Question Answering

Ph.D candidate in Computational Science and Engineering
Yonsei Univ.

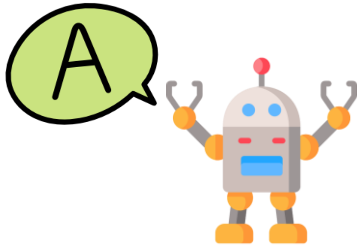
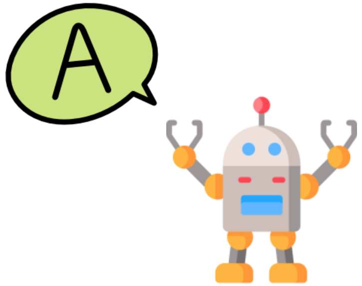
Jin-Duk Park

Reading group material

What is Question Answering (QA)?

	Question	Answer	
	What is 1+1 =?	2	
	Where is the capital of Korea?	Seoul	

What is Question Answering (QA)? (Cont'd)

	Question	Answer	
	What is 1+1 =?	2	
	Where is the capital of Korea?	Seoul	

Question Answering (QA) is a task that **answering a query** about a given context paragraph

What is Question Answering (QA)? (Cont'd)

Question Answering (QA) is a task that **answering a query** about a **given context paragraph**

Context paragraph

Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, … the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College…

Q:
What individual is the school named after?

A:
John Havard

SQuaD dataset

SQuAD2.0

The Stanford Question Answering Dataset

- **Open benchmark dataset (most widely used)**
- **Collected via crowdworkers**

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

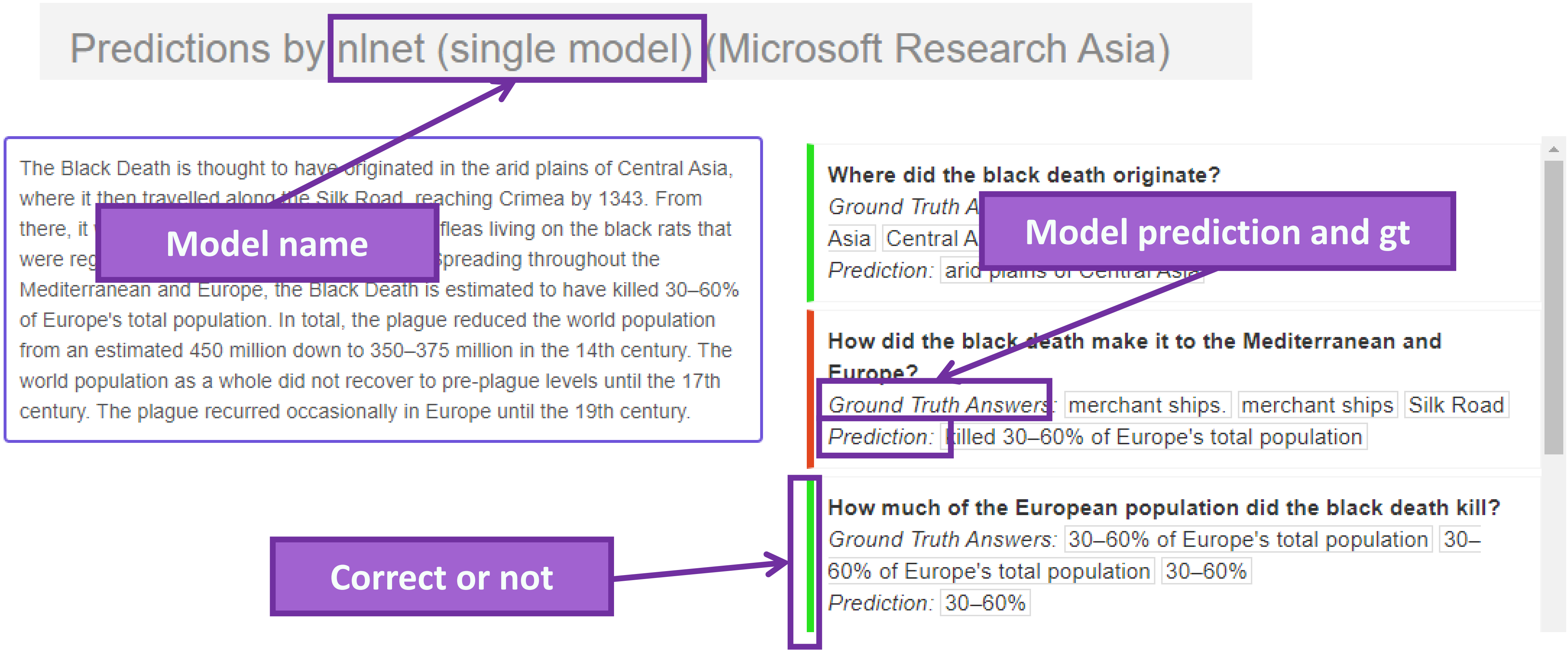
Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

https://rajpurkar.github.io/SQuAD-explorer/

SQuaD dataset (Cont'd)

[Web] *-> https://rajpurkar.github.io/SQuAD-explorer/*



Evaluation Metrics in QA

1. EM (Exact Match)

- A strict **all-or-nothing metric**
- If (model prediction) = (true answer), EM = 1 otherwise 0
- E.g.)
  >> Correct answer: Amazonia or the Amazon Jungle
  >> Prediction 1: Amazonia or the Amazon Jungle **-> EM = 1**
  >> Prediction 2: Amazonia **-> EM = 0**

2. F1 score

- TP, FP, FN are counted for **each token**

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

- E.g.)
  >> Correct answer: Amazonia or the Amazon Jungle
  >> Prediction: Amazonia or the Amazon Basin
  -> True positive(Amazonia or the Amazon)/False Positive:1(Basin)/False Negative:1 (Jungle)
  -> Precision = 0.8, Recall = 0.8

What is Question Answering (QA)? (Cont'd)

Question Answering (QA) is a task that **answering a query** about a **given context paragraph**

Context paragraph

Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, … the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College...

Q:
What individual is the school named after?

A:
John Havard

What is Question Answering (QA)? (Cont'd)

Question Answering (QA) is a task that **answering a query** about a **given context paragraph**

