MACHINE COMPREHENSION

Machine Comprehension

Chris Burges 2013

“A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question.”

Problem Setting

context

question → Document or a Passage → answer

What is the capital of Australia ?

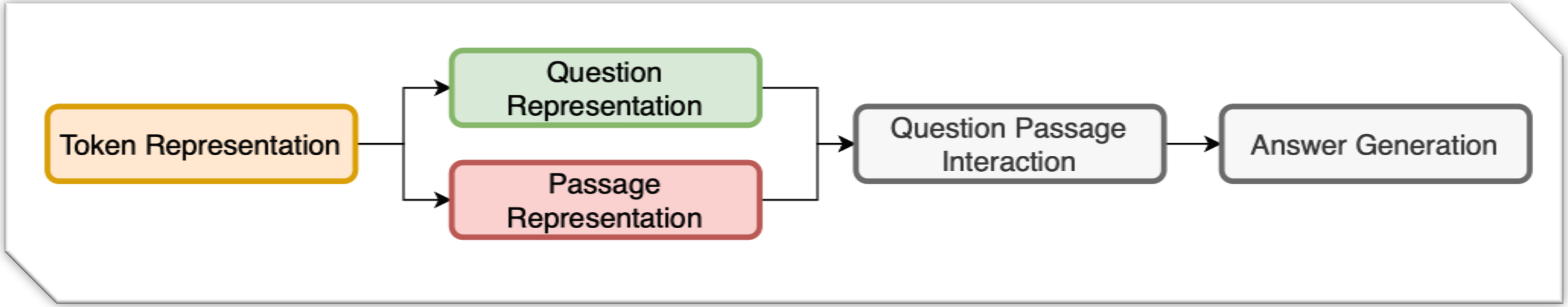
As the seat of the government of Australia, Canberra is home to many important institutions of the federal government, national monuments and museums. Canberra is also the capital of the country.

Canberra

- Extractive phrase
- Sentence

Canberra is also the capital of the country.

The MRC Pipeline



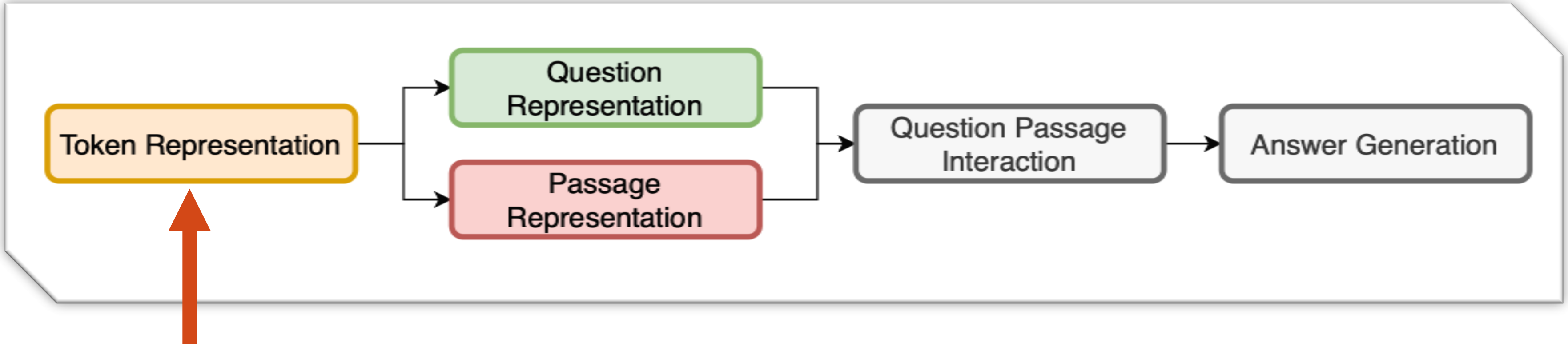
- Words, characters, sub-words embeddings
- Contextual Embeddings
- Other features – Matching, Alignment, Language structure

- Sequential representation
- Contextual representation
- Attentive reading

- Attentive reading
- Attention flows
- Multiple input passes inputs
- Re-representation of question and passages

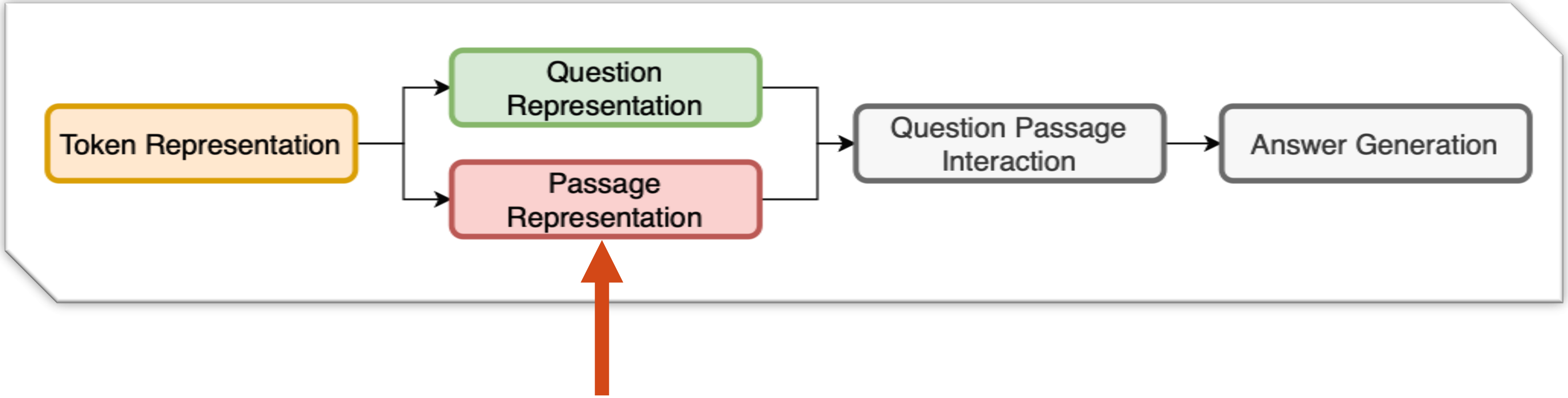
- Token prediction
- Span prediction
- Free-form generation

