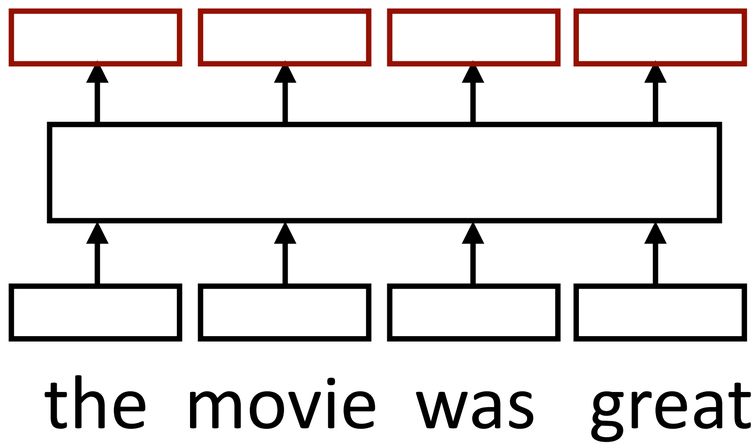
What do Transformers produce?

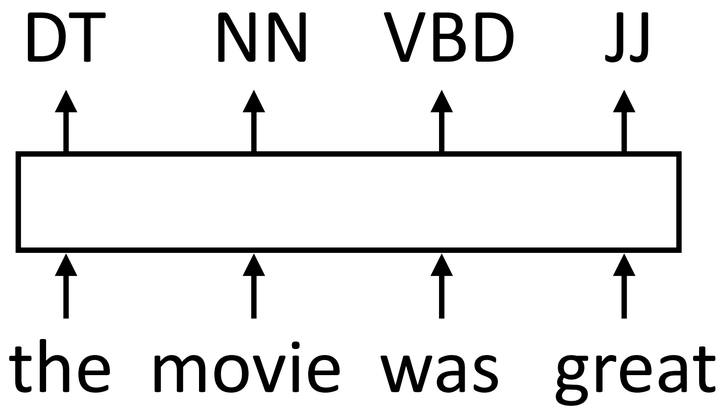


- Encoding of each word — can pass this to another layer to make a prediction (like predicting the next word for language modeling)

- Like RNNs, Transformers can be viewed as a transformation of a sequence of vectors into a sequence of context-dependent vectors