Context paragraph

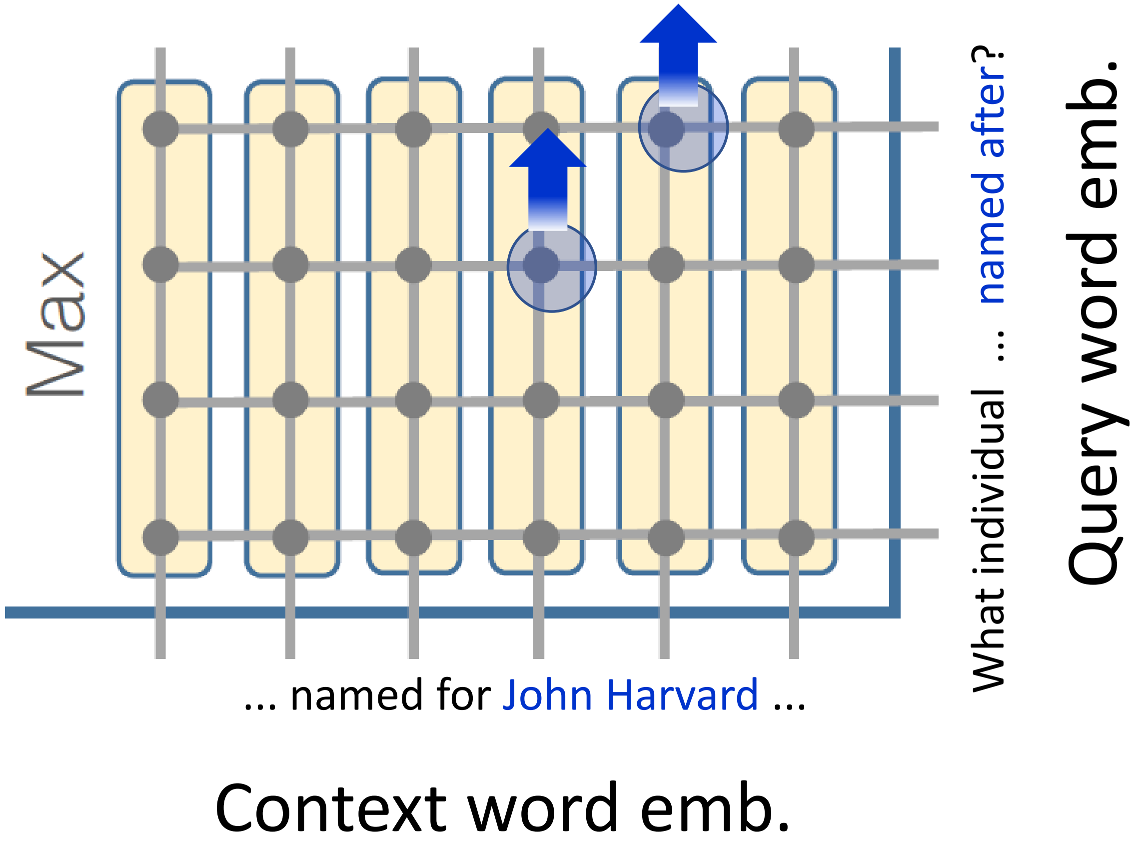
Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, … the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College

Attention is important for QA task

Q:
What individual is the school named after?

A:
John Havard

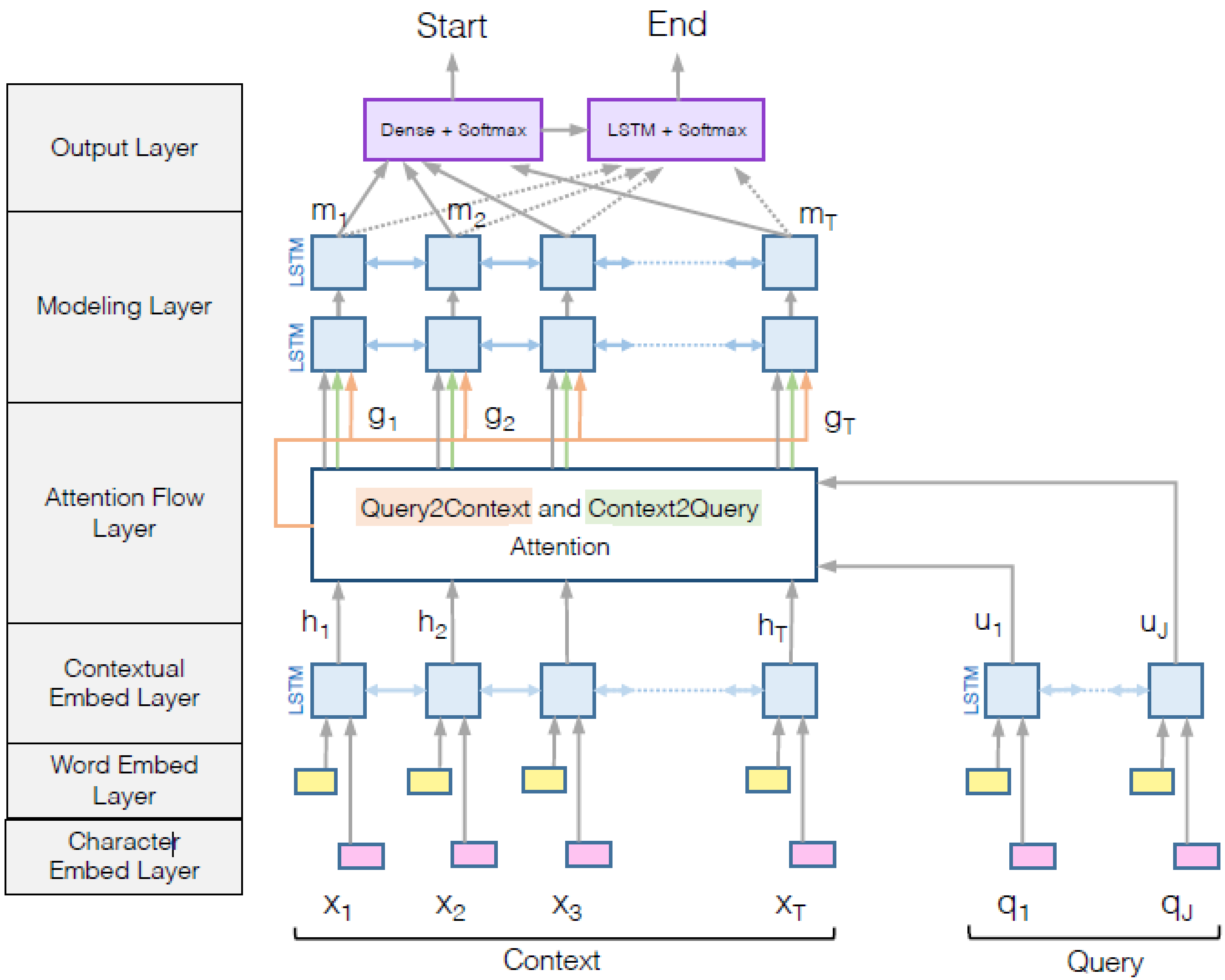
Attention Between Query and Context



- **Attention-based model: natural design choice**
- **Attention between query and context**

  Words in the context that have **high attention:** highly probable to be an **answer**

