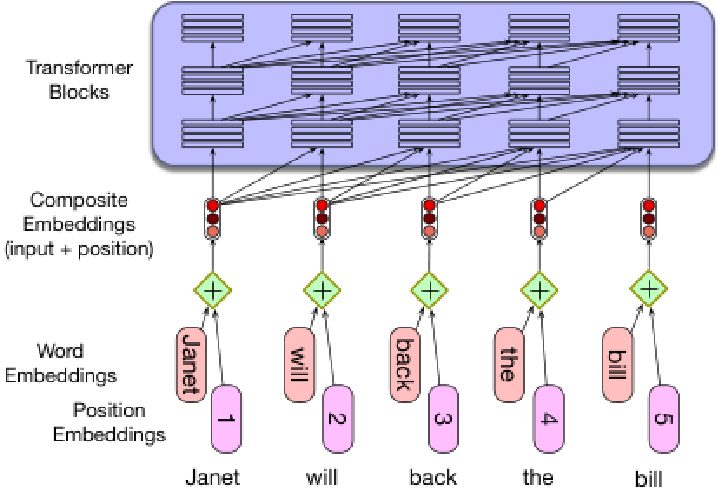
Multihead Self-Attention: Capturing Distinct Concerns

Figure 9.19. Separate heads (separate query-key-value matrices) for syntax, semantics, discourse, . . .



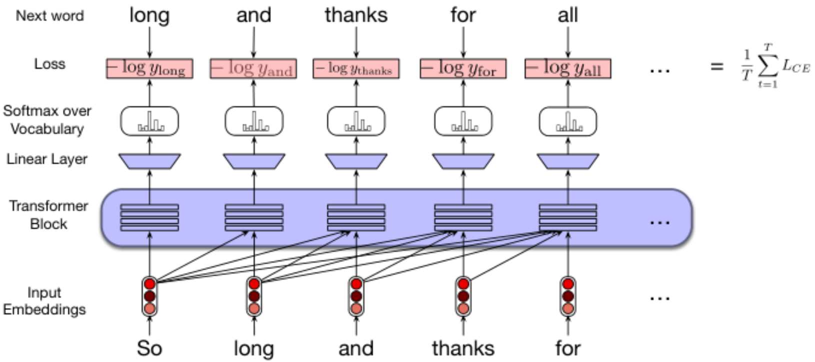
Positional Embeddings to Model Word Order

Figure 9.20. Learn embeddings for each position similarly to token embeddings add position embeddings to embeddings of the respective tokens



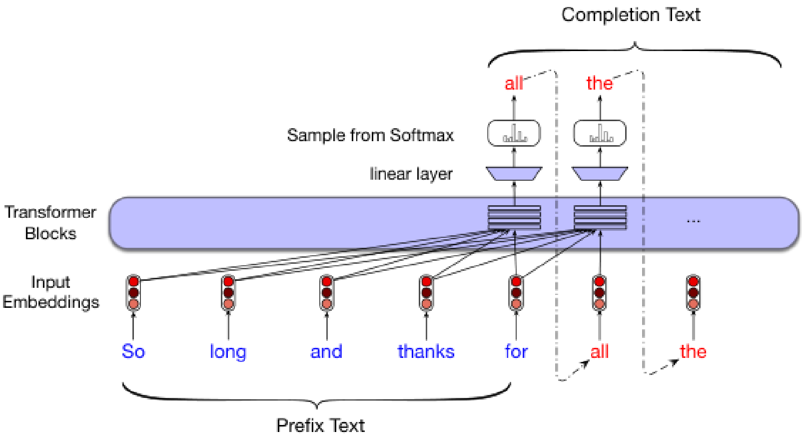
Training a Transformer as a Language Model

Figure 9.21



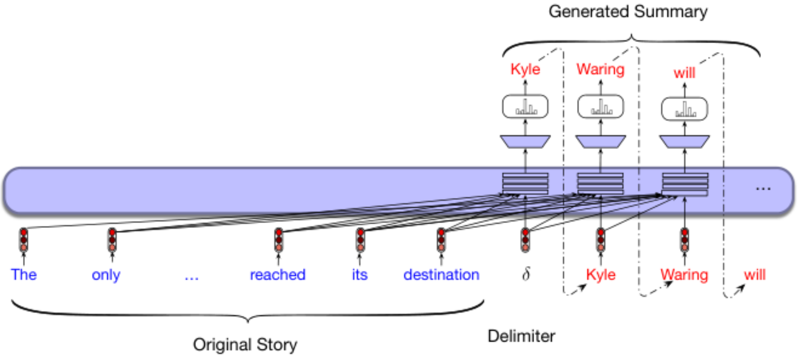
Autoregressive Text Completion with Transformers

Figure 9.22



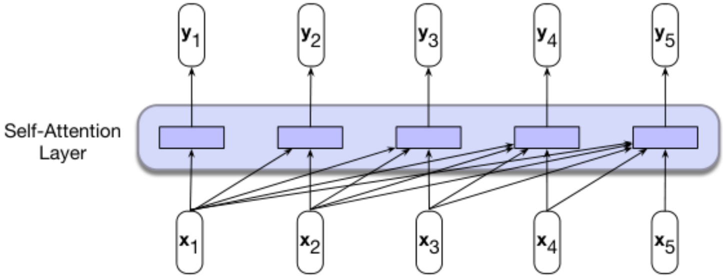
Summarization with Transformers

Figure 9.24. Train with actual story-summary pairs



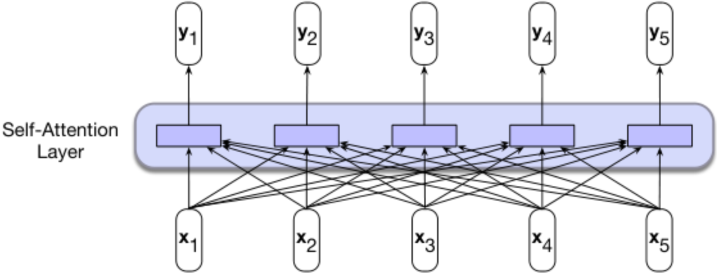
Causal, Backward Looking Transformer

Figure 11.1 (= 9.15). Causal because it doesn't look at "future" tokens



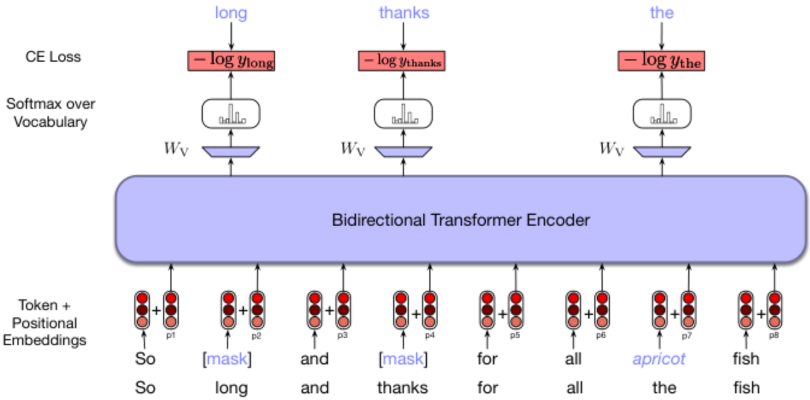
Bidirectional Self-Attention Model

Figure 11.2. Looks at future (subsequent) tokens



Masked Language Model Training

Figure 11.5. In BERT, 15% tokens are sample, of which 80% become [MASK], 10% become another random toke, 10% remain unchanged



Next Sentence Prediction

Figure 11.7