Token Representation



Conventional	Linguistic	Contextual
• One Hot • Word Embeddings • Sub-word, Character Embeddings	• POS • Named Entity • Query Category	• Bi-LSTM • BERT • ELMO

Question/Passage Representation



CNN

- CNN. for doc rep.
- Cross-attention
- …

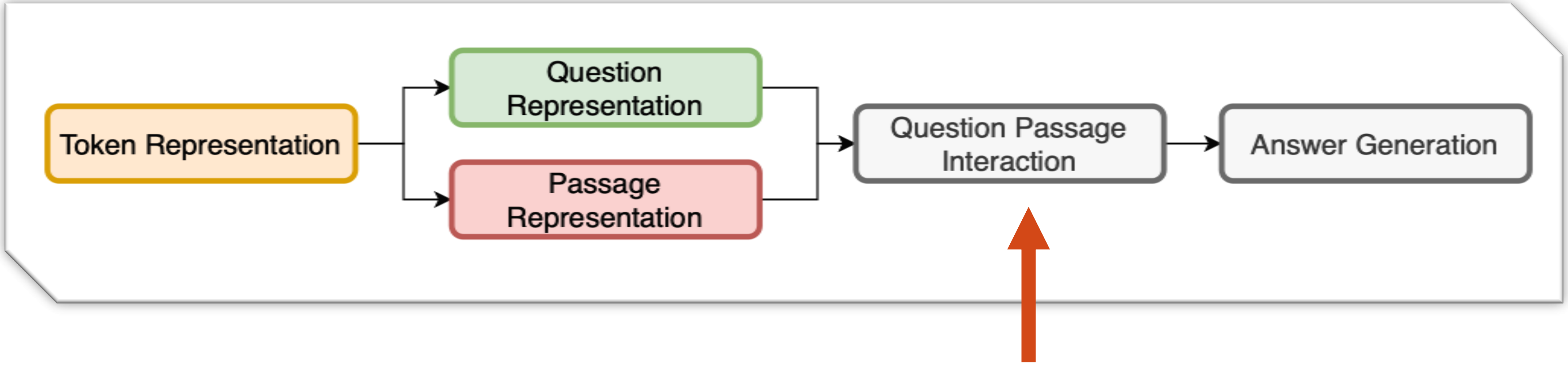
RNN

- LSTM
- Bi-LSTM
- Bi-GRU

Transformers

- GPT2
- BERT
- XLNET

Question And Passage Interaction



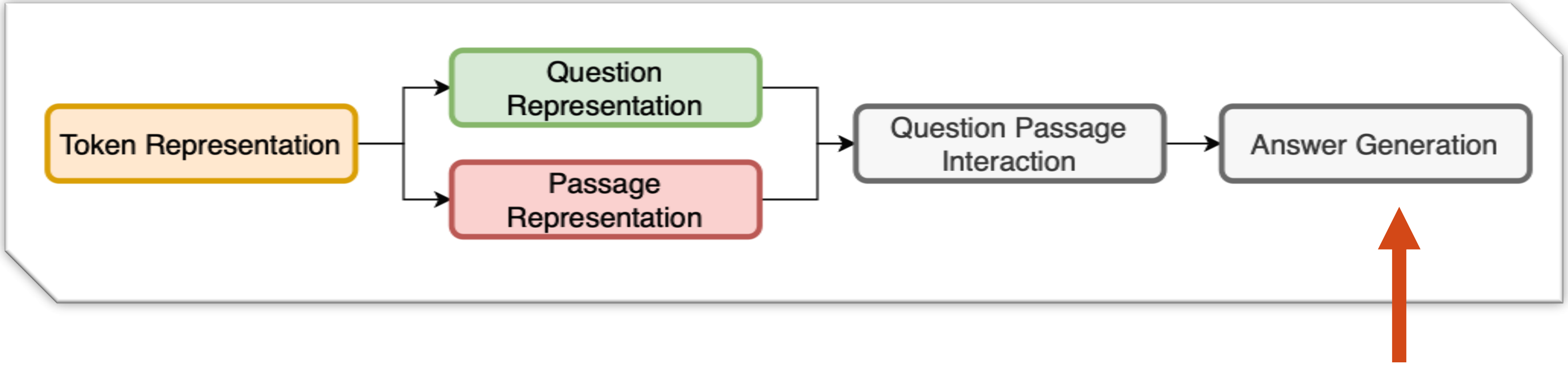
Attention

- Uni directional
- Bi-directional
- Cross-attention

Interaction Type

- Single Hop Interaction
- Multi hop Interaction

Answer Generation



Cloze	MCQ	Free Text	Span Pred.
• Word-level prediction (BC)	• Choosing one answer among many (MC)	• Generative models	• Predict begin and end of sequence