Overview of BIDAF

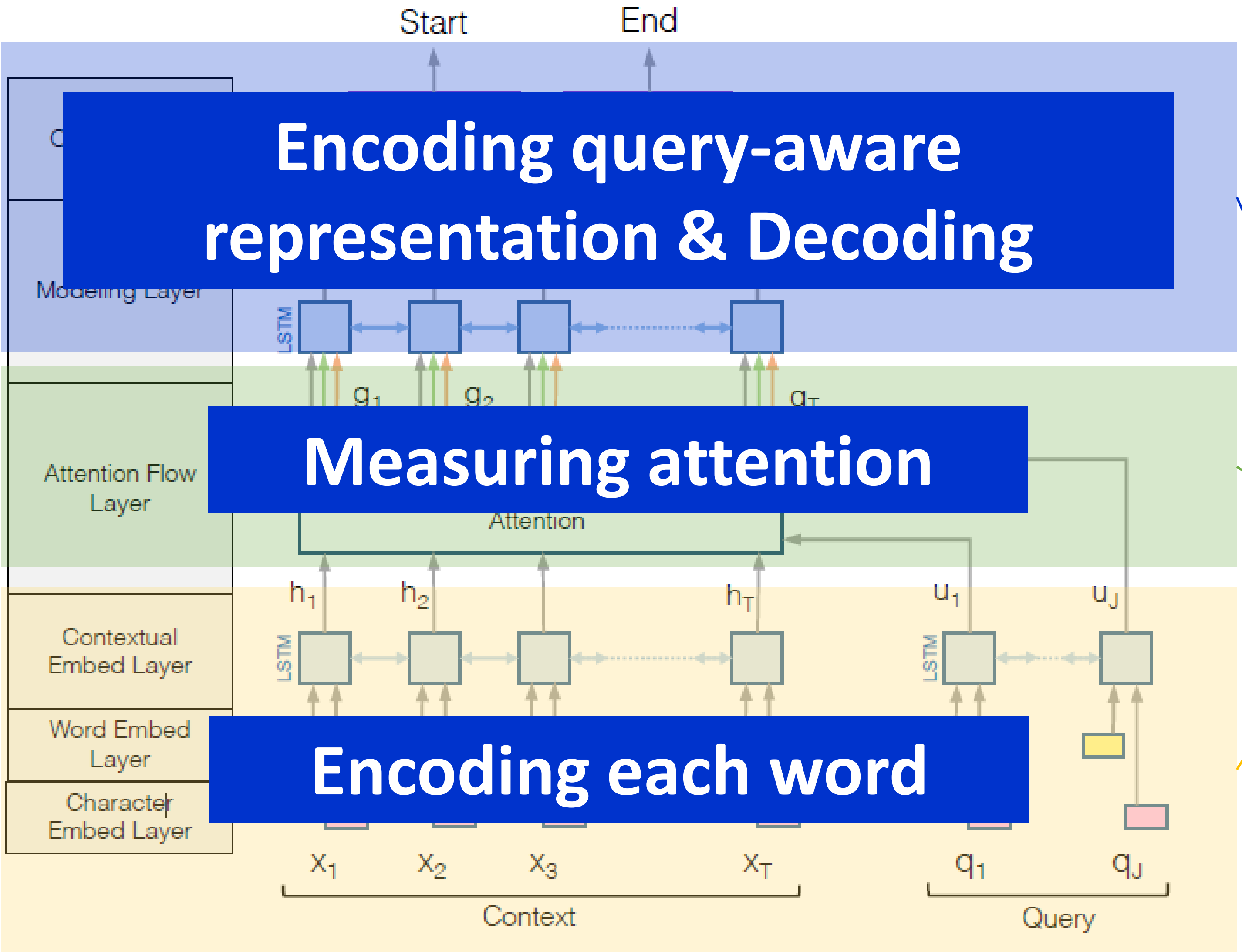


- **Model architecture of BIDAF** (bidirectional attention flow model)

- **Consists of 6 steps of layers**
  - Character embedding layer
  - Word embedding layer
  - Contextual embedding layer
  - Attention flow layer
  - Modeling layer
  - Output layer

Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." ICLR 2017

Overview of BIDAF



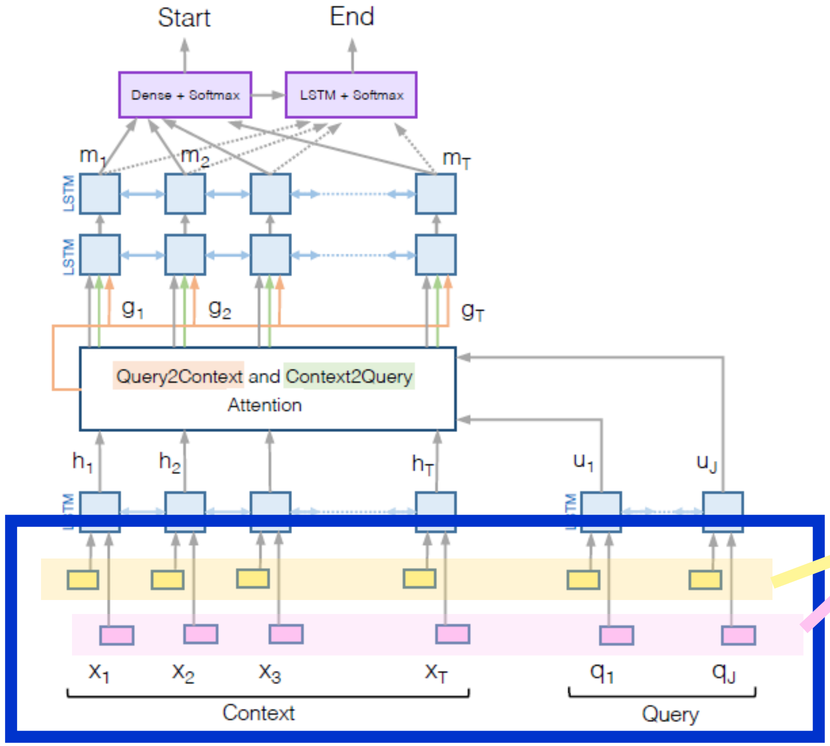
- **Model architecture of BIDAF** (bidirectional attention flow model)
- **Consists of 6 steps of layers**
  - Character embedding layer
  - Word embedding layer
  - Contextual embedding layer
  - Attention flow layer
  - Modeling layer
  - Output layer

(lets take a look one-by-one)

- ▶ Character embedding layer
- ▶ Word embedding layer
- Contextual embedding layer
- Attention flow layer
- Modeling layer
- Output layer

BIDAF (Seo et al., ICLR 2017)

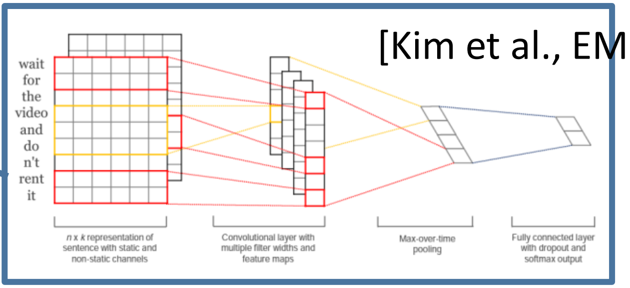
Architecture of BIDAF



- **Character & Word embedding**

Character Embedding

Char-CNN



[Kim et al., EMNLP 2014]

Word Embedding

GLOVE

X_{ik}: # Occurrence of word i in the context k

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

[Pennington et al., EMNLP 2014]

- Maps each word to a vector space via **CNN** and **linear regressor**, respectively

- **Pretrained static** word vector

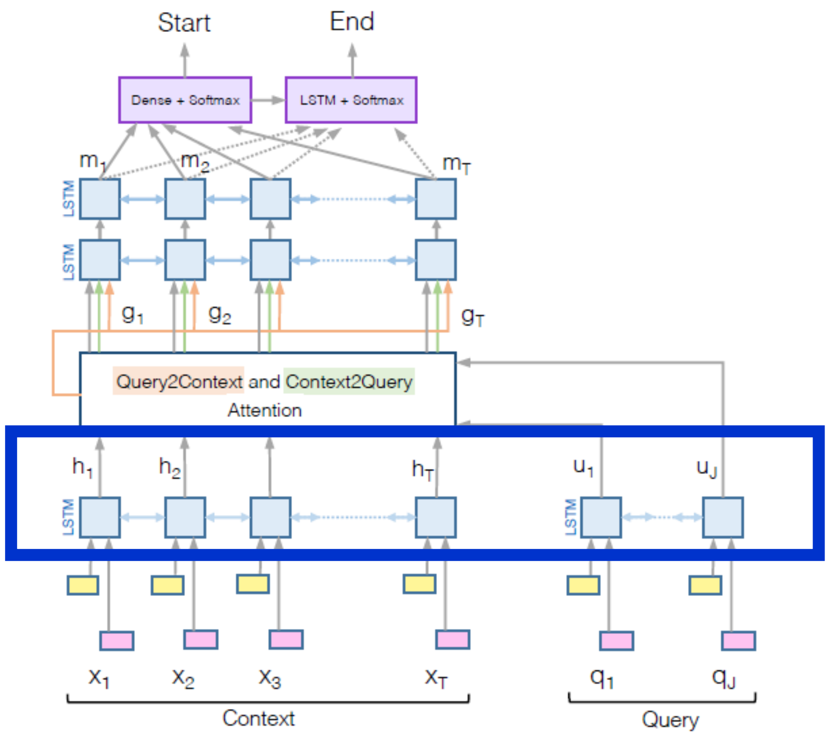
Yoon Kim. Convolutional neural networks for sentence classification. In EMNLP, 2014.