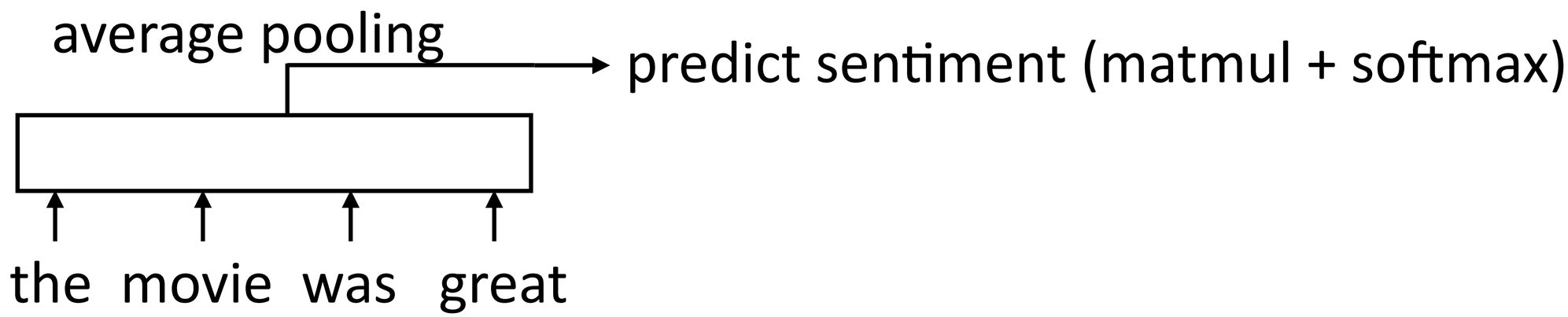
Transformer Uses

- Transducer: make some prediction for each element in a sequence

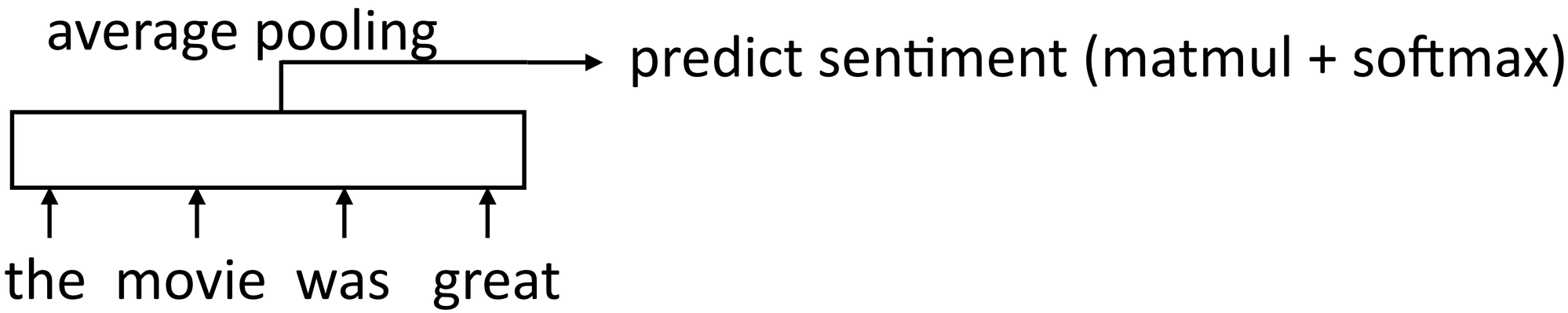


output **y** = score for each tag, then softmax

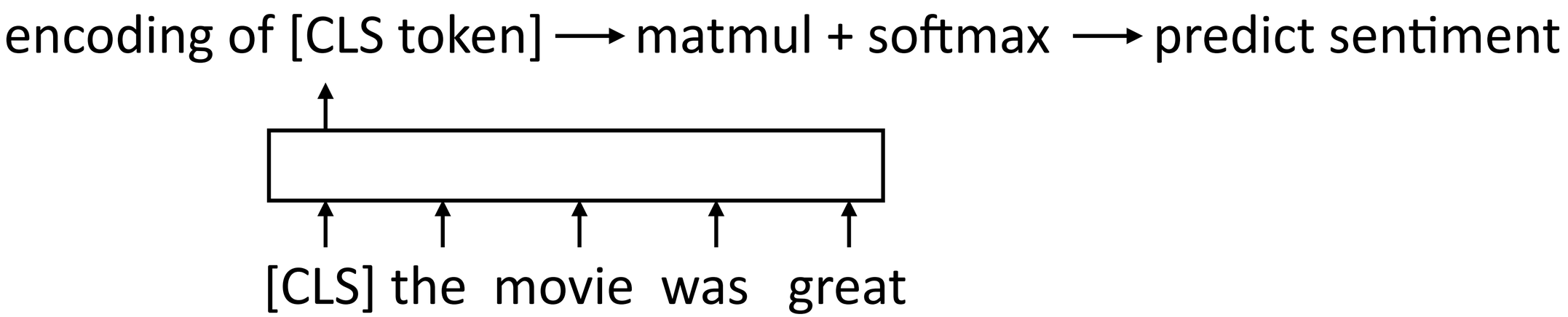
- Classifier: encode a sequence into a fixed-sized vector and classify that



Transformer Uses



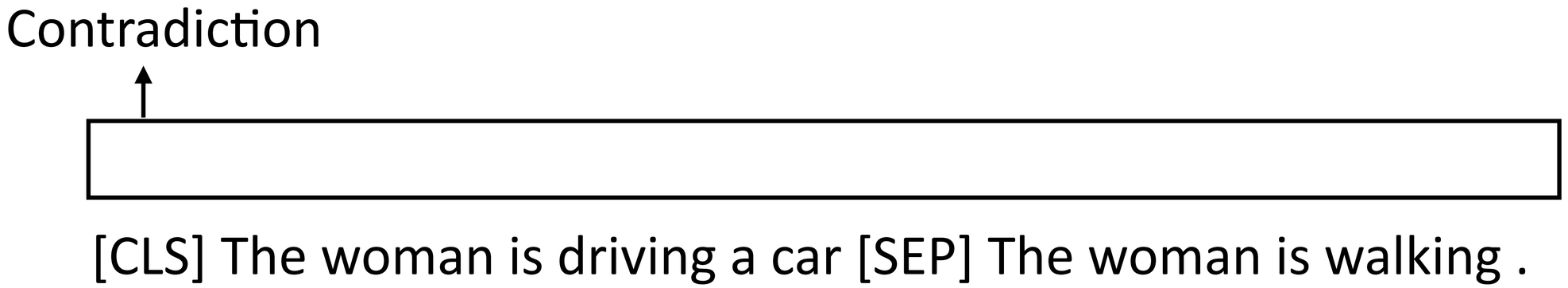
- Alternative: use a placeholder [CLS] token at the start of the sequence.



- Because [CLS] attends to everything with self-attention, it can do the pooling for you!

Transformer Uses

- Sentence **pair** classifier: feed in two sentences and classify something about their relationship



- Transformers are particularly good at sentence **pair** tasks because they can capture alignment between words