Attention Mechanism

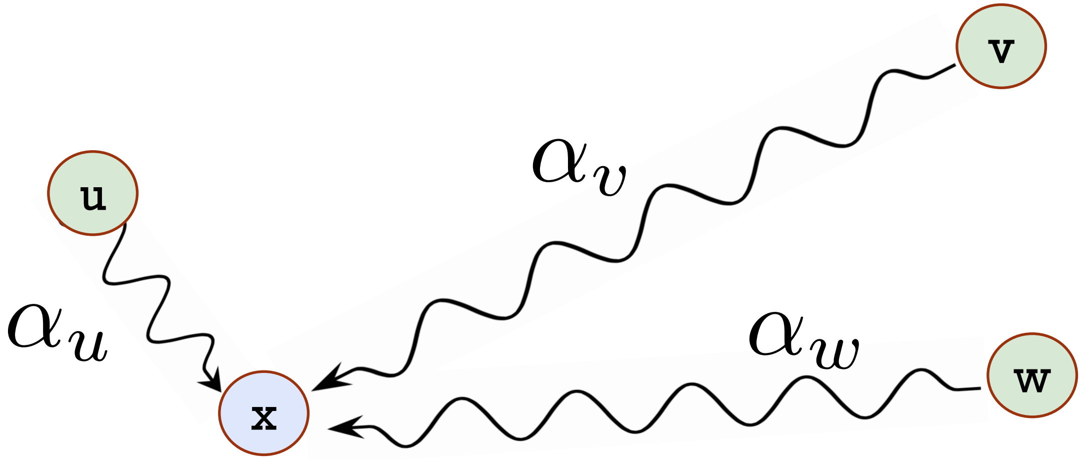
- Attention is used to represent tokens, question and passages
  - How do we re-represent otherwise independent token representations ?
  - How do we leverage contextualization ?

- Hard attention
- **Soft Attention**
- Co-attention
- Self-attention

Attention – Influence Point Of View

Attention encodes how much influence the context u has on x

$$\alpha_u = \left(\frac{e^{\mathbf{x}.\mathbf{u}}}{e^{\mathbf{x}.\mathbf{u}} + e^{\mathbf{x}.\mathbf{v}} + e^{\mathbf{x}.\mathbf{w}}} \right)$$



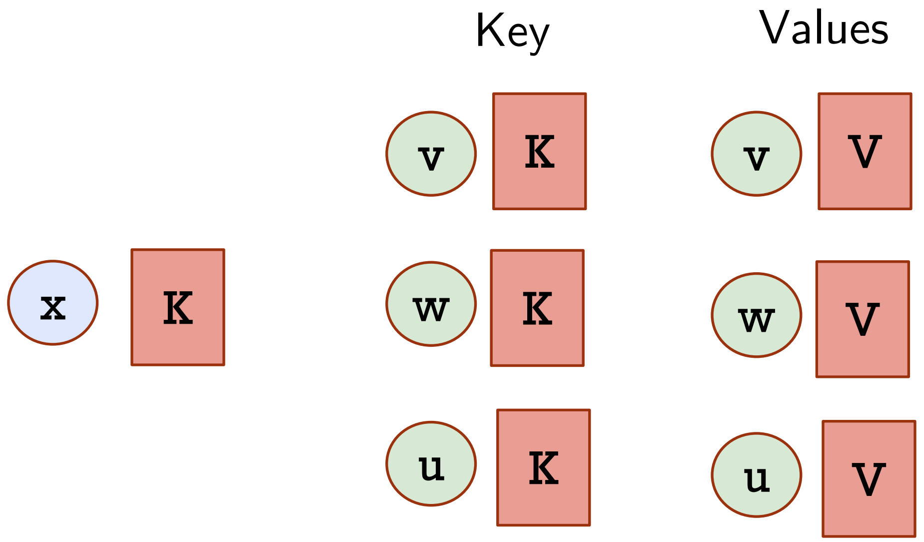
$$\mathbf{x}' = \alpha_u \mathbf{u} + \alpha_v \mathbf{v} + \alpha_w \mathbf{w}$$

- Typically x and context vectors are first projected through a learnable matrix W

Attention Mechanism – Memory Point Of View

Attention retrieves values from a continuous memory using fuzzy matching

- Assume vectors are stored in memory referenced by Key matrix K
- Thought expt: for 1-hot vectors = hashmaps
- Instead Kx retrieves from this continuous memory as a weighted sum over all values



Attention weight

$$\alpha_u = \frac{e^{K\mathbf{x}\cdot K\mathbf{u}}}{e^{K\mathbf{x}\cdot K\mathbf{u}} + e^{K\mathbf{x}\cdot K\mathbf{v}} + e^{K\mathbf{x}\cdot K\mathbf{w}}}$$

$$\mathbf{x}' = \alpha_u \mathbf{u} + \alpha_v \mathbf{v} + \alpha_w \mathbf{w}$$

Attentive Reader

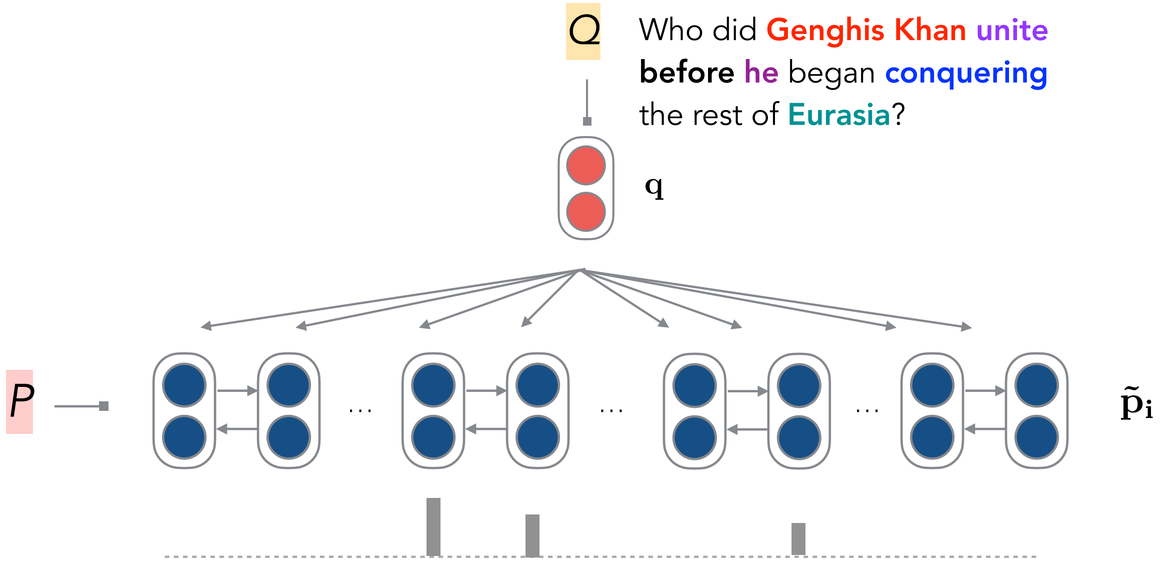
[Chen '16] Attentive reader

Tokens
- Word embeddings
- Exact matching features
- POS,NEF, TF features
- Alignment

Passage
- Bi-LSTM

Query-Passage Interaction
- Soft Attention

Answer Gen.
- Span Pred.