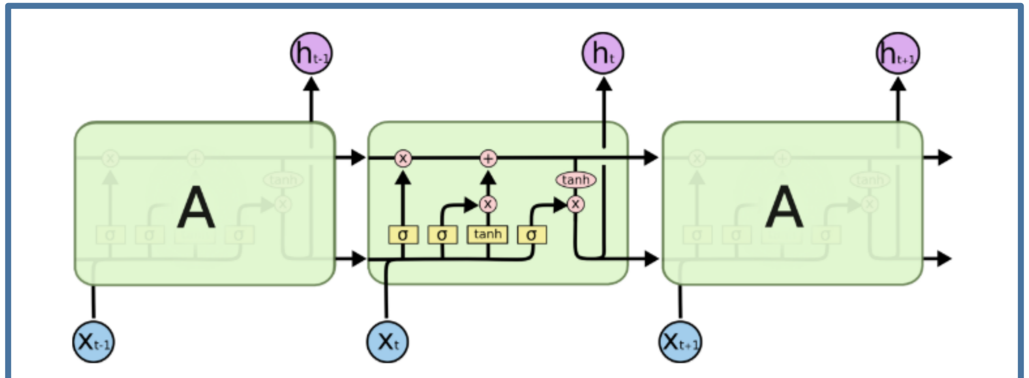
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

BIDAF (Seo et al., ICLR 2017)

Architecture of BIDAF

- **Contextual embedding layer**
  - LSTM is used for temporal interactions between words
  - LSTM is placed in both directions (2 LSTM), and outputs are concatenated





The repeating module in an LSTM contains four interacting layers.

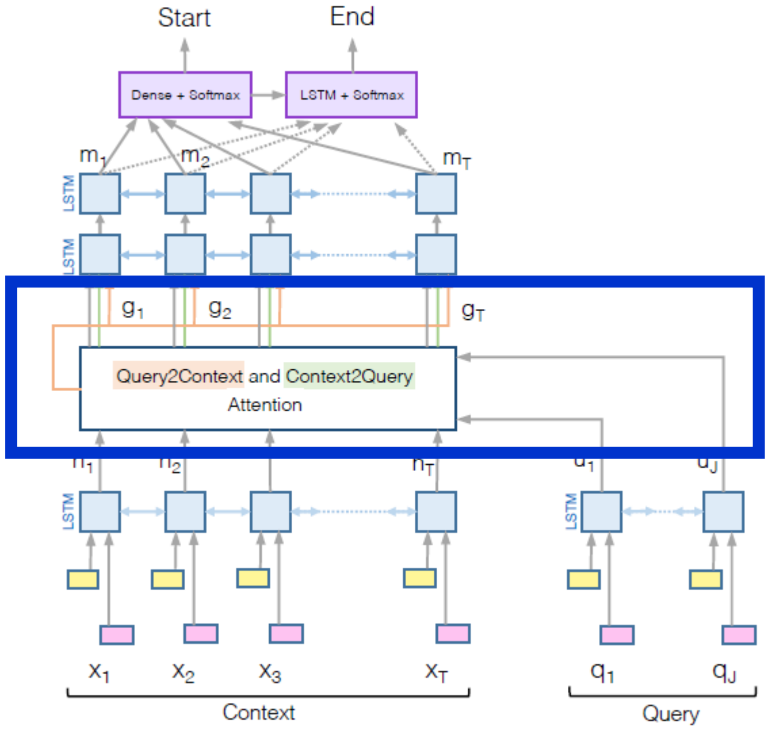
https://hackernoon.com/understanding-architecture-of-lstm-cell-from-scratch-with-code-8da40f0b71f4

* Architecture of LSTM

BIDAF (Seo et al., ICLR 2017)

Architecture of BIDAF

- Character embedding layer
- Word embedding layer
- Contextual embedding layer
- ▶ Attention flow layer
- Modeling layer
- Output layer

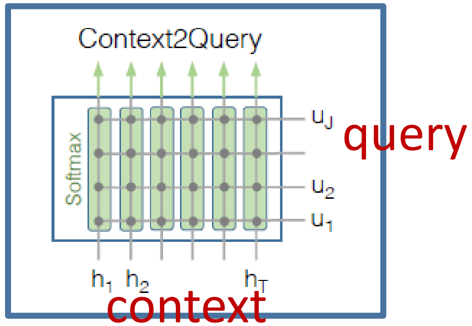


- **Similarity matrix**

$$\boxed{\mathbf{S}_{tj} = \alpha(\mathbf{H}_{:t}, \mathbf{U}_{:j}) \in \mathbb{R}}$$

α: trainable function

- **Context2Query**

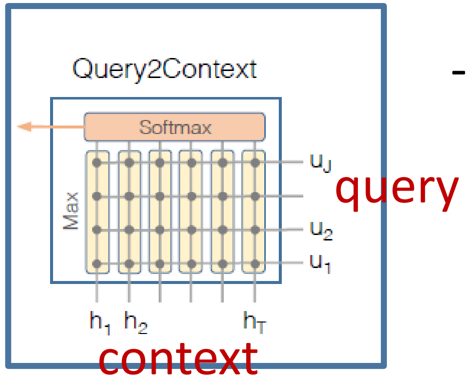


- Signifies which **query words** are most relevant to each **context word**

$$\mathbf{a}_t = \mathrm{softmax}(\mathbf{S}_{t:})$$

$$\tilde{\mathbf{U}}_{:t} = \sum_j \mathbf{a}_{tj} \mathbf{U}_{:j}$$

- **Query2Context**



- Signifies which **context words** are most relevant to each **query word**

$$\mathbf{b} = \mathrm{softmax}(\max_{col}(\mathbf{S}))$$

$$\tilde{\mathbf{h}} = \sum_t \mathbf{b}_t \mathbf{H}_{:t}$$

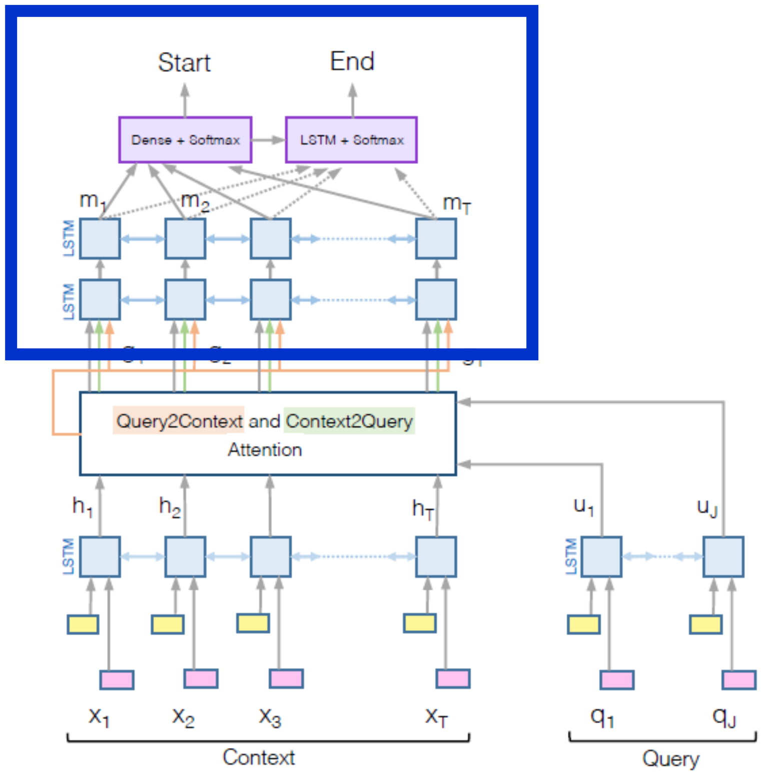
- **Query-aware representation**

$$\boxed{\mathbf{G}_{:t} = \boldsymbol{\beta}(\mathbf{H}_{:t}, \tilde{\mathbf{U}}_{:t}, \tilde{\mathbf{H}}_{:t}) \in \mathbb{R}^{d_{\mathbf{G}}}}$$

β: trainable function (neural network)

Architecture of BIDAF

- Character embedding layer
- Word embedding layer
- Contextual embedding layer
- Attention flow layer
- ▶ Modeling layer
- ▶ Output layer



- **Modeling and output Layer**
  - Bi-directional LSTM is again used for modeling layer, to predict $\boldsymbol{M}$
  - Output layer: to find the **start & end indices**

Start index: $\mathrm{p}^1 = \mathrm{softmax}(\mathbf{w}_{(\mathbf{p}^1)}^\top [\mathbf{G}; \mathbf{M}]),$

End index: $\mathrm{p}^2 = \mathrm{softmax}(\mathbf{w}_{(\mathbf{p}^2)}^\top [\mathbf{G}; \mathbf{M}^2])$

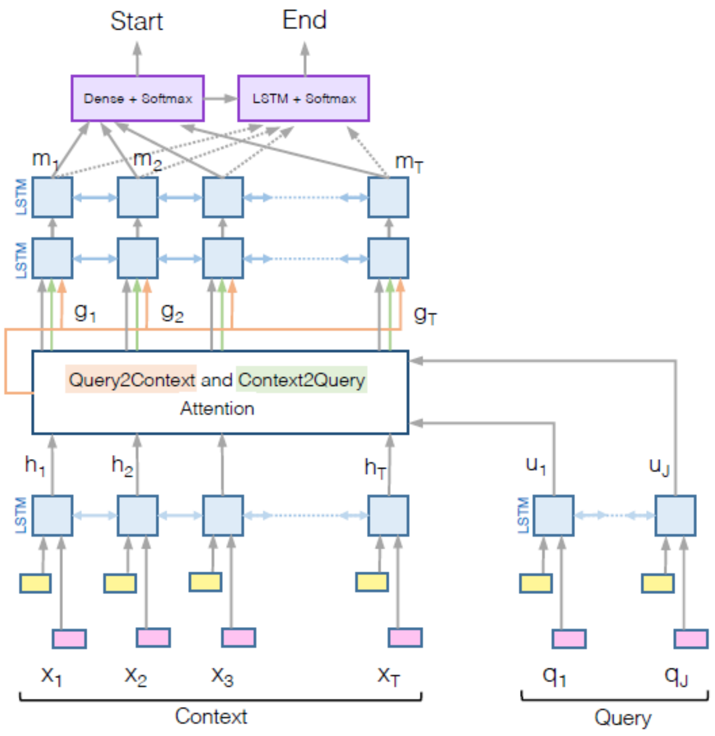
- **Training**
  - Training with the ground-truth index y

$$\boxed{L(\theta) = -\frac{1}{N}\sum_i^N \log(\mathrm{p}_{y_i^1}^1) + \log(\mathrm{p}_{y_i^2}^2)}$$

➡ $\boldsymbol{p}_*$:*-th value of $\boldsymbol{p}$

Evaluation of BIDAF

Ablation study shows contribution of each module



	EM	F1	F1
No char embedding	65.0	75.4	-1.9
No word embedding	55.5	66.8	-10.5
No C2Q attention	57.2	67.7	-9.6
No Q2C attention	63.6	73.7	-3.6
Dynamic attention	63.5	73.6	-3.7
BIDAF (single)	67.7	77.3	
BIDAF (ensemble)	72.6	80.7	

(b) Ablations on the SQuAD dev set

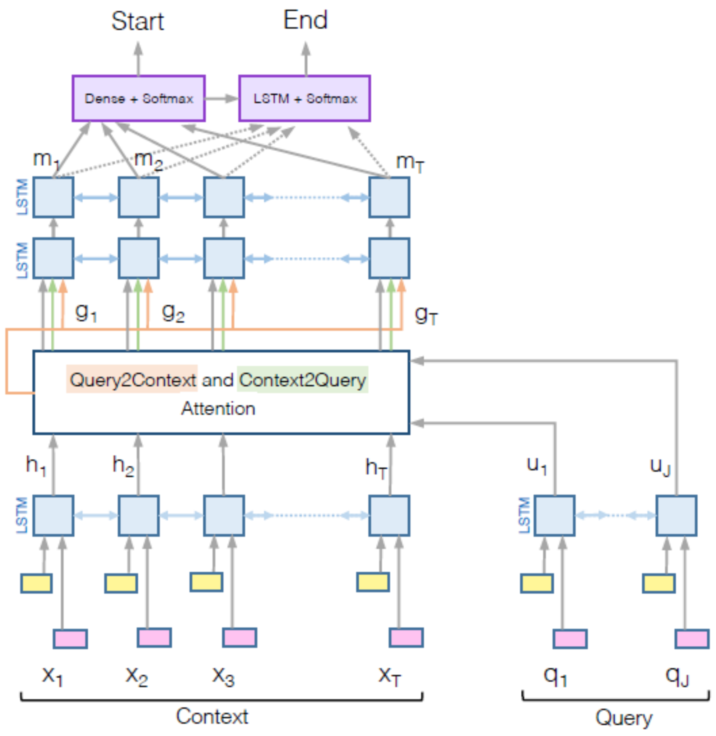
* Ensemble: An identical architecture is utilized

Contribution ranking of modules

Word emb > C2Q att > Modeling layer > Q2C att > Char emb

Evaluation of BIDAF

Ablation study shows contribution of each module



	EM	F1	F1
No char embedding	65.0	75.4	-1.9
No word embedding	55.5	66.8	-10.5
No C2Q attention	57.2	67.7	-9.6
No Q2C attention	63.6	73.7	-3.6
Dynamic attention	63.5	73.6	-3.7
BIDAF (single)	67.7	77.3	
BIDAF (ensemble)	72.6	80.7	

(b) Ablations on the SQuAD dev set

Q. Do we really need this parts?
1. Word emb w/ higher dimension would be more helpful
2. Using Q2C may enough

* Ensemble: An identical architecture is utilized

Contribution ranking of modules

Word emb > C2Q att > Modeling layer > Q2C att > Char emb

Evaluation of BIDAF

The results on SQuAD dataset

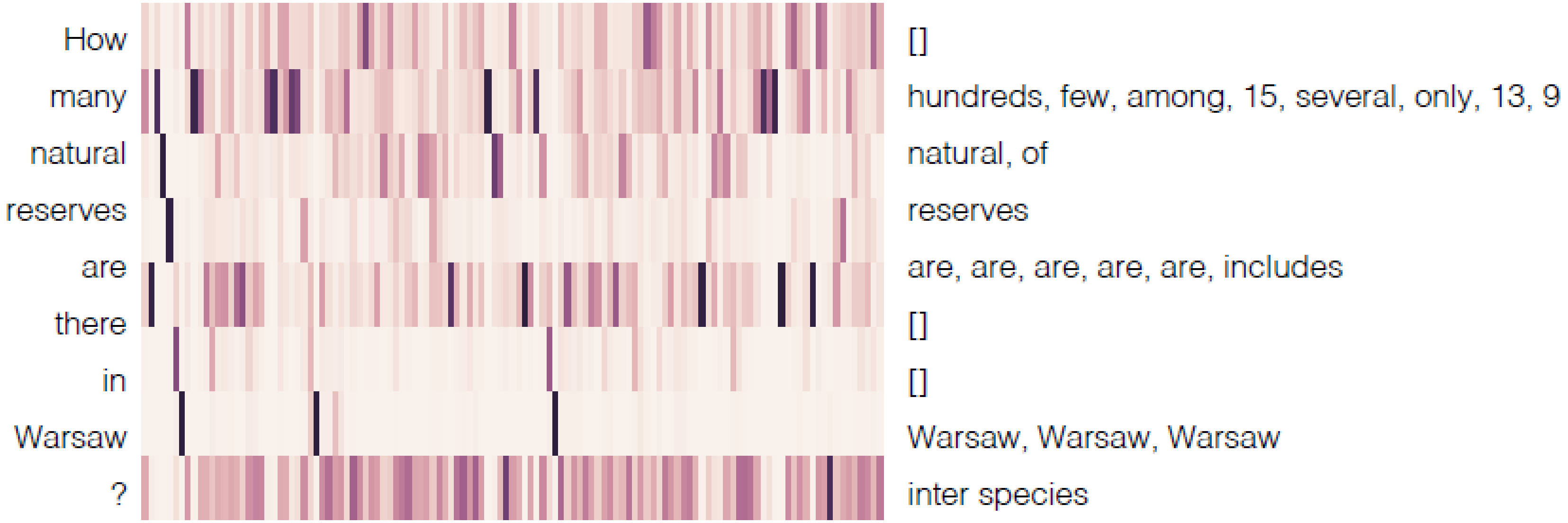
	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BIDAF (Ours)	68.0	77.3	**73.3**	**81.1**