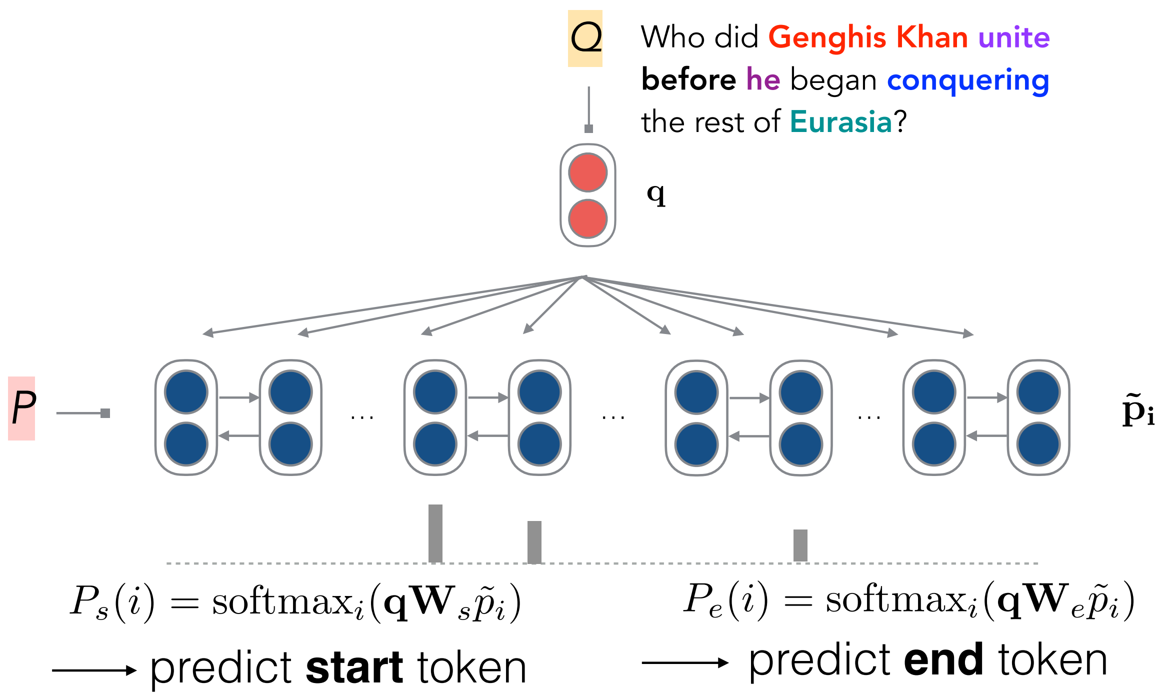


$$P_s(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$$
→ predict **start** token

$$P_e(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$$
→ predict **end** token

Attentive Reader

[Chen '16]