(a) Results on the SQuAD test set

- BIDAF (ensemble) achieves SOTA performance
- Attention-based method is effective in QA task

Interpretation via Attention Maps

There are 13 natural reserves in Warsaw–among others, Bielany Forest, Kabaty Woods, Czerniaków Lake . About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the otter, beaver and hundreds of bird species. There are also several lakes in Warsaw – mainly the oxbow lakes, like Czerniaków Lake, the lakes in the Łazienki or Wilanów Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent–the majority are emptied before winter to clean them of plants and sediments.



Layer	Query	Closest words in the Context using cosine similarity
Word	When	when, When, After, after, He, he, But, but, before, Before
Contextual	When	When, when, 1945, 1991, 1971, 1967, 1990, 1972, 1965, 1953
Word	Where	Where, where, It, IT, it, they, They, that, That, city
Contextual	Where	where, Where, Rotterdam, area, Nearby, location, outside, Area, across, locations
Word	Who	Who, who, He, he, had, have, she, She, They, they
Contextual	Who	who, whose, whom, Guiscard, person, John, Thomas, families, Elway, Louis

BERT: A Pretraining and fine-tuning-based approach

"BERT: Pre-training of Deep Bidirectional Transformers [Vaswan et al., NeurIPs 2017] for Language Understanding"

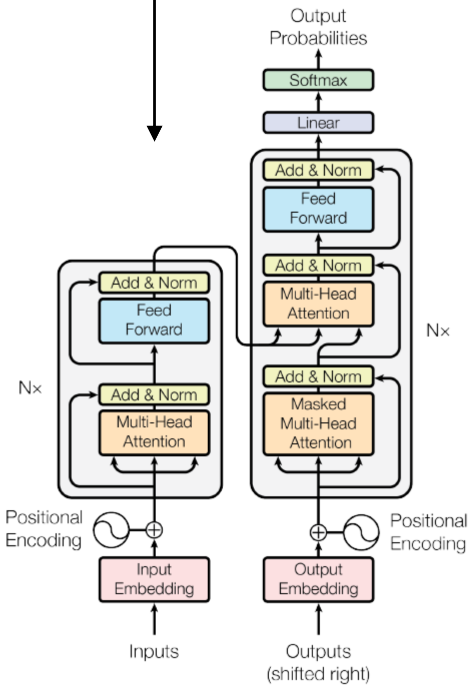
Vaswan et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

BERT: A Pretraining and fine-tuning-based approach (Cont'd)

"BERT: Pre-training of Deep Bidirectional **Transformers** [Vaswan et al., NeurIPs 2017] for Language Understanding"

Transformer is a **self-attention**-based backbone model

* However, we will not delve into Transformer in this talk



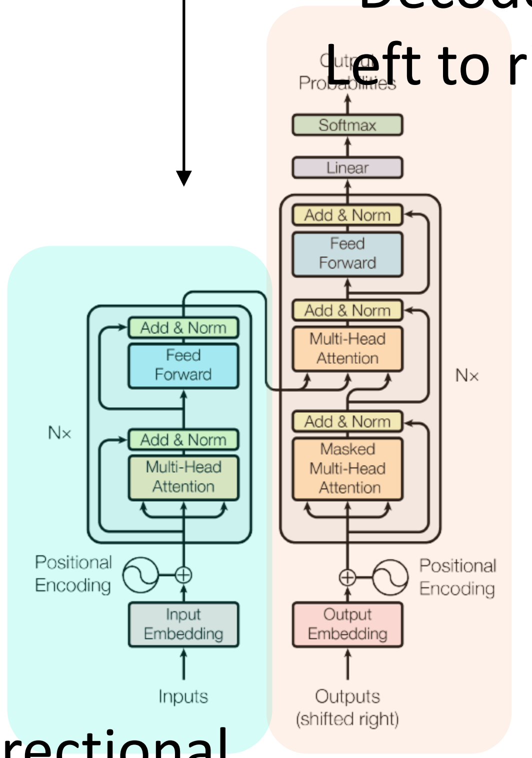
Vaswan et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

BERT: A Pretraining and fine-tuning-based approach (Cont'd)

"BERT: Pre-training of Deep **Bidirectional** Transformers [Vaswan et al., NeurIPs 2017] for Language Understanding"

Decoder:
Left to right

BERT only uses **encoder part** of transformer !
: bidirectional



Encoder: bi-directional

Vaswan et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

BERT: A Pretraining and fine-tuning-based approach (Cont'd)

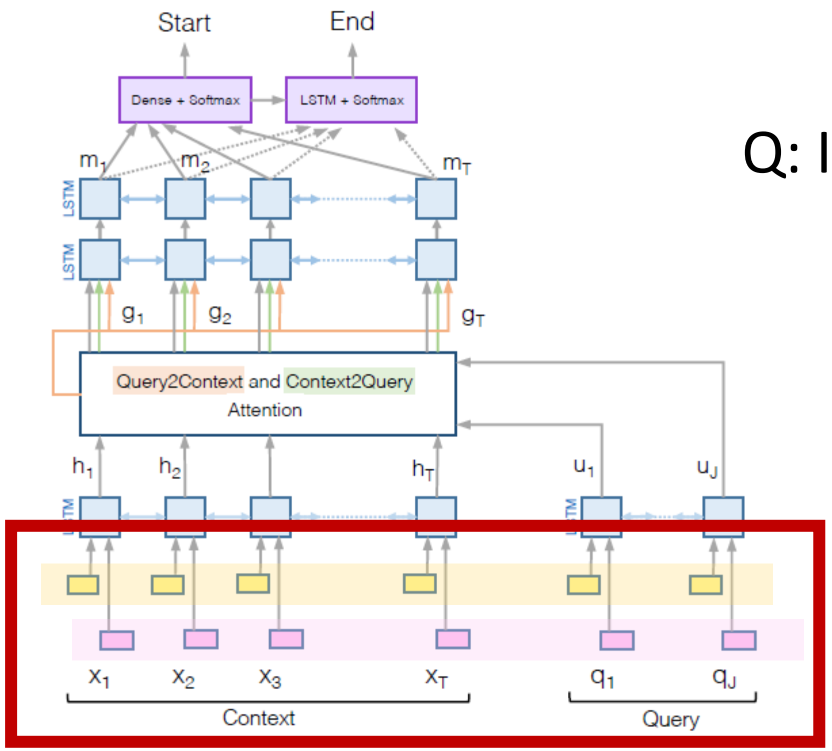
"BERT: **Pre-training** of Deep Bidirectional Transformers [Vaswan et al., NeurIPs 2017] for Language Understanding"