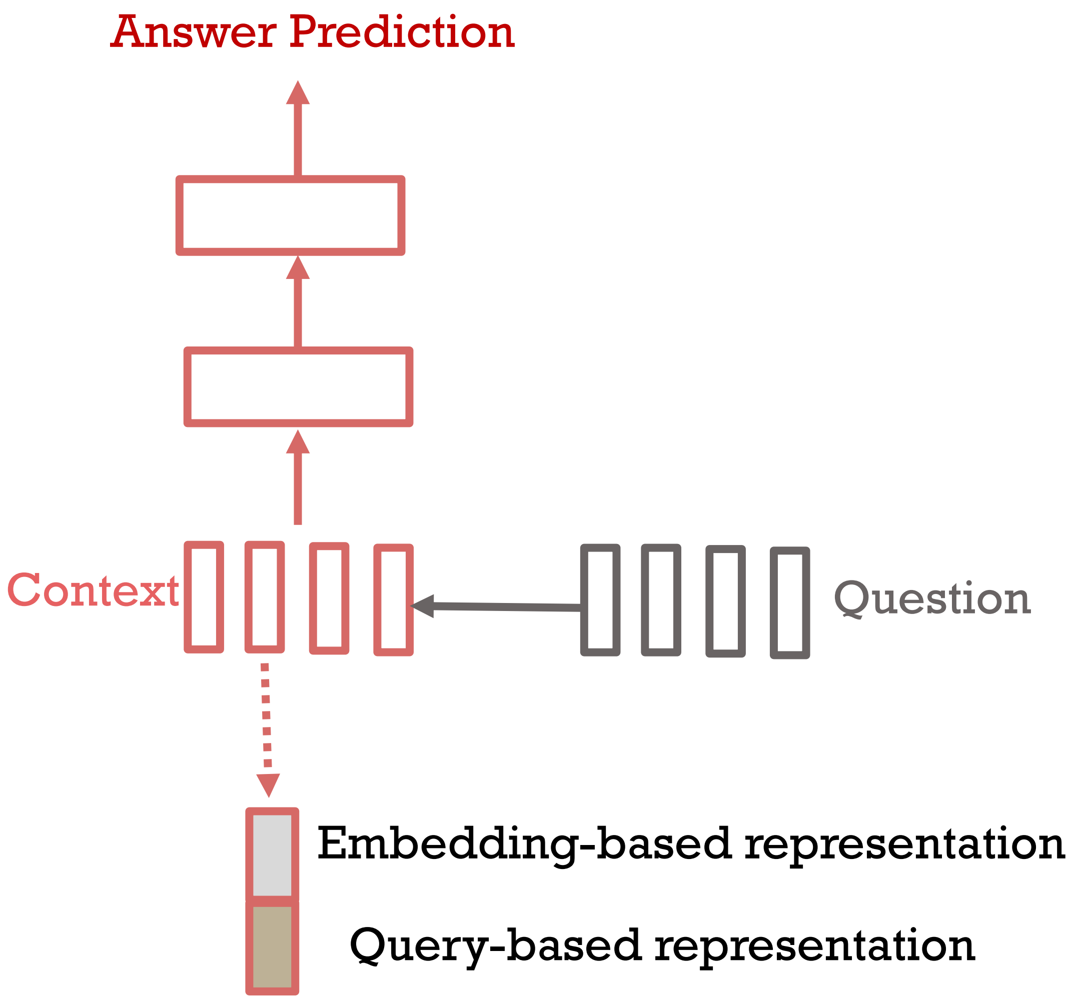


$$\Pr(a|q, p_i) = P_s(a_s)P_e(a_e)$$

Input Representation

DrQA [Chen '17]

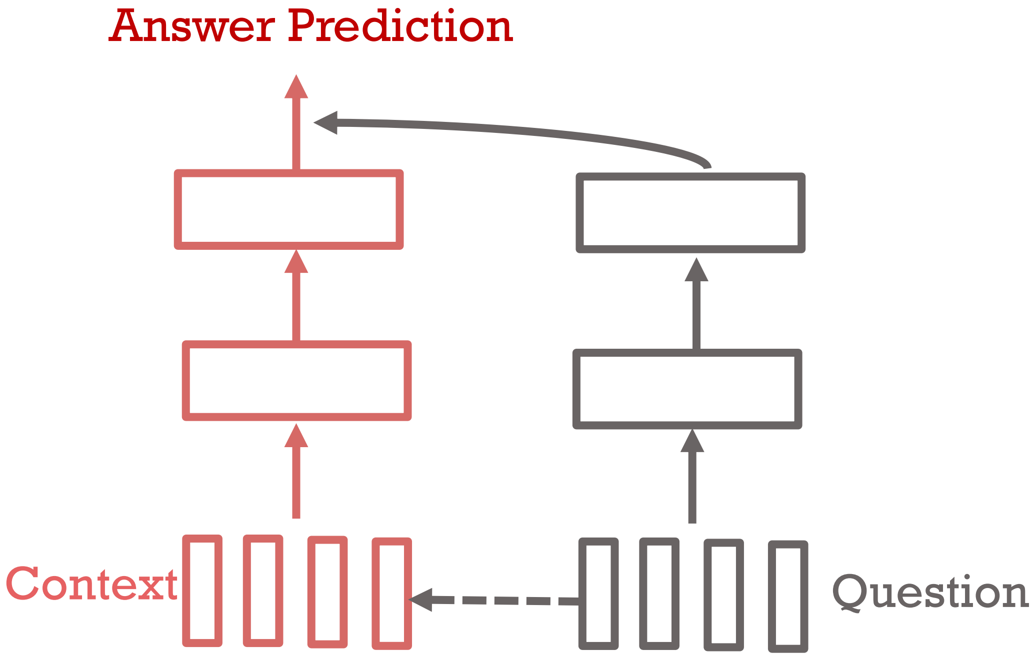
- Context words are represented based on similarity with the query
- Semantic similarity
  - Word embeddings
- Matching similarity
  - Direct word-level matching
  - Weighted matching
  - Attention mechanism



Late Interaction

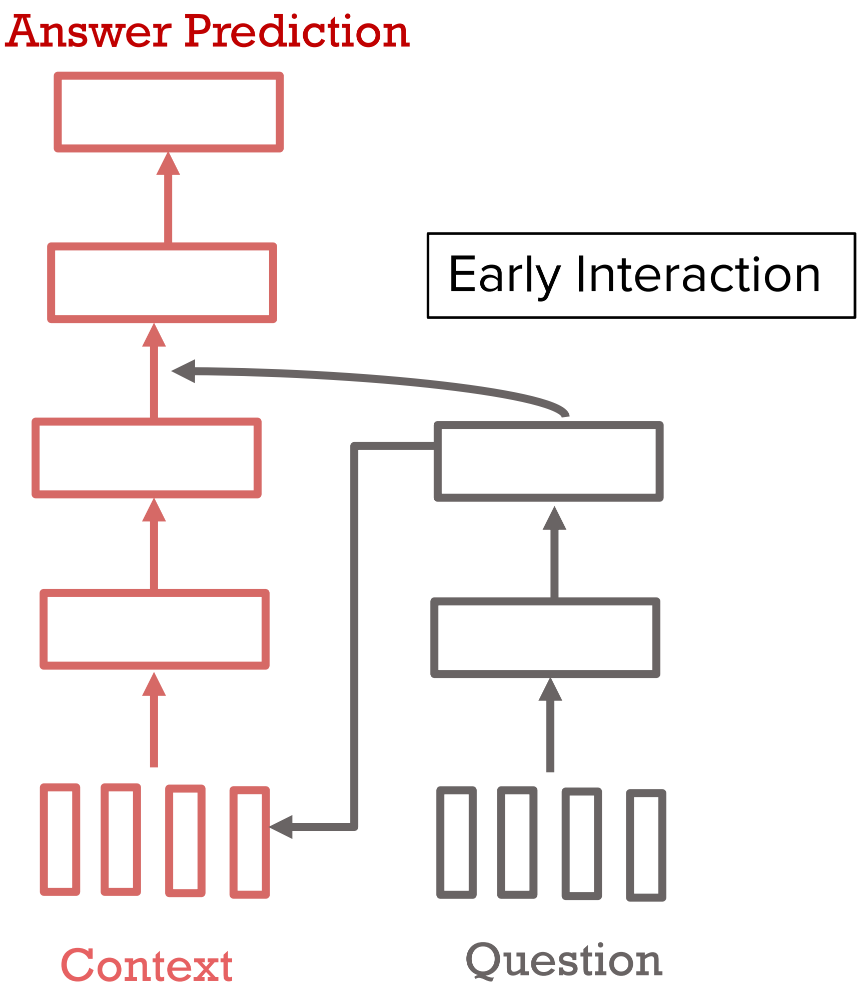
MatchLSTM [Wang & Jiang '16], DCN [Xiong '16], BiDAF [Seo '17]