1. BERT achieves state-of-the-art method with fine-tuning with small modification (e.g., additional 1-layer)!

2. We will focus only on QA task & its results

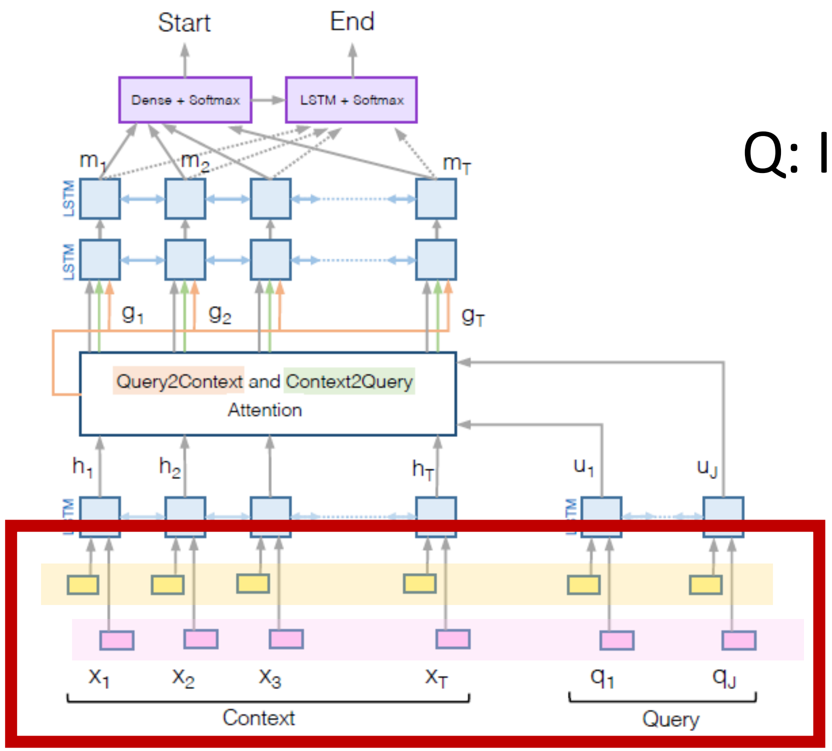
Vaswan et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

In Comparison with BIDAF



Q: Isn't it true that **BIDAF also uses pre-trained embeddings**?

In Comparison with BIDAF (Cont'd)

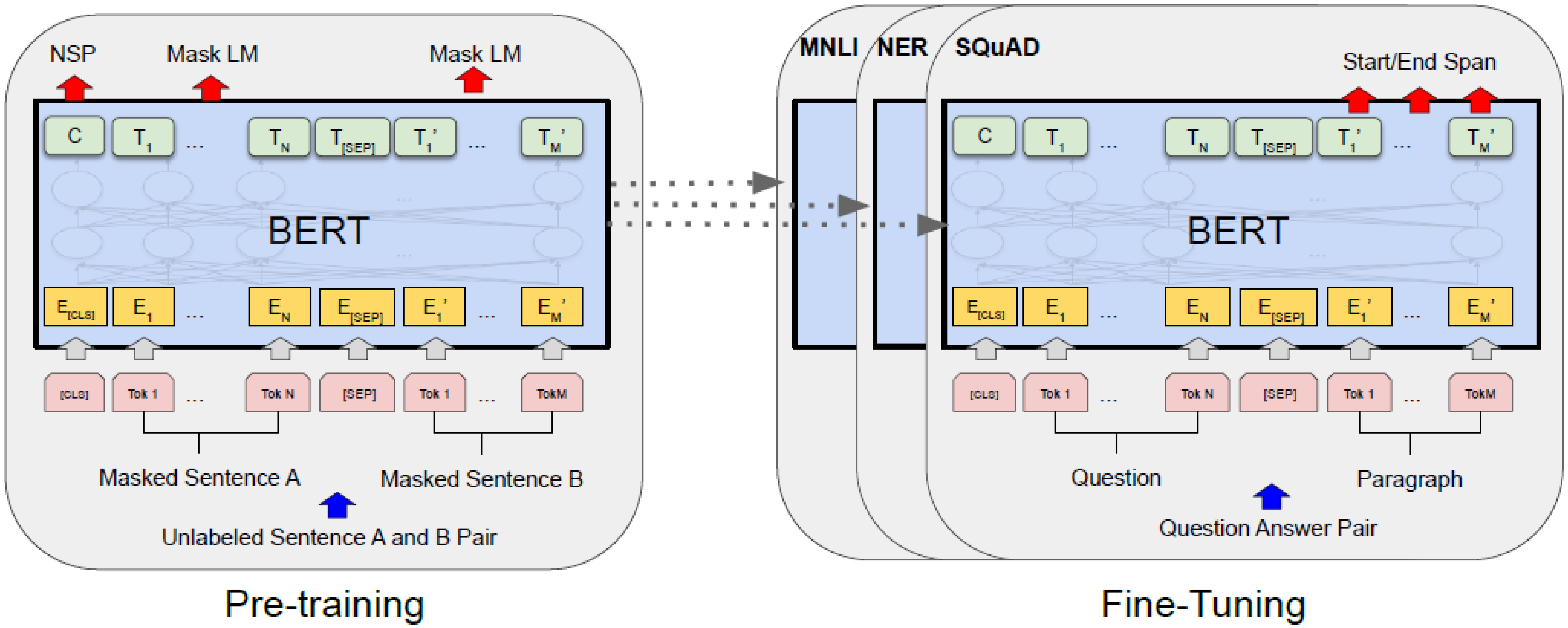


Q: Isn't it true that **BIDAF also uses pre-trained embeddings**?

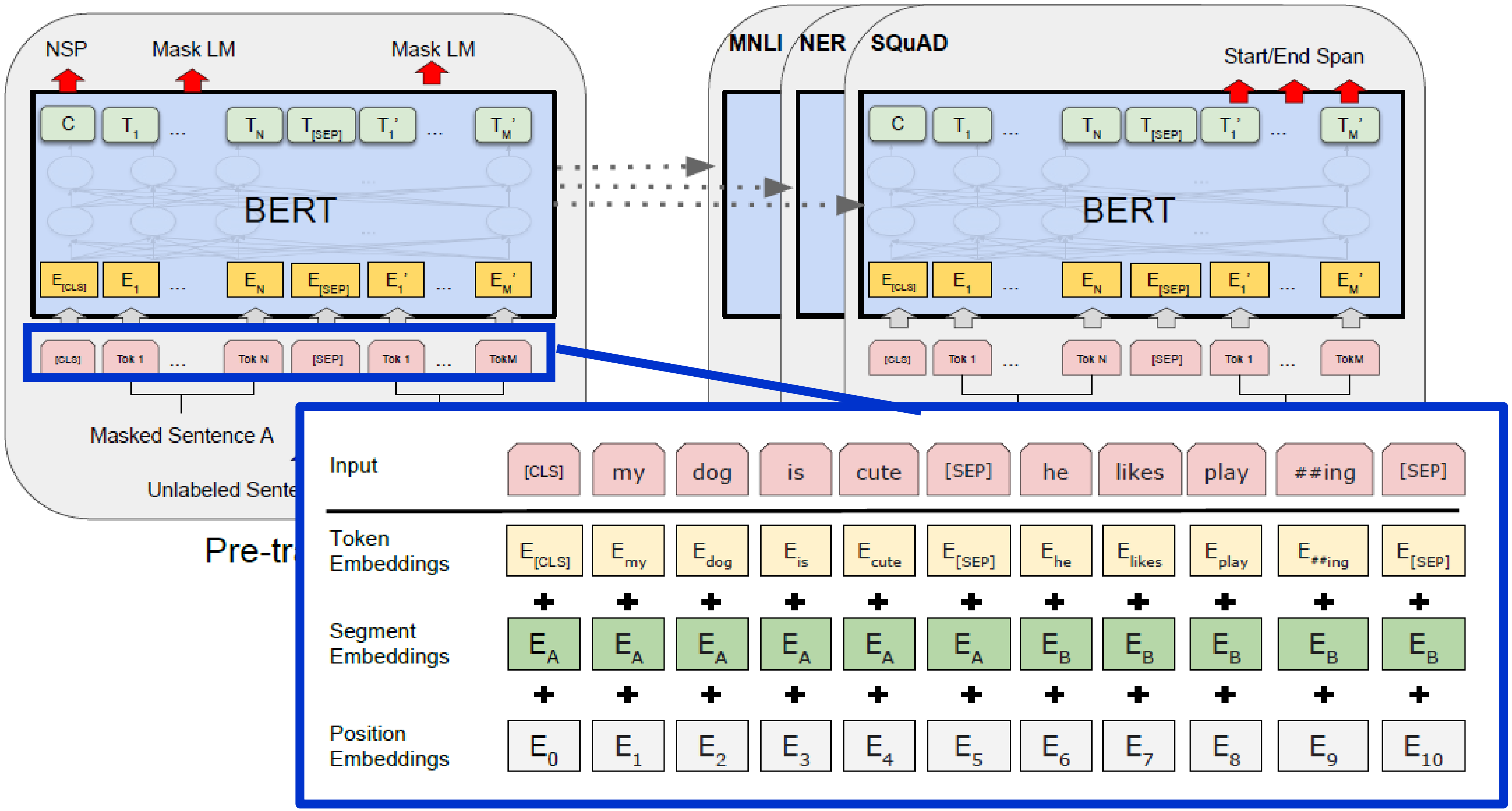
A: Yes. But BIDAF is a **two-stage** model

- (C and Q) text pairs are independently encoded
- Then, bidirectional cross attention is applied
- + not generally applicable to varying task

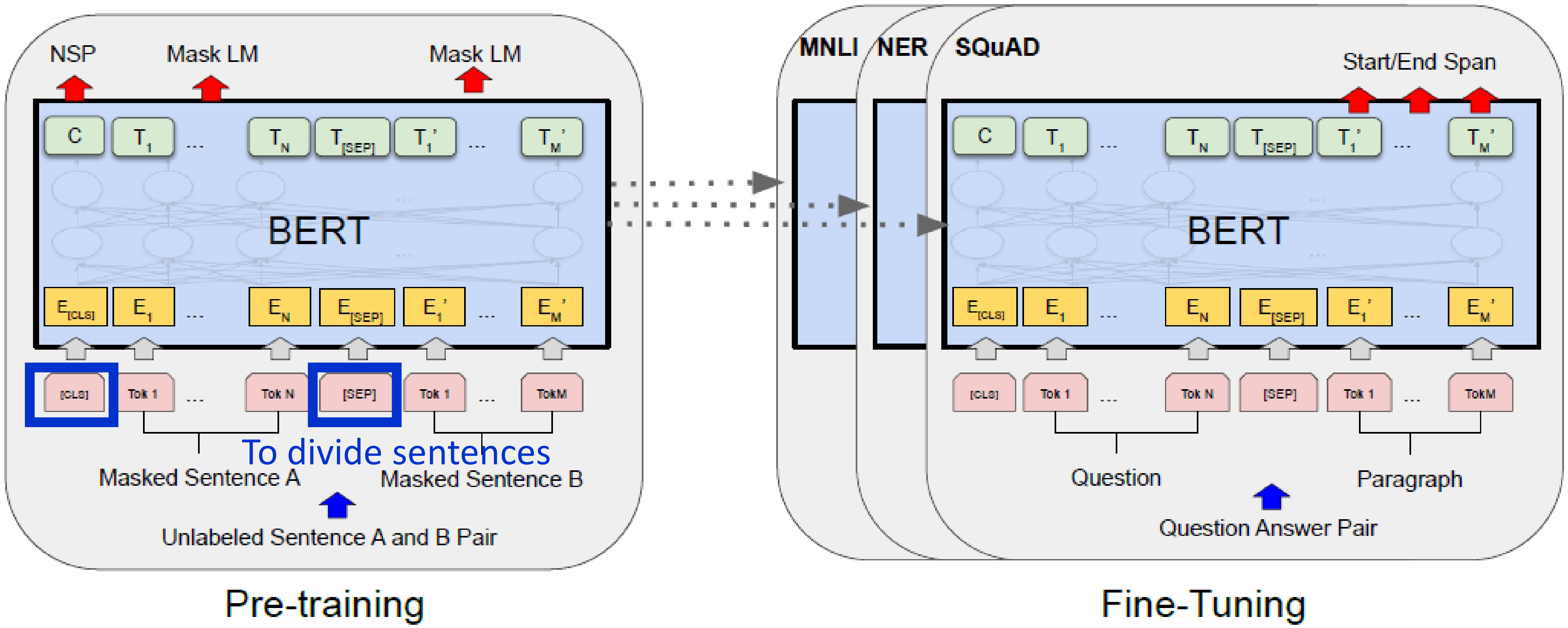
Overall Procedure of BERT



Overall Procedure of BERT (Cont'd)



Overall Procedure of BERT (Cont'd)



- BERT unifies the 2-stage architecture (such as BIDAF) via self-attention
- Masked LM (random masking) & next sentence prediction (NSP) is used for pre-training BERT model

Overall Procedure of BERT (Cont'd)

- **Masked LM**
  - Mask out some portion of labels, and predict the masked words
  - 15% is a normal case

The man went to the [MASK] to buy a [MASK] of milk

store gallon

- **NSP**
  - Learn relationships between sentences
  - Predict whether a given sentence is next sentence or not
  - Binary classification

Sentence A	The men went to the store	The men went to the store
Sentence B	He bought a gallon of milk	I like the cat
Label	TRUE	FALSE

Overall Procedure of BERT (Cont'd)

- **Masked LM** for **word-level** prediction, **NSP** to predict **sentence-level** prediction

- Masked LM and NSP is **jointly trained** in pretraining steps

```
https://github.com/codertimo/BERT-pytorch/blob/master/bert_pytorch/trainer/pretrain.py

Line 100 ~
            # 1. forward the next_sentence_prediction and masked_lm model
            next_sent_output, mask_lm_output = self.model.forward(data["bert_input"], data["segment_la

            # 2-1. NLL(negative log likelihood) loss of is_next classification result
            next_loss = self.criterion(next_sent_output, data["is_next"])

            # 2-2. NLLLoss of predicting masked token word
            mask_loss = self.criterion(mask_lm_output.transpose(1, 2), data["bert_label"])

            # 2-3. Adding next_loss and mask_loss : 3.4 Pre-training Procedure
            loss = next_loss + mask_loss
```

Evaluation of BERT on QA task

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\mathrm{BERT}_{\mathrm{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

- QA results on SQuAD 2.0 dataset
- Single model still achieves SOTA performance
- Almost reaches at human annotation results

Evaluation of BERT on QA task (Cont'd)

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)	
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5	
No NSP	83.9	84.9	86.5	92.6	87.9	-4.7
LTR & No NSP	82.1	84.3	77.5	92.1	77.8	-14.3
+ BiLSTM	82.1	84.1	75.7	91.6	84.9	-6.7

- NSP: next sentence prediction while pre-training
- LTR: Only left to right
- BiLSTM: Use BiLSTM instead of bidirectional transformer

Takeaways

- Attention is important for machine comprehension
- Bidirectional encoding is important for machine comprehension
- BERT can be more generally applicable for various downstream tasks

"Success is not final, failure is not fatal:
it is the courage to continue that counts."
- Winston Churchill

Thank you!

jindeok6@yonsei.ac.kr