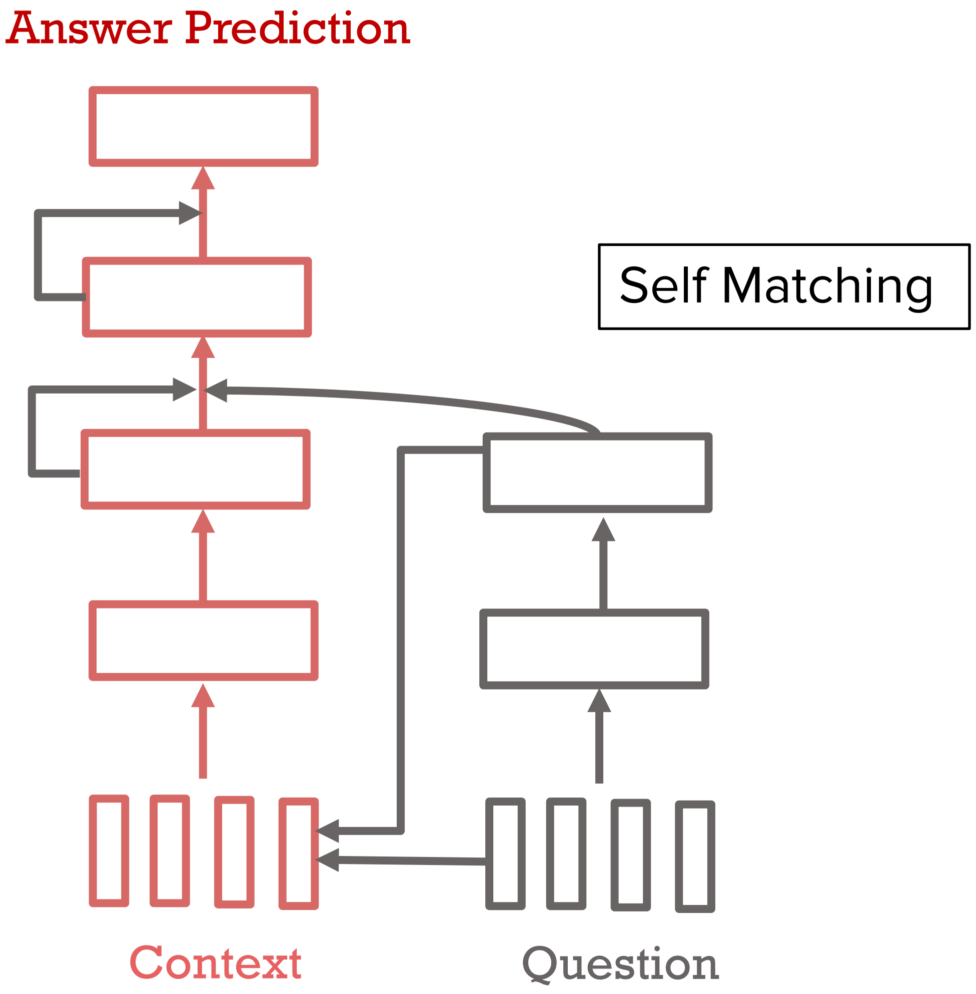
- First encode question and context sufficiently
- Choice of encoders
  - Bi-LSTMs
  - Conv Nets
- Most popular Model
  - Bi-directional attention flow [Seo '17]



Rasor [Lee '16], fastQA [Weissenborn'17]

Other Variants

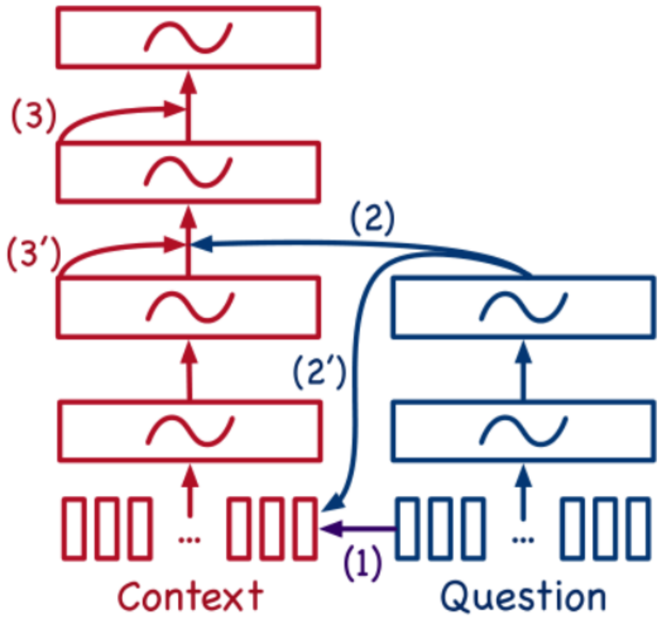




Attention Based Architectures

- 2016 – 2017 – Multitude of attention based architecture

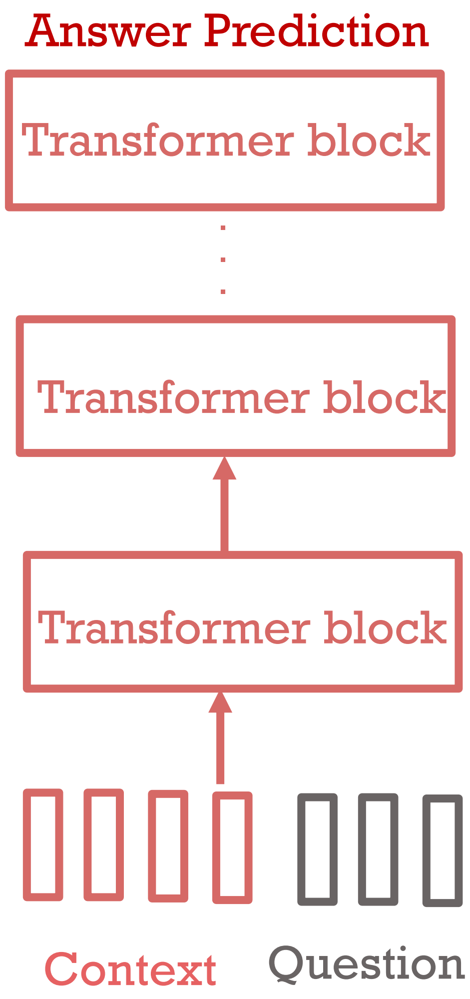
Architectures	**(1)**	**(2)**	**(2')**	**(3)**	**(3')**
Match-LSTM (Wang and Jiang, 2016)		✓			
DCN (Xiong et al., 2017)		✓			✓
FastQA (Weissenborn et al., 2017)	✓				
FastQAExt (Weissenborn et al., 2017)	✓	✓		✓	
BiDAF (Seo et al., 2017)		✓			✓
RaSoR (Lee et al., 2016)	✓		✓		
DrQA (Chen et al., 2017)	✓				
MPCM (Wang et al., 2016)	✓	✓			
Mnemonic Reader (Hu et al., 2017)	✓	✓		✓	
R-net (Wang et al., 2017b)		✓		✓	



(1) Word-level fusion, (2) high-level fusion, (2') high-level fusion (alternative), (3) self-boosted fusion, and (3') self-boosted fusion (alternative).

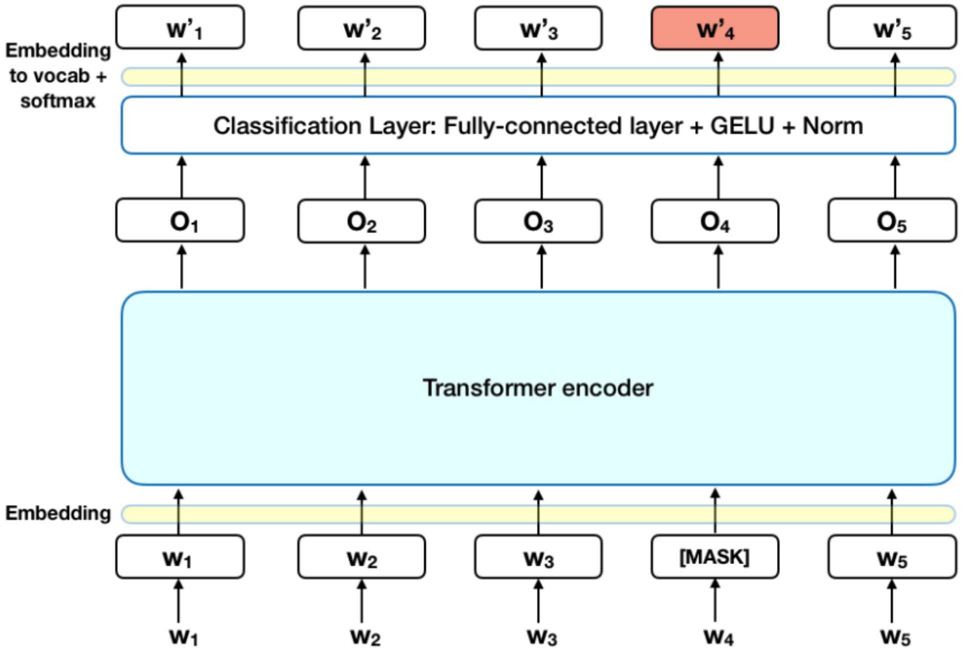
Contextual Language Models

- BERT – No Recurrence, only attention
- Re-representing each token based on the context
- Shows the most promising performance



BERT

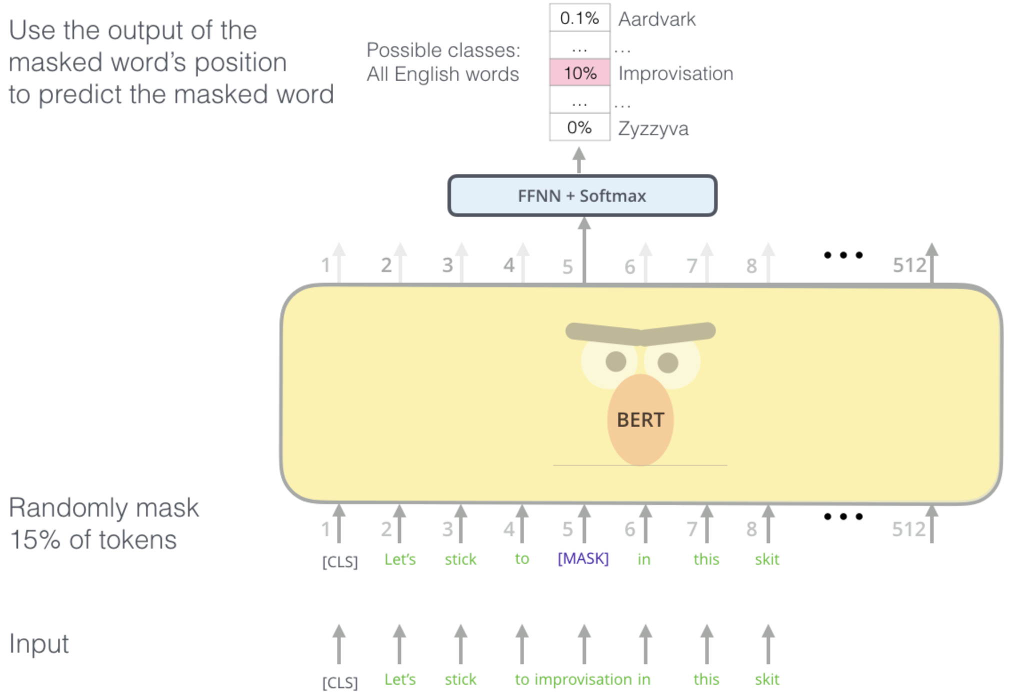
- Bi-directional : Transformer encoder reads the entire sequence of words at once.
  - Learns the context of a word based on all of its surroundings (left and right of the word).



BERT– Masked Language Model

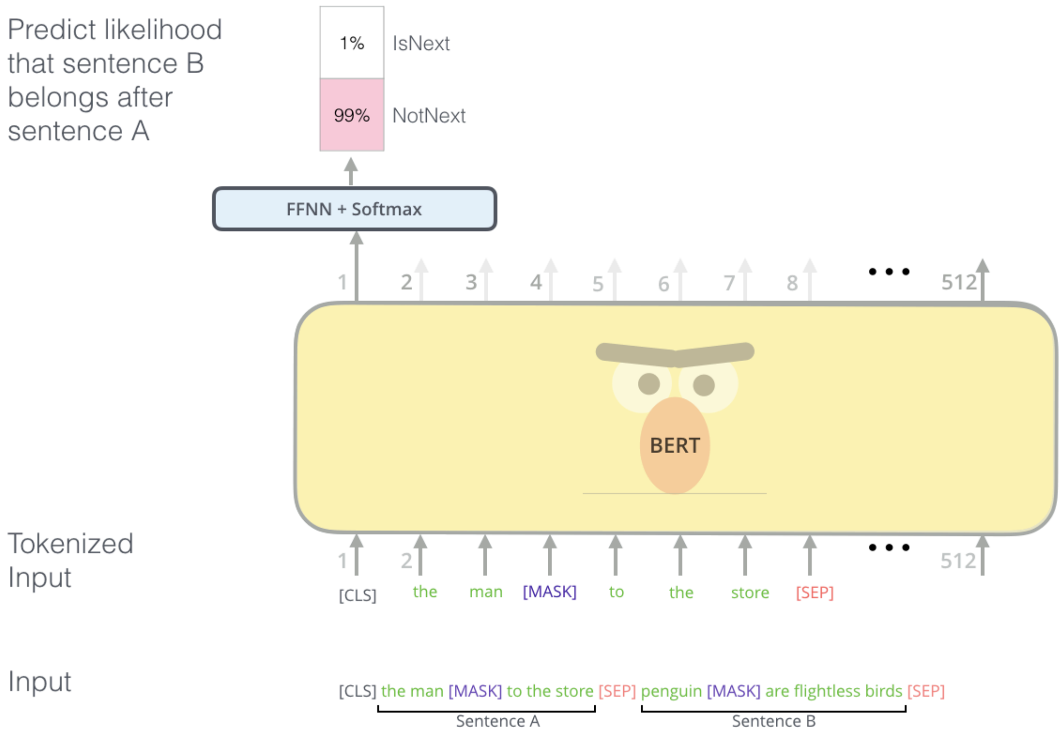
Masked word prediction

- Given a sentence with some words masked at random, can we predict them?
- Randomly select 15% of tokens to be replaced with “<MASK>”



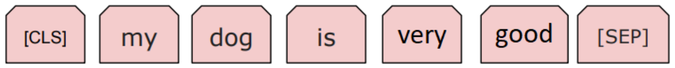
Next Sentence Prediction

- Given two sentences, does the first follow the second? Teaches BERT about relationship between two sentences
- 50% of the time the actual next sentence, 50% random

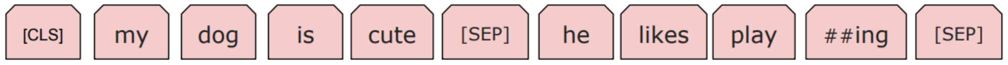


BERT Fine Tuning

- Inputs to BERT – [CLS] <token embeddings> [SEP] …

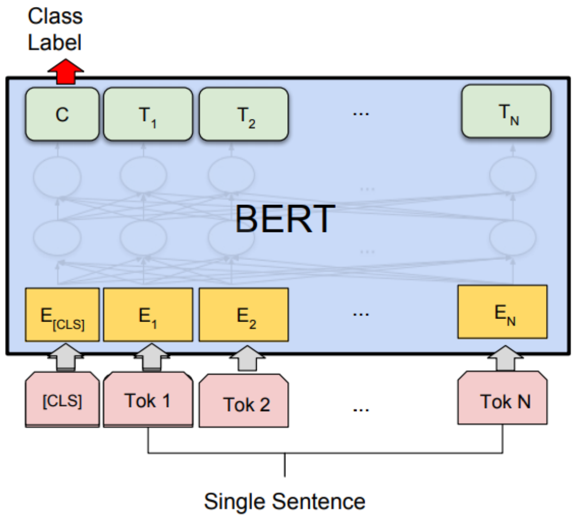


Single sentence input



Single sentence input

- Classification tasks such as sentiment analysis are done similarly to Next Sentence classification, by adding a classification layer on top of the Transformer output for the [CLS] token.



BERT Fine Tuning

Q&A model can be trained by learning two extra vectors that mark the beginning and the end of the answer.